



**MDMC**

Master in Data Management  
and Curation

## MASTER in DATA Management And CuRATION

# Implementation of FAIR Principles for XPS Data at MODA Laboratory through NeXus File Format

*Supervisor(s):*

*Dr. Tommaso Rodani*

*Prof. Emiliano Di Gennaro*

*Candidate:*

*Mohadeseh Zandavifard*

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, Mohadeseh Zandavifard, declare that this thesis entitled, 'Implementation of FAIR Principles for XPS Data at MODA Laboratory through NeXus File Format' and the work presented therein are my own.

I certify that:

- This work was performed wholly or principally during MDMC internship in MODA LABORATORY of SPIN-CNR.
- 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 Claude (<https://claude.ai/>) and ChatGPT (<https://chatgpt.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.
- 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: Mohadeseh Zandavifard



Date: 29/04/2025

*“If we knew what it was we were doing, it would not be called research, would it?”*

*— Albert Einstein*

# Acknowledgments

I would like to express my deepest gratitude to all those who supported and guided me throughout the development of this thesis.

First and foremost, I warmly thank **Dr. Mariarita De Luca**, coordinator of the MDMC program, for her constant presence, thoughtful encouragement, and insightful support throughout every stage of this academic journey.

I am especially grateful to my supervisors, **Dr. Tommaso Rodani** and **Prof. Emiliano Di Gennaro**, for their invaluable mentorship, constructive feedback, and steady encouragement. I also wish to sincerely thank **Dr. Federica Bazzocchi** for her essential support and contributions. Their expertise has played a pivotal role in shaping the direction and substance of this work.

I extend special thanks to **Dr. Mohammadhassan Valadan** and the team at the **MODA Laboratory, CNR-SPIN, Naples**, for their collaboration and openness in sharing data and insights. Their contributions significantly enhanced my understanding of X-ray Photoemission Spectroscopy and the real-world challenges of scientific data management.

I am also grateful to the instructors of the **Master's course in Data Management and Curation at SISSA**, whose teachings on FAIR principles and Open Science provided the conceptual foundation for this thesis.

I deeply appreciate the company and support of my fellow students and colleagues, particularly **Rosario Forlenza, Mattia Marella, Matteo Francesco Bontorno**, and **Dario De Angelis**, for their patience, technical assistance, and constant encouragement. Their help in tackling the practical aspects of data formats and digital tools was invaluable.

On a personal note, I am profoundly thankful to my family for their unconditional love, patience, and support—especially during moments of stress and the many long nights spent at the computer.

Finally, I dedicate this thesis to the light of my life, **my daughter, Diana**. Your joy, curiosity, and laughter are my greatest inspiration.

# Abstract

The digital transformation of scientific research has placed increasing emphasis on data that is not only accessible but also well-structured, reusable, and interoperable. In this context, the FAIR principles—Findability, Accessibility, Interoperability, and Reusability—have become foundational for modern data management practices, especially in large-scale collaborative research infrastructures. This thesis addresses the implementation of FAIR data practices in the context of X-ray Photoelectron Spectroscopy (XPS) experiments conducted at the MODA Laboratory of the CNR-SPIN institute in Naples, Italy. Specifically, it focuses on the development of a Python-based data conversion pipeline that translates raw experimental XPS data into the NeXus file format, a community standard that ensures metadata-rich, hierarchically structured, and machine-readable files.

The research is framed within the broader landscape of Open Science and European research infrastructure initiatives such as NFFA-DI (Nano Foundries Fine Analysis - Digital Infrastructure), which seek to harmonize data handling across multiple scientific domains. The MODA lab, involved in both experimental research and digital innovation, presents an ideal case study for piloting the application of FAIR principles to surface science data. Challenges in current data management practices - such as inconsistent metadata recording, non-standard file formats, and limited integration with centralized data repositories - are systematically addressed in the development of this pipeline.

The proposed workflow is built around the `pynxtools[xps]` package, extended through custom scripting to enable the extraction of critical metadata from both ELNs (Electronic Lab Notebooks) and instrument-generated files (.vms format). The user is guided through a semi-automated process that includes file selection, metadata generation or reuse, and NeXus file creation. To improve usability and reproducibility, the pipeline includes interactive prompts for choosing between existing and new metadata, applies file naming conventions, and preserves version control by appending timestamps. This design ensures that each NeXus file is uniquely identifiable and traceable to its experimental origin.

The integration of metadata from eLabFTW, an open-source ELN currently being adopted at MODA, is also discussed as a forward-looking extension. By

aligning the metadata structure in the NeXus files with the ontology used in the ELN, the system aims to establish a robust chain of provenance that spans from experimental design to data publication. The eventual goal is to make XPS data not only FAIR within the MODA lab but also compatible with national and international data ecosystems such as NOMAD and the NFFA-DI central data hub.

This work demonstrates that the adoption of FAIR principles is not only technically feasible in a real laboratory setting but also enhances scientific efficiency and transparency. It shows how thoughtfully designed software tools can bridge the gap between experimental research and data management. By facilitating structured data sharing and long-term preservation, this pipeline lays the groundwork for reproducible, collaborative surface science in the age of digital research infrastructures.

# Contents

|                                                                         |           |
|-------------------------------------------------------------------------|-----------|
| <b>Author's Declaration .....</b>                                       | <b>3</b>  |
| <b>Acknowledgments .....</b>                                            | <b>4</b>  |
| <b>Abstract.....</b>                                                    | <b>5</b>  |
| <b>Contents .....</b>                                                   | <b>7</b>  |
| <br>                                                                    |           |
| <b>Chapter 1 Introduction to Open Science and FAIR Principles .....</b> | <b>10</b> |
| 1.1. Open Science Framework .....                                       | 10        |
| 1.2. FAIR Principles in Detail .....                                    | 11        |
| 1.2.1. Findable .....                                                   | 11        |
| 1.2.2. Accessible.....                                                  | 11        |
| 1.2.3. Interoperable.....                                               | 12        |
| 1.2.4. Reusable .....                                                   | 12        |
| 1.3. The NFFA-DI Project .....                                          | 12        |
| 1.4. Importance of Standardization in Nanoscience .....                 | 14        |
| 1.5. Challenges in Implementing FAIR Data Practices .....               | 15        |
| 1.5.1. Technical Challenges .....                                       | 15        |
| 1.5.2. Cultural and Organizational Challenges .....                     | 16        |
| 1.5.3. Cost and Sustainability Concerns .....                           | 17        |
| 1.6. Conclusion .....                                                   | 17        |

|                                                                                         |               |
|-----------------------------------------------------------------------------------------|---------------|
| <b>Chapter 2 MODA Laboratory Environment and X-ray Photoemission Spectroscopy .....</b> | <b>19</b>     |
| 2.1. Introduction to MODA Laboratory at CNR-SPIN .....                                  | 19            |
| 2.2. Research Activities and Materials Growth .....                                     | 20            |
| 2.2.1. Pulsed Laser Deposition (PLD) .....                                              | 20            |
| 2.2.2. Material Characterization Techniques .....                                       | 20            |
| 2.3. X-ray Photoemission Spectroscopy (XPS) .....                                       | 21            |
| 2.3.1. Physical Principles of XPS .....                                                 | 21            |
| 2.3.2. The XPS Instrument at MODA Laboratory .....                                      | 21            |
| 2.3.3. XPS Measurement Process .....                                                    | 22            |
| 2.3.4. Information Obtained from XPS Analysis .....                                     | 22            |
| 2.3.5. XPS Data Structure and Interpretation .....                                      | 23            |
| 2.3.6. Applications of XPS in Materials Science .....                                   | 24            |
| 2.3.7. Challenges in XPS Data Management .....                                          | 24            |
| 2.4. Data Management Practices and Developments at MODA .....                           | 25            |
| 2.4.1. Integration of eLabFTW for Digital Documentation .....                           | 26            |
| 2.4.2. Data Storage, Formats, and Organization .....                                    | 26            |
| 2.4.3. PLD Data Integration with NOMAD .....                                            | 26            |
| 2.5. NFFA-DI Integration at MODA .....                                                  | 27            |
| 2.5.1. Data Flow to NFFA-DI Central Services .....                                      | 27            |
| 2.5.2. Standards Compliance Requirements .....                                          | 27            |
| 2.6. The Need for Standardized XPS Data Formats .....                                   | 28            |
| 2.7. Challenges in the Laboratory Environment .....                                     | 29            |
| 2.7.1. Balancing Research Productivity and Data Management .....                        | 29            |
| 2.7.2. Technical Infrastructure Limitations .....                                       | 29            |
| 2.7.3. Knowledge Transfer and Training .....                                            | 30            |
| 2.8. Anticipating Future Developments at MODA .....                                     | 30            |
| 2.8.1. Expansion of Instrumentation .....                                               | 30            |
| 2.8.2. Managing Growing Data Volumes .....                                              | 31            |
| 2.9. Research Focus and Infrastructure at MODA .....                                    | 31            |
| 2.10. Conclusion .....                                                                  | 32            |
| <br><b>Chapter 3 .....</b>                                                              | <br><b>33</b> |
| <b>Implementation of FAIR Principles for XPS Data .....</b>                             | <b>33</b>     |
| 3.1. Introduction .....                                                                 | 33            |
| 3.2. Overview of the Conversion Pipeline .....                                          | 33            |
| 3.3. Components of the Data Conversion Pipeline .....                                   | 34            |
| 3.3.1. Input Data Sources .....                                                         | 34            |
| 3.3.2. Configuration Files .....                                                        | 35            |
| 3.3.3. Conversion Software .....                                                        | 35            |
| 3.4. Implementation Details .....                                                       | 36            |
| 3.4.1. Python Converter Script .....                                                    | 36            |
| 3.4.2. Metadata Collection .....                                                        | 38            |
| 3.4.3. Configuration File Structure .....                                               | 40            |

|                                                     |                                                                     |    |
|-----------------------------------------------------|---------------------------------------------------------------------|----|
| 3.5.                                                | FAIR Principles Implementation .....                                | 41 |
| 3.5.1.                                              | Findability.....                                                    | 41 |
| 3.5.2.                                              | Accessibility .....                                                 | 42 |
| 3.5.3.                                              | Interoperability .....                                              | 43 |
| 3.5.4.                                              | Reusability .....                                                   | 43 |
| 3.6.                                                | Usage Example.....                                                  | 44 |
| 3.7.                                                | Challenges and Solutions .....                                      | 46 |
| 3.8.                                                | Future Improvements and Integration with Laboratory Workflows<br>47 |    |
| 3.9.                                                | Conclusion .....                                                    | 48 |
| Chapter 4 Conclusions and Future Work .....         |                                                                     | 49 |
| 4.1.                                                | Summary of Achievements.....                                        | 49 |
| 4.2.                                                | Impact on Research Practices .....                                  | 50 |
| 4.3.                                                | Limitations of the Current Implementation .....                     | 51 |
| 4.4.                                                | Future Directions .....                                             | 52 |
| 4.5.                                                | Concluding Remarks .....                                            | 54 |
| Appendices.....                                     |                                                                     | 56 |
| Appendix A: Installation and Setup Guide .....      |                                                                     | 56 |
| Appendix B: NeXus File Structure and Hierarchy..... |                                                                     | 57 |
| Bibliography .....                                  |                                                                     | 58 |

# **Chapter 1**

## **Introduction to Open Science and FAIR Principles**

### **1.1. Open Science Framework**

Science is changing rapidly. The ways researchers conduct, share, and evaluate their work are evolving dramatically. Open Science represents a transformative movement aimed at making scientific research accessible to everyone [1]. It's revolutionary. This approach extends beyond traditional publication methods. It promotes transparency. It encourages accessibility. It fosters collaboration throughout the entire research process.

During my Master course in Data Management and Curation at SISSA, I gained insights into Open Science. It wasn't just theoretical stuff. Open Science offers practical frameworks that enhance scientific outcomes considerably [2]. There are many stories about the problems with restricted access; when researchers can't access previous studies or raw data, they often repeat work unnecessarily [3]. They waste time and miss crucial opportunities to build upon existing knowledge. This creates inefficiency in the scientific enterprise.

The movement for Open Science promotes several interconnected practices. Open Access to scientific publications ensures everyone can read research findings. Open Research Data allows verification and reuse of experimental results. Open Source software enables anyone to use and improve scientific tools. Open Peer Review makes evaluation transparent. Citizen Science brings non-professionals into the research process. These elements work together to create a more transparent scientific ecosystem.

My limited experience at the MODA laboratory (Modular Facility for Oxide Deposition and Analysis) in CNR-SPIN in Naples gave me sights of how these principles might apply to experimental physics research. I wasn't there very often. But I noticed documentation was minimal. Data typically resided on local computers without standardized global naming conventions. Metadata was sparse. Understanding experimental conditions often required talking directly to

whoever did the experiment originally. However, there is the movement to standardize documentation of lab notes, and infrastructures are being made to have an electronic lab notebook. We should also consider that the transformation won't be immediate. Change happens slowly in established laboratories. But even small shifts from personal notes to structured systems represented meaningful progress toward openness.

## **1.2. FAIR Principles in Detail**

The FAIR principles - Findable, Accessible, Interoperable, and Reusable - were developed in response to the growing complexity of scientific data and the urgent need for better data management practices [4]. As research increasingly relies on digital resources and collaborations across disciplines and institutions, the ability to share, interpret, and reuse data has become a cornerstone of effective and transparent science. FAIR does not refer to a specific technology or standard but instead outlines a set of guiding principles meant to improve how research data are handled, shared, and preserved.

The FAIR principles are not only about improving data storage or access but about rethinking the entire research process to make it more transparent, efficient, and reproducible. These principles are especially relevant in collaborative research infrastructures and large-scale data platforms, which depend on well-managed and shareable data to function effectively.

### **1.2.1. Findable**

The first of these principles, Findable, emphasizes that data should not be hidden in isolated folders or locked behind obscure systems. Instead, datasets should be described with sufficiently rich metadata that clearly define what the data are, how they were generated, and under what conditions they can be used. This metadata should be indexed in searchable resources, allowing both humans and machines to discover them. Assigning globally unique and persistent identifiers, such as DOIs, plays an important role in making data reliably findable across systems and over time [5].

### **1.2.2. Accessible**

Being Accessible means that once someone has located a dataset, they should be able to retrieve it or at the very least, obtain information about how to access it. Importantly, accessibility does not always mean that data must be freely available to everyone. Rather, the conditions of access should be clearly stated,

whether the data are open, restricted, or require special permissions. Protocols for accessing the data should be transparent and robust, and crucially, the metadata should remain accessible even if the data themselves are no longer available. This ensures that the scientific record remains complete and traceable.

### **1.2.3. Interoperable**

The principle of Interoperability addresses the problem of data compatibility. Scientific research is increasingly interdisciplinary, which means data often needs to be combined or compared across different fields, tools, and platforms. Without common formats or shared vocabulary, this integration becomes extremely difficult. Interoperability encourages the use of standard data formats and formally defined terminologies, often maintained through domain-specific ontologies. These standards enable data to be used in automated workflows and to be interpreted correctly by software systems beyond their original context.

### **1.2.4. Reusable**

Finally, data must be Reusable. This involves more than just being able to access or read a file. To be genuinely reusable, data must be described in a way that someone outside the original research group can understand and evaluate [6]. This includes providing sufficient detail about the methods used, the context of data collection, and the data's intended scope. A clearly stated license is also essential, as it defines how the data can be reused, shared, or modified. Furthermore, aligning with community standards ensures that datasets are compatible with the tools and practices already familiar to researchers in that domain, making their reuse more practical and impactful.

## **1.3. The NFFA-DI Project**

The NFFA-DI (Nano Foundries Fine Analysis - Digital Infrastructure) project is a national Italian initiative with European significance, aiming to reshape how scientific data in nanoscience is managed, shared, and reused [7]. It addresses the critical need for a coordinated, FAIR-compliant digital infrastructure that can support the increasingly data-intensive nature of nanoscience research [8]. The NFFA-DI project seeks to create open-access, interoperable platforms that connect advanced experimental facilities with robust data services, ultimately fostering a more integrated and transparent research environment [9].

At its core, NFFA-DI supports open science by developing tools, protocols, and standards that enable seamless collaboration among distributed research

facilities. It facilitates the harmonization of data management practices across diverse experimental domains, such as spectroscopy, microscopy, and materials synthesis. The goal is to ensure that the data produced through these techniques is not only stored and preserved, but also structured in a way that makes it accessible and reusable by other researchers, now and in the future.

A central challenge in nanoscience is the complexity of its experimental workflows, which often involve multiple instruments and produce heterogeneous data formats. The NFFA-DI project responds to this challenge by organizing working groups dedicated to specific classes of techniques. One such group focuses on spectroscopy, particularly X-ray Photoemission Spectroscopy (XPS), and is engaged in developing standardized metadata schemas and data models. These efforts aim to make experimental data interoperable between laboratories, instruments, and software environments [10].

Among the partner institutions involved in NFFA-DI is CNR-SPIN (Istituto di Superconduttori, Materiali Innovativi e Dispositivi), a research institute under the Italian National Research Council. Within CNR-SPIN, the MODA laboratory in Naples plays an active role in applying and testing the project's data infrastructure principles, particularly in relation to experimental workflows involving surface and interface characterization techniques.

My experience with the MODA lab was closely tied to the foundations I acquired during a Master course focused on data management and FAIR principles. This course emphasized the importance of structured metadata, standardized data formats, and reproducibility in modern scientific practice. During my visits to the laboratory, I was able to observe real-world research processes and explore how concepts from the course could be translated into practical workflows. In particular, I attempted to understand how XPS data might be adapted to fit emerging standards for metadata and structure, such as those proposed within the NFFA-DI framework. Although my time at the lab was limited, the experience allowed me to bridge the gap between theoretical learning and hands-on application.

Through this interaction with the NFFA-DI project and the MODA lab, I gained a deeper appreciation for the challenges and opportunities involved in building a sustainable, FAIR-aligned data infrastructure for complex scientific domains like nanoscience. My ongoing work reflects an effort to continue this learning process, contributing - where possible - the broader goals of interoperability, openness, and data-driven discovery that NFFA-DI promotes.

## 1.4. Importance of Standardization in Nanoscience

In the context of modern scientific research, the importance of standardization is obvious, particularly in fields as diverse and technique intensive as nanoscience. Standardization plays a fundamental role in enabling the implementation of FAIR (Findable, Accessible, Interoperable, Reusable) principles, which are now widely recognized as essential for sustainable, open, and collaborative science. Without common protocols, terminologies, and data structures, the utility of research data is limited, and cross-disciplinary collaboration becomes increasingly difficult [11].

Nanoscience presents a unique set of challenges in this regard. Unlike more consolidated scientific fields, nanoscience encompasses a wide range of instruments, methodologies, and experimental conditions. Each laboratory may use highly customized setups or specialized data acquisition systems, which results in a abundance of data formats and storage practices. This fragmentation often leads to isolated datasets, inaccessible outside their original context, and incompatible with external tools or infrastructures.

Numerous studies from adjacent fields such as genomics and astronomy have shown the transformative power of standardization [12, 13]. In those domains, the introduction of shared data models and standardized metadata schemas has enabled large-scale data integration, facilitating discoveries that would have been impossible through isolated analyses. These examples emphasize the network effects generated when communities adopt common standards: as more researchers contribute to and draw from standardized datasets, the cumulative value and reusability of the data increase significantly.

These considerations are particularly relevant for techniques like X-ray Photoemission Spectroscopy (XPS), which is widely used in nanoscience for surface characterization. Standardizing the format and description of XPS data offers numerous benefits [14]. When data is consistently structured and accompanied by comprehensive metadata, reproducibility is enhanced, as others can understand and replicate the experimental conditions. Standard formats also enable the use of shared analysis tools, allowing results to be verified and extended by different groups. Moreover, consistent data structures are essential for the application of machine learning methods, which rely on large, harmonized training datasets. For long-term preservation, the use of non-proprietary, well-documented formats ensures that data can remain accessible and interpretable even as software environments evolve.

Despite these advantages, the implementation of standards is not without challenges. It requires agreement across institutions, sustained community engagement, and sometimes, the redesign of local practices [15]. Difficulties arise when attempting to accommodate the full complexity of experimental

methods while still maintaining usability. Standards that are too detailed may overwhelm researchers with excessive documentation requirements, while those that are too simple risk omitting critical contextual information. Achieving an effective balance between these extremes is essential, and this task must be informed by both domain knowledge and the practical realities of laboratory workflows.

In response to these challenges, projects like NFFA-DI have taken a leading role in fostering consensus around data models and metadata for specific nanoscience techniques. By involving both data scientists and domain experts, working groups have been formed to define and test standards for techniques such as spectroscopy, including XPS. Their work provides the foundation for interoperable data infrastructures that can support long-term scientific collaboration across Europe and beyond.

In summary, standardization is not merely a technical detail but a foundational element for enabling open science in nanoscience. It supports data sharing, enhances scientific reproducibility, and opens the door to advanced computational methods. While the path to widespread adoption may be complex, the long-term benefits make it an essential objective for any effort seeking to modernize research data management.

## **1.5. Challenges in Implementing FAIR Data Practices**

While the FAIR principles offer a compelling framework for modern scientific data management, their practical implementation continues to present significant challenges across research environments. These difficulties arise from both technical limitations and broader cultural, organizational, and economic constraints. In particular, within the context of spectroscopy data in nanoscience, such as X-ray Photoemission Spectroscopy (XPS), the path to FAIR compliance is often hindered by the complexity of experimental data, legacy formats, and fragmented workflows.

### **1.5.1. Technical Challenges**

From a technical standpoint, the conversion of existing experimental data into FAIR-compliant formats requires specialized domain knowledge combined with data science expertise. In spectroscopy, data are frequently generated in proprietary formats such as .vms files commonly used for XPS measurements that lack comprehensive public documentation. These files cannot easily be

parsed or integrated into standardized systems without the development of custom extraction tools, which must often be designed from scratch. The creation of such tools requires familiarity with both the experimental methodology and the target data models used for FAIR representation, such as NeXus or other community-endorsed standards.

Another technical barrier relates to the infrastructure necessary for storing and preserving FAIR data. The inclusion of rich metadata, versioning, and multiple export formats tends to increase both the size and complexity of stored data objects. In many laboratories, existing storage systems are optimized for short-term, locally accessible use, often focused on the immediate needs of active researchers rather than long-term preservation or external sharing. Upgrading these systems to support long-term accessibility and interoperability would require significant investment in both hardware and data management infrastructure, including the implementation of repositories capable of persistent identification, metadata indexing, and secure access protocols.

## **1.5.2. Cultural and Organizational Challenges**

Beyond technical considerations, many of the most persistent obstacles to FAIR adoption hinder from cultural and institutional norms. In academic research, performance and recognition continue to be closely tied to publications and citation counts, rather than to the quality or accessibility of the underlying data [16]. As a result, incentives for rigorous data management are often weak or nonexistent. The time and effort required to produce well-documented, reusable datasets are typically viewed as secondary to the primary objectives of experimentation and analysis.

Resistance to documentation practices is frequently observed, particularly in fast-paced research environments where immediate experimental outcomes are prioritized. The preparation of metadata, the adherence to file-naming conventions, and the validation of data completeness are sometimes perceived as administrative burdens rather than essential components of scientific transparency. Shifting this perception would require a fundamental change in research culture—one that values data management as an integral part of the scientific process.

In addition, concerns related to intellectual property and academic competition present further barriers to data sharing. In highly competitive research areas, the fear often discourages researchers from making data openly available prior to publication [17]. This has led to the widespread adoption of embargo strategies, where data are preserved and described but only made accessible after an agreed delay. While such approaches help mitigate risk, they depend on institutional policies and technical systems that are not always in place.

### **1.5.3. Cost and Sustainability Concerns**

The financial implications of FAIR implementation also deserve serious consideration. Proper data management over the full lifecycle of a research project from acquisition and processing to preservation and sharing, requires dedicated resources, including skilled personnel, digital infrastructure, and ongoing technical support [18]. However, data management costs are often underestimated or omitted entirely in research proposals. Funding allocations tend to prioritize instrumentation and research staff, leaving little room for the infrastructural investments required to sustain FAIR-compliant workflows.

Moreover, it is increasingly recognized that one-time efforts to “make data FAIR” are insufficient. Long-term success in this area depends on embedding FAIR practices within standard operating procedures and institutional routines. This involves not only the creation of tools and templates but also continuous training, oversight, and strategic planning. Sustainable data stewardship must be viewed not as an isolated initiative but as a systemic shift in research practice; one that requires strong leadership, clear policies, and dedicated funding mechanisms.

## **1.6. Conclusion**

Open Science and FAIR principles represent both philosophy and practical framework. They aim to improve scientific research fundamentally. As I learned during my Master course, the transition to FAIR data practices requires technical solutions. It demands cultural changes. It needs organizational support. These elements must work together for successful implementation.

The following chapters will explore detailed applications of these principles to XPS data at the MODA laboratory. Chapter 2 will describe the laboratory environment and its data management ecosystem and will explain X-ray Photoemission Spectroscopy physics and techniques. This background will clarify data standardization challenges. Chapter 3 will detail technical implementation of converting XPS data to the NeXus format following the FAIRmat model [19]. Chapter 4 will discuss conclusions and future directions.

Through this thesis, I document both technical processes and broader context. My experience, though limited in practical application, has shown that data management transcends administrative function. It represents a fundamental component of good science. The principles explored here apply across scientific domains. They offer pathways to more reproducible, efficient, and collaborative research practices.

The journey from traditional data management to FAIR compliance continues. It faces obstacles and requires commitment. But its potential benefits for scientific progress make it worthwhile. My hope is that this thesis contributes to that journey, even if my personal contribution to the laboratory's transformation remained modest due to limited onsite experience.

# Chapter 2

## MODA Laboratory Environment and X-ray Photoemission Spectroscopy

### 2.1. Introduction to MODA Laboratory at CNR-SPIN

The MODA laboratory is housed within the CNR-SPIN Institute, part of Italy's National Research Council. MODA stands for “*Modular Facility for Oxide Deposition and Analysis*” [20]. The laboratory specializes in the growth and characterization of complex oxide materials and thin films, which have applications in electronics, energy storage, and medical devices.

The CNR-SPIN Institute focuses specifically on Superconductors, Innovative Materials, and Devices. The MODA laboratory was established with the goal of creating world-class capabilities for oxide materials research. The laboratory name directly reflects its dual focus: material creation through molecular beam epitaxy and analysis through spectroscopic characterization.

The laboratory is structured into distinct sections to serve specific research functions:

- The Thin Film Growth Area contains equipment for creating precisely controlled material layers, including the Pulsed Laser Deposition (PLD) system.
- The Materials Characterization Area features the X-ray Photoelectron Spectroscopy (XPS) instrument and other analytical tools.
- The Sample Preparation Area provides clean space for preparing substrates and samples.
- Data Analysis Workstations occupy another area, running specialized analysis software.

MODA laboratory participates in various research projects and collaborates extensively with universities, research institutions, and industry partners across Europe. This collaborative network enables knowledge sharing and resource pooling.

## **2.2. Research Activities and Materials Growth**

The research at MODA concentrates on growing and characterizing oxide thin films and heterostructures. These materials exhibit fascinating properties such as superconductivity, ferroelectricity, and unique magnetic behaviors, which make them particularly interesting for scientific study and practical applications.

### **2.2.1. Pulsed Laser Deposition (PLD)**

The primary method used at MODA for growing thin films is Pulsed Laser Deposition (PLD). In this process, a high-power laser focuses onto a target material inside a vacuum chamber [21]. The laser creates a plasma plume, and material from this plume deposits onto a heated substrate. Layer by layer, a thin film forms with precision down to atomic levels.

The PLD system offers flexibility for research, as it can grow various materials by changing targets and conditions. However, this versatility creates data management challenges. Each experiment involves numerous parameters, such as: Laser energy, Pulse frequency, Target-to-substrate distance, Substrate temperature, Chamber pressure, Gas composition, Deposition time. These parameters affect the resulting film properties and must be documented completely for reproducibility. The experimental process is often iterative, with researchers adjusting parameters based on results from previous attempts.

### **2.2.2. Material Characterization Techniques**

After growing films, researchers characterize them using various analytical techniques. The primary characterization tool is X-ray Photoelectron Spectroscopy (XPS), which provides detailed information about chemical composition and electronic states. Other characterization tools at MODA include Atomic Force Microscopy (AFM) for surface analysis, Electrical transport measurements, Optical spectroscopy, X-ray Diffraction (XRD) for studying crystal structures. Each technique generates unique data formats, including different file types and metadata requirements. Some use proprietary software, while others employ open standards. This diversity presents significant challenges for creating a unified approach to data management.

## 2.3. X-ray Photoemission Spectroscopy (XPS)

X-ray Photoemission Spectroscopy, sometimes called ESCA (Electron Spectroscopy for Chemical Analysis), is a fundamental technique for understanding the surface chemistry of materials [22]. It analyzes surface composition, revealing not only which elements exist on a material's surface but also how these elements bond with each other.

### 2.3.1. Physical Principles of XPS

XPS is based on the photoelectric effect, which was explained by Einstein in 1905. When X-rays bombard a sample surface, they can eject electrons from the inner shells of atoms [23]. These X-rays have precise energy levels and carry enough energy to cause electrons to escape both the atom and the material surface. The escaping electrons are called photoelectrons.

In atoms, electrons occupy different energy "shells" or levels. Each element has its own distinct arrangement of electrons in these shells. When X-rays knock out an inner shell electron, the ejected electron's energy depends on two factors: the X-ray energy and the binding energy that was holding the electron in place.

XPS measures the kinetic energy of these ejected electrons. Since the X-ray source energy is known, the binding energy can be calculated using this equation:

$$\text{Binding Energy} = \text{X-ray Energy} - \text{Kinetic Energy} - \text{Work Function}$$

The "work function" term is a small correction factor related to the instrument. Binding energies create distinctive patterns, with each element producing characteristic peaks. Different chemical states of the same element show subtle but measurable differences in these peaks. This creates a chemical "fingerprint" that identifies not just which elements are present but also their chemical environment.

### 2.3.2. The XPS Instrument at MODA Laboratory

The XPS system at MODA laboratory consists of several critical components:

- The X-ray source generates photons that trigger the photoelectric effect. At MODA, an aluminum anode is used, producing X-rays with approximately 1486.6 electron volts (eV) of energy.
- The sample stage holds the material being analyzed and can be moved or

rotated to examine different areas or angles.

- The electron energy analyzer separates ejected electrons according to their kinetic energy and counts how many electrons possess each specific energy level.
- The detector counts electrons and transmits this information to a computer system, which produces the final spectrum.

The entire system operates under ultra-high vacuum conditions, which are essential because the technique only analyzes the uppermost few nanometers of a sample. Any contamination from air molecules would compromise the results. Samples must be compatible with ultra-high vacuum conditions, which creates certain limitations. Liquids cannot be analyzed directly, and materials that release trapped gases are problematic. For the thin films studied at MODA, these constraints generally do not pose problems as they are solid and stable under vacuum.

### **2.3.3. XPS Measurement Process**

The XPS measurement process begins with precise sample preparation to avoid contamination, as even a fingerprint could introduce foreign elements that would appear in the analysis. Once prepared, the sample enters a preparation chamber first, where an initial vacuum forms. It is then transferred to the main analysis chamber where ultra-high vacuum conditions exist.

When measurement begins, the X-ray source is activated, and the sample is positioned correctly relative to the analyzer. The computer system controls all analyzer settings, determining which energies to measure and for how long. A typical analysis includes two types of scans:

1. A survey scan, which covers a wide range of binding energies (usually from 0 to 1100 eV) and offers an overview of all elements present in the sample.
2. Detailed scans, which focus on narrow energy ranges where particular elements show peaks. The exact shape and position of these peaks reveal information about chemical bonding.

A complete XPS analysis can take several hours, especially when detailed information about multiple elements is needed. Longer measurement times improve signal quality by collecting more photoelectrons.

### **2.3.4. Information Obtained from XPS Analysis**

XPS spectra contain rich information that requires both theoretical knowledge and practical experience to interpret properly. The analysis provides several types of information:

1. Element identification: Each element has characteristic binding energy peaks, making identification straightforward in principle.
2. Quantitative analysis: Peak intensity correlates with element concentration, though sensitivity factors must be applied to convert peak areas to actual atomic percentages.
3. Chemical state analysis: When atoms form chemical bonds, electron distribution changes, shifting core electron binding energies by small amounts. These "chemical shifts" reveal whether an element exists as a metal, oxide, nitride, or other chemical form [24].

For example, in oxide films, XPS can be used to study oxygen vacancies that affect electronic properties, making the material either more insulating or more conductive depending on their presence.

### **2.3.5. XPS Data Structure and Interpretation**

The raw data from XPS consists primarily of electron counts detected at various binding energies. These counts form the y-axis of the spectrum, while the binding energy forms the x-axis [25]. At MODA laboratory, the XPS system stores data in .vms files, a proprietary format from the instrument manufacturer that requires specialized software to open.

When accessed with the manufacturer's software, these files reveal several components:

- The raw counts measured by the detector, including background signal along with characteristic peaks
- Information about measurement conditions such as X-ray power, detector settings, and analysis area
- Calibration data that converts channel numbers in the analyzer to actual binding energies
- Records of any processing steps applied to the data, such as smoothing or background subtraction

Interpreting XPS data involves several steps:

1. Calibration of the binding energy scale using a reference peak of known energy (often carbon 1s from surface contamination) [26]
2. Background subtraction to remove contributions from electrons that lost energy through collisions before reaching the detector (commonly using the Shirley background method)
3. Peak fitting to separate overlapping peaks or identify multiple chemical states of the same element

Different approaches to data processing, particularly regarding peak fitting, can lead to slightly different interpretations of the same raw data. This highlights the importance of documenting exact processing methods.

## 2.3.6. Applications of XPS in Materials Science

XPS is applied to various research questions in materials science, including:

1. Composition analysis: Verifying that intended elements exist in correct proportions, which is fundamental for thin films grown at MODA. This analysis confirms whether the stoichiometry (element ratio) matches the target composition.
2. Contamination detection: Identifying even minute amounts of contaminants that can significantly affect material properties.
3. Interface analysis: By combining XPS with ion beam sputtering, which gradually removes surface layers, researchers create depth profiles showing composition changes from surface into material. This technique helps understand boundaries between different materials, which often determine functional properties of devices.
4. Surface reaction studies: By performing XPS before and after exposing materials to certain gases or liquids, researchers can identify chemical state changes indicating reactions have occurred. This approach provides insights into catalysis, corrosion, and other surface phenomena.

## 2.3.7. Challenges in XPS Data Management

Standardizing X-ray Photoemission Spectroscopy (XPS) data involves several technique-specific challenges that must be addressed to enable FAIR data practices:

One common issue is the lack of standardized file structures. While .vms files commonly produced by XPS instruments, can be opened with a text editor and are not strictly proprietary or binary, they often contain instrument-specific conventions, non-standard metadata formatting, and variable structures depending on the measurement settings used. This lack of uniformity complicates automated parsing and can hinder interoperability across different platforms or software tools. Without standardized schemas, integrating such files into shared databases or converting them into formats like NeXus remains a manual and error-prone task.

Another challenge arises from the diversity and complexity of XPS experiments. Different scientific objectives necessitate varying measurement approaches. For instance, some analyses target specific elemental peaks, while others perform broad energy range surveys. Angle-resolved XPS introduces angular dependence, effectively adding another dimension to the dataset, while depth profiling techniques generate a series of spectra corresponding to various material depths. These variations make it difficult to apply a one-size-fits-all

data structure.

Data provenance presents additional difficulties. The final appearance of an XPS spectrum is influenced not only by physical sample characteristics and measurement parameters but also by data processing decisions such as background subtraction, peak fitting, or normalization methods. Without detailed documentation of these steps, reproducibility and interpretability suffer. Proper metadata capturing this provenance is essential for meaningful reuse.

Finally, there are integration challenges. XPS data rarely stands alone in nanoscience research. To fully understand material properties, XPS results must often be linked with data from other techniques (e.g., structural analysis, transport measurements, or growth process monitoring). For example, the chemical composition information derived from XPS is more valuable when considered alongside thin film growth parameters or device performance data. Enabling such integration requires harmonized metadata frameworks and cross-disciplinary data models.

## **2.4. Data Management Practices and Developments at MODA**

At the MODA laboratory (CNR-SPIN, Naples), multiple efforts are underway to modernize data management workflows and align them with the FAIR principles. These initiatives form part of the laboratory's contribution to the NFFA-DI project and reflect a growing commitment to reproducibility, accessibility, and long-term preservation of research data.

A unified data ecosystem is being developed through the integration of three complementary projects:

- Standardization of XPS data using the NeXus file format, which is the focus of this thesis
- Development of a graphical interface for improved experiment entry in the eLabFTW electronic laboratory notebook (ELN)
- Creation of a NOMAD plugin and parser for Pulsed Laser Deposition (PLD) data

These projects are designed to work together across different data types and processes, enabling the consistent documentation, organization, and sharing of experimental results. By ensuring that distinct techniques such as XPS and PLD follow compatible standards, the laboratory is building a coherent foundation for FAIR data workflows. The resulting ecosystem supports traceability, facilitates interoperability with external platforms like NOMAD, and prepares the ground for integration with the distributed infrastructure of NFFA-DI.

### **2.4.1. Integration of eLabFTW for Digital Documentation**

To improve experiment documentation and metadata management, MODA has adopted eLabFTW, an open-source ELN [27]. The platform is designed to support the structured digital recording of experimental activities, protocols, and instrumentation settings. Although still in the process of being fully integrated into daily operations, eLabFTW provides a centralized and searchable environment for laboratory data.

To support its adoption, a graphical user interface is being developed to simplify data entry. This interface is intended to enhance usability by providing visual representations of experimental setups, allowing researchers to select parameters through intuitive interactions. The goal is to increase consistency in experiment records and reduce the time burden of documentation.

### **2.4.2. Data Storage, Formats, and Organization**

Data generated at MODA is currently stored on the local computers connected to each instrument. For XPS in particular, data is initially saved in .vms files, which are accessible via text editors. However, this file format is not standardized for interoperability or long-term sharing.

To address this, XPS data and metadata are being converted into the NeXus file format as part of this thesis project. NeXus is an open, hierarchical file format well suited for storing structured experimental data [28]. By aligning with NeXus application definitions, XPS data becomes both machine-readable and compatible with international standards, facilitating its reuse and integration with associated platforms.

In parallel, efforts are being made to improve the general organization of experimental data within the lab. This includes establishing consistent folder hierarchies, adopting standard naming conventions, and including key metadata through structured documentation practices. These measures aim to increase data traceability and ensure that information can be effectively retrieved and reused.

### **2.4.3. PLD Data Integration with NOMAD**

Another ongoing project focuses on standardizing data related to the Pulsed Laser Deposition (PLD) process. A dedicated plugin and parser are being

developed to extract PLD-related data and metadata from local records and convert them into a format compatible with the NOMAD repository. This allows the data to be transferred to the central NOMAD server in Trieste, where it can be made available for broader scientific use.

This process involves mapping experimental parameters; collected, for example, via eLabFTW to the structured schema required by NOMAD. The automation of these tasks supports standardization while minimizing the burden on researchers. Like the XPS standardization work, this initiative exemplifies how technique-specific workflows can be integrated into a larger ecosystem of FAIR-aligned research infrastructure.

## **2.5. NFFA-DI Integration at MODA**

The laboratory's participation in the NFFA-DI project provides both opportunities and requirements. The project creates external motivation for improving data management, provides frameworks and standards for implementation, and imposes constraints and deadlines.

### **2.5.1. Data Flow to NFFA-DI Central Services**

NFFA-DI aims to create a distributed data infrastructure connecting participating laboratories. Data collected at MODA would eventually flow to central services in Trieste, where it becomes part of a larger searchable repository. This data flow involves multiple steps:

1. Local collection and documentation in eLabFTW
2. Conversion to standardized formats
3. Local validation and quality control
4. Transfer to central repositories with appropriate metadata
5. Integration with search and access services

In reality, each transition requires solving technical and organizational challenges. Format conversions, metadata mapping, authentication systems, and security protocols all present potential complications.

### **2.5.2. Standards Compliance Requirements**

NFFA-DI has developed guidelines for data standardization across experimental techniques. For spectroscopic data like XPS, the recommendation is the NeXus file format with appropriate application definitions. These requirements influence the direction of data standardization projects, as converted data files

need to comply with both general FAIR principles and specific NFFA-DI standards for interoperability.

NFFA-DI provides guidelines for implementing standardized data formats. These resources are valuable for understanding requirements and best practices. However, translating general guidelines to specific laboratory contexts presents challenges, as standards necessarily generalize across different situations while laboratory environments have unique characteristics and constraints.

## **2.6. The Need for Standardized XPS Data Formats**

Standardizing XPS data formats would benefit the research community in several ways:

1. Scientific reproducibility improves when data is shared in standard formats with comprehensive metadata [29]. Other researchers can understand exactly how measurements were performed and what results mean, even without direct communication with the original experimenter.
2. Data reuse becomes possible when formats are non-proprietary and well-documented. Historical data can answer new research questions if it remains accessible and interpretable over time. Researchers occasionally need to revisit measurements from previous years but sometimes struggle with opening older files or understanding measurement conditions.
3. Automated analysis benefits from standardized formats. As machine learning and artificial intelligence techniques gain popularity in materials science, having large datasets in consistent formats becomes increasingly valuable. Standard formats enable the development of analysis tools that work across data from different instruments and laboratories.
4. Integration with materials databases requires standardized data structures. The NFFA-DI project aims to create merged nanoscience data databases, which become possible only if data from different facilities follows common standards. XPS data provides crucial material composition information for these databases. The NeXus format, based on HDF5, provides a flexible scientific data container while enforcing sufficient structure to ensure interoperability. The FAIRmat project has developed specific definitions for organizing XPS data within NeXus, which provide guidelines for implementation.

## **2.7. Challenges in the Laboratory Environment**

Realizing new data management practices in active research environments presents unique challenges. These challenges represent concrete barriers to implementation.

### **2.7.1. Balancing Research Productivity and Data Management**

MODA laboratory exists primarily for scientific research, not data management. Researchers receive evaluation based on scientific outputs such as publications, conference presentations, and new material discoveries. This creates natural tension between proper data documentation and conducting more experiments. Time spent organizing data means less time generating new findings.

Solutions that integrate seamlessly with existing practices are more likely to succeed. Adding separate documentation steps would likely face resistance. Data standardization could happen automatically during normal research activities; background processes could convert formats without researcher intervention, and metadata extraction could occur from existing documentation rather than requiring additional input.

### **2.7.2. Technical Infrastructure Limitations**

Like many research facilities, MODA operates with limited IT resources. The laboratory has specialized equipment for experiments, but the computing infrastructure remains relatively basic. Solutions need to accommodate several technical limitations:

- Limited server space
- No dedicated database administrator
- Basic network infrastructure
- Reliance on open-source solutions due to budget constraints

These factors push toward lightweight implementations that can run on available hardware. The design needs to function without requiring specialized IT training for maintenance, as the laboratory staff lacks time for complex system administration.

### **2.7.3. Knowledge Transfer and Training**

Another significant challenge involves ensuring continuity after project implementation. Implementing technical solutions provides limited value without accompanying knowledge transfer. Laboratory members need training to use new tools effectively, and documentation needs to explain processes clearly.

Simple, well-documented solutions are needed. Basic guides for processes such as using data conversion tools, following standardized naming conventions, adding appropriate metadata to experiments, and finding and accessing previously converted data could potentially support continued use of improved practices.

## **2.8. Anticipating Future Developments at MODA**

Looking ahead, the MODA laboratory is preparing for several important developments that will significantly impact its data management landscape. Practically addressing these upcoming changes ensures that the systems being developed today will remain flexible and extensible, capable of supporting not only current techniques like XPS and PLD, but also future experimental capabilities.

### **2.8.1. Expansion of Instrumentation**

MODA is planning the installation of a new X-ray Diffraction (XRD) system, which will introduce new types of experimental data requiring structured storage and metadata documentation. While the current efforts such as the NeXus-based XPS data standardization detailed in this thesis focus on specific instruments, they also serve as prototypes for future extensions.

To accommodate new instruments and experimental methods, software solutions must be designed with modularity and reusability in mind. This includes creating shared libraries for common tasks like metadata handling, timestamping, and file conversion. A well-documented and configurable codebase would allow future researchers to extend existing tools to handle additional techniques, minimizing duplication of effort and facilitating long-term sustainability.

The integration of future equipment beyond XRD - each with potentially unique data formats and metadata requirements - underscores the need for flexible and

evolving FAIR-compliant solutions. Building with change in mind is essential in a dynamic research environment.

## **2.8.2. Managing Growing Data Volumes**

As scientific investigations grow in complexity and scale, the amount of data generated will be increased as well. Modern materials science experiments often involve high-resolution measurements, larger sample sets, and multiple analysis techniques applied in combination. This results in exponential growth in data volume, making manual data handling increasingly impractical.

For example, systematic studies involving multiple thin film compositions analyzed by techniques such as XPS and AFM can produce hundreds of interrelated files per project. Organizing and interpreting these files without automation is both time-consuming and error-prone.

In response, future data workflows at MODA must prioritize batch processing capabilities, automated metadata extraction, efficient, standardized storage formats (e.g., NeXus), and integrated tracking of relationships between samples and measurements.

These enhancements will allow the laboratory to scale its data management practices alongside its research ambitions. They also reinforce the central FAIR principle that data should be reusable, not just by humans, but also by machines.

## **2.9. Research Focus and Infrastructure at MODA**

A closer look at MODA's scientific output and infrastructure reveals a highly specialized environment at the forefront of oxide thin film research. The lab focuses on interface engineering at the atomic scale, using sophisticated fabrication techniques such as Pulsed Laser Deposition (PLD) and advanced characterization tools like X-ray Photoemission Spectroscopy (XPS) and soon, X-ray Diffraction (XRD).

MODA's research has contributed to the understanding of emergent phenomena such as two-dimensional electron gases (2DEGs) at oxide interfaces materials with significant potential for future electronic devices. This scientific impact is evident in publications in high-impact journals and active collaborations with international partners.

The laboratory's role within the NFFA-DI project reflects recognition of its scientific and technical excellence. By contributing to a federated European research infrastructure, MODA supports open, interoperable science at a continental scale.

MODA also plays a strong educational role, training doctoral students and postdoctoral researchers from Italy and abroad. While this brings new perspectives, it also increases the need for consistent documentation practices and intuitive digital tools that can support both experienced scientists and newcomers.

## 2.10. Conclusion

The MODA laboratory demonstrates a modern materials science facility navigating the transition to FAIR-aligned data practices. Balancing ambitious local research with participation in broader initiatives like NFFA-DI, the laboratory faces the challenge of standardizing data across diverse instruments, proprietary formats, and evolving user needs.

The collaborative structure of data management efforts at MODA spanning XPS standardization, PLD plugin development, and ELN integration, illustrates how incremental, technique-specific improvements can contribute to a larger transformation. No single project can resolve all data challenges, but together, these initiatives form a cohesive foundation for long-term interoperability and reuse.

This thesis has contributed to this broader vision by focusing on the conversion of XPS data from the .vms format into standardized NeXus files, in accordance with FAIRmat guidelines. By capturing both the raw data and associated metadata in a machine-readable, extensible format, this work enhances the findability, accessibility, interoperability, and reusability of XPS data produced at MODA.

While the initial focus is XPS, the tools and practices developed here provide a flexible model for other techniques and future instruments. As MODA expands its capabilities and research scope, these data infrastructure improvements will support faster, more reliable, and more collaborative science.

# **Chapter 3**

## **Implementation of FAIR Principles for XPS Data**

### **3.1. Introduction**

In previous chapters, we explored the basics of Open Science, FAIR principles, and X-ray Photoemission Spectroscopy (XPS). This chapter focuses on the practical implementation of FAIR principles specifically for XPS data collected in the MODA laboratory. As discussed earlier, making data FAIR (Findable, Accessible, Interoperable, and Reusable) is essential for modern scientific research. However, implementation requires specific tools and methodologies tailored to particular scientific domains.

The challenge for XPS data lies in converting proprietary file formats from commercial instruments into standardized, machine-readable formats that adhere to FAIR principles. This chapter presents the methodology and tools developed during this thesis project to address this challenge.

Converting proprietary data formats isn't easy. Many scientific instruments output data in formats that only their software can read. This creates barriers to data sharing and reuse. The solution developed transforms these closed formats into open ones, although the implementation process faced several challenges.

During the limited time at the MODA laboratory, it was observed that researchers struggled with data management issues. Files were stored in proprietary formats. Metadata was sometimes incomplete or scattered across various documents.

### **3.2. Overview of the Conversion Pipeline**

To implement FAIR principles for XPS data, a conversion pipeline was developed

that transforms data files from the laboratory's XPS instrument into standardized NeXus files. NeXus is an internationally recognized data format based on HDF5, specifically designed for scientific data storage.

The conversion pipeline consists of three main components:

First, the input data sources include proprietary .vms files from the XPS instrument and experimental metadata stored in ELN YAML files. The .vms files contain the raw spectroscopic data and some basic instrument settings, while the YAML files provide additional contextual information necessary for full FAIR compliance.

Second, the conversion tools comprise a Python script utilizing the pynxtools library with its XPS plugin. This script acts as a user-friendly interface to the underlying conversion process, guiding users through file selection and metadata collection.

Finally, the output data consists of standardized NeXus files (.nxs) following the NXxps application definition. These files contain both the experimental data and comprehensive metadata in a structured format that adheres to FAIR principles.

The basic flow starts with raw data acquisition on the XPS instrument, which generates .vms files. These files, along with metadata either from existing YAML files or created through interactive prompts, are processed by the converter script. The script then produces standardized NeXus files that can be shared, archived, or further analyzed using standard tools.

## **3.3. Components of the Data Conversion Pipeline**

### **3.3.1. Input Data Sources**

The conversion process starts with two primary data sources. The first source consists of VMS files (.vms), which are proprietary format files generated by the Scienta Omicron XPS instrument in the MODA laboratory. These files contain the raw spectroscopic data and some metadata about the experiment, such as energy ranges, count times, and basic instrument settings. However, they often lack comprehensive contextual information necessary for full FAIR compliance.

The second source comprises Electronic Laboratory Notebook (ELN) files (.yaml), which contain additional metadata about the experiment that might not be included in the VMS files. This includes information about the sample, instrument settings, experimenter details, and experimental conditions. These YAML files can either be created before the experiment and reused for multiple measurements or generated through the converter script's interactive prompts. The combination of these two data sources ensures that both the actual experimental data and its context are preserved in the conversion process.

### 3.3.2. Configuration Files

A crucial component of the conversion pipeline is the configuration file (MODA\_config.json). This JSON file maps the information from VMS and ELN files to the appropriate fields in the NeXus format according to the NXxps application definition. The configuration file is extensively structured to ensure proper organization of data in the output file.

The configuration includes data field mappings between source files and NeXus structure, specifying which fields from the input sources should populate each NeXus field. It also contains unit specifications for physical quantities, ensuring that all measurements have appropriate units associated with them. Furthermore, coordinate system transformations are defined to maintain spatial relationships in the data. The structural organization of the output NeXus file is also determined by the configuration, ensuring consistency across converted files.

The configuration file plays a vital role in ensuring that the converted data follows the standardized structure required for interoperability. It essentially acts as a translation layer between the proprietary data format and the standardized NeXus format, handling the complexities of data reorganization and enrichment with additional contextual information.

The creation of the configuration file required careful consideration of the NXxps standard and the specifics of the XPS data.

### 3.3.3. Conversion Software

For the conversion process, a Python script was developed that leverages the pynxtools library with its XPS plugin. The pynxtools library is a Python package designed to facilitate the conversion of scientific data into NeXus format. The XPS plugin extends this functionality specifically for XPS data, providing specialized readers and writers for this type of spectroscopic data.

The Python script performs several key functions:

1. It checks for required dependencies and files, ensuring that the environment is properly set up for conversion.
2. It allows users to select the input VMS file for conversion from those available in the current directory.
3. It enables users to select an existing ELN metadata file or create a new one through guided prompts that collect essential information about the experiment.
4. Once the input files are selected, the script handles the conversion execution by calling the pynxtools dataconverter with appropriate parameters.
5. Finally, it manages the output by generating a uniquely named output file and providing conversion status information to the user.

The script is designed to be user-friendly, with clear prompts and error handling to guide through the conversion process. This is particularly important in a laboratory environment where researchers may have varied technical backgrounds.

## 3.4. Implementation Details

### 3.4.1. Python Converter Script

The Python script developed serves as a user-friendly interface for the conversion process. It guides users through the necessary steps and handles the complexity of the conversion behind the scenes. The script begins by checking for required dependencies and files, ensuring that the environment is properly set up for conversion.

*Code snippet 1:*

```
#  Pre-check dependencies and required files
REQUIRED_FILES = ["MODA_config.json"]
REQUIRED_PACKAGES = ["pynxtools", "yaml"]

for req in REQUIRED_FILES:
    if not os.path.isfile(req):
        print(f"\033[91m Required file '{req}' is missing. Please
add it to this folder.\033[0m")
        sys.exit(1)
```

This initial block “Code snippet 1” checks if the necessary configuration file is present. Without this file, the conversion cannot proceed, so the script exits with an error message if it's missing.

This block that is “Code snippet 2” checks for required Python packages. If any are missing, the script provides installation instructions and exits, preventing errors due to missing dependencies.

*Code snippet 2:*

```
# ✅ Check for required packages
REQUIRED_MODULES = {
    "pynxtools": "pip install git+https://github.com/FAIRmat-
NFDI/pynxtools-xps.git@main",
    "yaml": "pip install pyyaml",
    "pandas": "pip install pandas",
    "numpy": "pip install numpy",
    "matplotlib": "pip install matplotlib",
    "scipy": "pip install scipy"
}

missing = []

for module, install_cmd in REQUIRED_MODULES.items():
    try:
        importlib.import_module(module)
    except ImportError:
        missing.append((module, install_cmd))

if missing:
    print("\n❌ Missing required packages:\n")
    for module, install_cmd in missing:
        print(f" - {module}: {install_cmd}")
    print("\n👉 Please install the missing packages above and re-run
the script.")
    sys.exit(1)
```

The file selection functions (Code snippet 3) utilize Python's glob module to find files with specific extensions in the current directory. They present these files to the user and allow selection through numbered options. This approach makes the script accessible even to users with limited command-line experience.

This function scans the current directory for .vms files and presents them as numbered options. It validates user input to ensure a valid selection is made.

The script also includes error handling to deal with common issues such as missing files or invalid selections. This robustness is essential for practical use in a laboratory setting where unexpected situations may arise.

Code snippet 3:

```
# Function to choose a .vms file
def choose_vms_file():
    vms_files = glob.glob("*.vms")
    if not vms_files:
        print("\033[91m❌ No .vms files found in this
directory.\033[0m")
        sys.exit(1)
    print("\n\033[94m📁 Available .vms files:\033[0m")
    for i, f in enumerate(vms_files):
        print(f"[{i+1}] {f}")
    while True:
        choice = input("\nEnter the number of the .vms file to use:
").strip()
        if choice.isdigit() and 1 <= int(choice) <= len(vms_files):
            return vms_files[int(choice) - 1]
        print("\033[91mInvalid choice. Please select a valid number
from the list.\033[0m")
```

## 3.4.2. Metadata Collection

A critical aspect of implementing FAIR principles is ensuring comprehensive metadata. To facilitate this, functionality was implemented to create new ELN metadata files when needed. The `create_eln_yaml` function in the script provides an interactive way to collect this metadata (Code snippet 4).

Code snippet 4:

```
# Create a new ELN YAML file
def create_eln_yaml():
    print("\n\033[92m📄 Creating ELN metadata for the XPS
experiment...\033[0m")

    source_probe_types = [
        'Synchrotron X-ray Source', 'Rotating Anode X-ray', 'Fixed Tube
X-ray', 'UV Laser',
        'Free-Electron Laser', 'Optical Laser', 'UV Plasma Source',
        'Metal Jet X-ray',
        'HHG laser', 'UV lamp', 'Monochromatized electron source']
    collectioncolumn_schemes = [
        'angular dispersive', 'spatial dispersive', 'momentum
dispersive', 'non-dispersive']
    energydispersion_schemes = [
        'tof', 'hemispherical', 'double hemispherical', 'cylindrical
mirror',
        'display mirror', 'retarding grid']
```

```

el_n = {
    'title': ask("🔥 Experiment title", "XPS Experiment"),
    'user': {
        'name': ask("👤 User name"),
        'affiliation': ask("🏠 Affiliation")
    },
    'instrument': {
        'vendor': ask("💡 Instrument vendor", "Scienta Omicron"),
        'model': ask("🔧 Instrument model", "OMICRON model"),
        'source_probe': {
            'type': ask_from_list("💡 Source probe type",
source_probe_types, "Fixed Tube X-ray")
        },
        'electronanalyzer': {
            'collectioncolumn': {
                'scheme': ask_from_list("🔍 Collection column
scheme", collectioncolumn_schemes, "angular dispersive")
            },
            'energydispersion': {
                'scheme': ask_from_list("🔍 Energy dispersion
scheme", energydispersion_schemes, "hemispherical")
            }
        }
    },
    'sample': {
        'name': ask("📄 Sample name", "SPIN-Napoli-sample-name"),
        'identifier': ask("📄 Sample identifier", "Napoli-sample-
identifier")
    }
}

timestamp = datetime.now().strftime("%Y%m%d_%H%M%S")
el_n_filename = f"el_n_{timestamp}.yaml"
with open(el_n_filename, 'w') as f:
    yaml.dump(el_n, f, sort_keys=False)
print(f"\n\033[92m✅ ELN file saved as: {el_n_filename}\033[0m\n")
return el_n_filename

```

The function guides users through a series of prompts for essential metadata, including information about the experiment, user, instrument, and sample. For each prompt, it provides clear descriptions and default values where appropriate. This makes the process less intimidating for users who might not be familiar with all the required metadata fields.

The collected information is structured according to the requirements of the NeXus format and saved in a YAML file with a timestamp-based filename. This approach ensures that metadata files can be easily associated with the experiments they describe and can be reused for multiple conversions if appropriate.

The metadata collection process is designed to be comprehensive yet not overly burdensome. It focuses on collecting information that is essential for FAIR

compliance while not overwhelming users with excessive details. The balance between completeness and usability is crucial for practical adoption of the tool in a research environment.

The selection of metadata fields was based on the NXxps standard recommendations, focusing on essential information that would make the data truly reusable. Default values are provided for fields that are unlikely to change between experiments, such as instrument vendor and analyzer type, to streamline the process.

### 3.4.3. Configuration File Structure

The configuration file (MODA\_config.json) is structured according to the NeXus hierarchy, with mappings to source data fields. This hierarchical structure mirrors the organization of the output NeXus file, making it easier to understand and maintain.

The configuration uses a path-based approach, where each entry represents a path in the NeXus file hierarchy. For example, `"/ENTRY"` represents the top-level entry in the NeXus file, while `"/ENTRY/INSTRUMENT[instrument]"` represents the instrument section within that entry. Each path is associated with a set of mappings that specify how to populate the corresponding NeXus fields.

The mapping syntax uses special prefixes to indicate data sources. The `"@eln"` prefix indicates that the data should be taken from the ELN YAML file. The `"@attrs"` prefix indicates that the data should be extracted from attributes in the VMS file. The `"@data"` prefix is used for data arrays from the VMS file. This flexible mapping system allows for complex transformations between the input and output formats.

```
"/ENTRY":{
  "@default":"data",
  "title":"@eln",
  "start_time":["@attrs:time_stamp', '@eln']",
  "end_time":"@eln",
  "method":["@attrs:analysis_method', '@eln']",
  "transitions":"@attrs:transitions",
  "program_name":"@eln",
  "experiment_institution":"@attrs:institute_id",
  "identifier_experiment":["@attrs:experiment_id', '@eln']"
}
```

This selection from the configuration file shows how various fields in the NeXus file are mapped to data sources. For example, `"title"` is taken from the ELN file, while `"start_time"` has two possible sources: first, it tries to use the `"time_stamp"` attribute from the VMS file, and if that's not available, it uses the value from the ELN file.

The configuration also includes special directives for handling units, defaults, and

transformations. For example, the `"/@units"` suffix is used to specify the units for a particular field. The `"!"` prefix is used to indicate that a value should be negated, which is useful for handling coordinate system differences between the input and output formats.

```
"AXISNAME[y_rotation]":"!@attrs:source_polar_angle",  
"AXISNAME[y_rotation]/@units":"@attrs:source_polar_angle/@units",  
"AXISNAME[y_rotation]/@transformation_type":"rotation",  
"AXISNAME[y_rotation]/@vector":[0, -1, 0],  
"AXISNAME[y_rotation]/@depends_on":"."
```

This passage shows a coordinate transformation definition. The transformation is defined as a rotation around the `[0, -1, 0]` vector, with additional metadata specifying the transformation type and dependencies.

This structured mapping ensures that all relevant information is correctly placed in the standardized NeXus format. The configuration file essentially encodes the domain knowledge required to transform XPS data from its proprietary format to a standardized, FAIR-compliant format.

## 3.5. FAIR Principles Implementation

Let's examine how this conversion pipeline implements each of the FAIR principles for XPS data.

### 3.5.1. Findability

The conversion to NeXus format enhances findability in several ways. First, the NXxps application definition provides a well-defined structure that makes it easier to locate specific data elements. This standardized structure means that users familiar with the NeXus format can quickly navigate to the information they need, regardless of the original format of the data.

Second, the ELN YAML files and the conversion process ensure comprehensive metadata, including experiment title, user information, sample details, and instrument parameters. This rich metadata makes it possible to search for specific experiments based on various criteria, enhancing the discoverability of the data.

Third, the conversion script generates timestamped output filenames to ensure uniqueness, and sample identifiers are included in the metadata. These identifiers can be used to reference the data in publications or databases, further enhancing its findability.

*Code snippet 5:*

```
# Generate unique output filename
def get_output_filename(input_file):
    base = os.path.splitext(input_file)[0]
    timestamp = datetime.now().strftime("%Y%m%d_%H%M%S")
    output = f"{base}_converted_{timestamp}.nxs"
    count = 1
    while os.path.exists(output):
        output = f"{base}_converted_{timestamp}_{count}.nxs"
        count += 1
    return output
```

This function generates unique filenames based on the input filename and current timestamp, preventing accidental overwriting of files and creating a clear link between the original data and the converted file (Code snippet 5).

The structured nature of the NeXus format also allows for the inclusion of additional identifiers or links to external resources, such as publication DOIs or database entries. This enhances the discoverability of the data within the wider scientific ecosystem.

With such improved metadata and standardized structure, it becomes possible to search for XPS data and experiments by sample name, user, experiment date, or other criteria, significantly improving findability.

## 3.5.2. Accessibility

The NeXus format improves accessibility of XPS data through several mechanisms. As an open and well-documented format based on HDF5, NeXus files can be accessed through numerous software tools across different platforms. This eliminates the need for software that might be expensive or limited to specific operating systems, making the data more widely accessible.

The self-contained nature of NeXus files, which include all necessary metadata within the file, reduces dependence on external documentation. This means that even if the data is separated from its original context, such as laboratory notebooks or publications, it can still be fully understood and used. This is particularly important for long-term preservation of data, where contextual information might otherwise be lost.

The compatibility of NeXus files with standard HDF5 viewers and specialized scientific software further enhances accessibility. Users can choose from a wide range of tools to access and analyze the data, depending on their specific needs and expertise. This flexibility is crucial for promoting the use of the data across different disciplines and research groups.

The software for XPS data might not be universally available. By converting to an open format, the data becomes accessible to a much wider audience using free tools

such as HDFView. This homogenize access to the data and facilitates collaboration with researchers who might not have access to the specialized software.

### 3.5.3. Interoperability

Interoperability is enhanced through the adoption of standardized formats and vocabularies. By following the NXxps application definition, the data structure conforms to community standards, making it compatible with other tools and datasets that adhere to the same standards. This standardization facilitates the integration of XPS data with data from other techniques or sources.

The configuration file explicitly specifies units for physical quantities, ensuring correct interpretation of the data regardless of the software used to access it. This is crucial for scientific data, where misinterpretation of units can lead to significant errors in analysis and interpretation.

```
"incident_energy":["@attrs:excitation_energy',  
'@attrs:characteristic_energy']",  
"incident_energy/@units":["@attrs:excitation_energy/@units',  
'@attrs:characteristic_energy/@units', 'eV']",
```

This excerpt from the configuration file shows how units are explicitly specified for the incident energy, with fallback options if the primary source is not available.

The detailed coordinate system information included in the configuration allows proper spatial interpretation of the data. This is particularly important for XPS data, where the orientation of the sample relative to the X-ray source and analyzer can significantly affect the results. By standardizing the coordinate system, the data can be more easily compared with other experiments or theoretical models.

The use of standardized terms and relationships between data elements, as defined in the NXxps application definition, further enhances interoperability. This shared vocabulary ensures that different researchers and software tools can understand and process the data in a consistent manner.

While the NXxps standard may not be perfect for all use cases, adopting a standard approach is inherently better than custom solutions, as it creates a common language for data exchange. This standardization allows for automated processing of files without human intervention, significantly enhancing interoperability.

### 3.5.4. Reusability

The conversion pipeline promotes reusability by including detailed information about experimental conditions, instrument settings, and sample properties. This comprehensive metadata allows other researchers to understand the context in which the data was collected and to evaluate its relevance for their own research

questions.

The recording of information about the original data source and the conversion process provides important background information. This transparency is essential for scientific reproducibility, as it allows others to trace back the origin of the data and understand any transformations it has undergone.

The opportunity to include licensing information in the metadata clarifies the terms under which the data can be reused. This legal clarity is important for facilitating data sharing and collaboration, as it removes uncertainty about what users are allowed to do with the data.

The preservation of error estimates and other quality indicators from the original data enhances its reusability by allowing users to assess its reliability and suitability for specific analyses. This quality information is crucial for scientific data, where understanding the limitations and uncertainties of the data is as important as the data itself.

Reusability has practical benefits in a research environment. For example, when new researchers join the laboratory, they can quickly access and understand existing data without extensive training. When comparing results across experiments, the standardized format makes it easier to identify relevant parameters and experimental conditions, facilitating data reuse.

## 3.6. Usage Example

To illustrate the practical application of the conversion pipeline, let's walk through a typical usage scenario. The process begins when a researcher has completed an XPS experiment and has one or more .vms files containing the data. The researcher then runs the Python converter script in the directory containing these files.

Upon starting the script, the user is presented with a list of available .vms files in the current directory. For example, the script might display:

```
Available .vms files:
[1] sample1_survey.vms
[2] sample1_C1s.vms
[3] sample2_survey.vms
Enter the number of the .vms file to use: 2
```

The user selects the file they want to convert by entering the corresponding number. In this case, the user has selected "sample1\_C1s.vms", which contains a high-resolution scan of the carbon 1s peak.

Next, the script asks the user to select an existing ELN metadata file or create a new one. If the user has already created a metadata file for this sample or experiment, they can reuse it. Otherwise, they can choose to create a new one by answering a series of prompts:

```

📄 Available YAML metadata files:
[1] eln_20250410_143022.yaml
[2] eln_20250410_150147.yaml
[N] + Create a new ELN YAML file
Choose a file by number or 'N' to create new: N

✏️ Creating ELN metadata for the XPS experiment...
🚩 Experiment title [default: XPS Experiment]: C1s Core Level Study
👤 User name: John Smith
🏢 Affiliation: MODA Laboratory

```

The user continues to answer prompts about the instrument, source, and sample details. These prompts are designed to collect all the essential metadata required for FAIR compliance. Once the metadata collection is complete, the script saves it to a new YAML file with a timestamp-based filename.

With both the data file and metadata file selected, the script proceeds with the conversion. It constructs and executes the dataconverter command with appropriate parameters:

```

⚙️ Running conversion...
dataconverter sample1_C1s.vms eln_20250410_152733.yaml -c
MODA_config.json --nxd1 NXxps --reader xps --ignore-undocumented --
output sample1_C1s_converted_20250410_152733.nxs

```

The script then displays a summary of the conversion, including the input files used, the configuration file, and the output file:

```

📄 Summary:
- Input VMS file: sample1_C1s.vms
- ELN metadata : eln_20250410_152733.yaml
- Config file  : MODA_config.json
- Output NeXus : sample1_C1s_converted_20250410_152733.nxs
🎉 Conversion successful!

```

The result of this process is a standardized NeXus file that adheres to FAIR principles and can be easily shared, archived, or analyzed with standard tools. The user can now proceed to analyze the data using the software of their choice, or share it with collaborators or repositories with confidence that it includes all the necessary context for proper interpretation.

The script is designed to be user-friendly, even for those with limited programming experience. Clear prompts and error handling guide the user through the process,

minimizing the technical knowledge required. The ability to reuse metadata files for multiple scans of the same sample also streamlines the workflow for typical laboratory scenarios.

## 3.7. Challenges and Solutions

During the implementation of this conversion pipeline, several challenges were encountered. One significant challenge was related to the limitations of the .vms format, which does not include all necessary metadata for comprehensive FAIR data. This is a common issue with such formats, which are often designed for specific software applications rather than for data sharing and preservation.

To address this challenge, supplementary ELN YAML files were implemented to capture additional metadata through user input. This approach allows for the collection of important contextual information that might not be recorded by the instrument software. The interactive prompts in the script guide users through the process of providing this additional metadata, ensuring that it is comprehensive and consistent.

Another challenge involved the standardization of coordinate systems for XPS data. XPS experiments involve complex spatial relationships between the sample, X-ray source, and analyzer, which must be properly represented for the data to be correctly interpreted. Different instruments and software may use different conventions for these coordinate systems, making standardization difficult.

To solve this issue, detailed coordinate system transformations were defined in the configuration file. These transformations ensure that the spatial relationships in the data are properly represented in the NeXus format, regardless of the conventions used in the original data. This standardization is crucial for interoperability, as it allows data from different instruments to be compared and combined.

User experience was also a significant consideration, particularly given that some laboratory users might have limited programming experience. To make the tool accessible to a wide range of users, a user-friendly script with clear prompts and error handling was developed. The script guides users through the conversion process step by step, minimizing the technical knowledge required to use it. Error messages are designed to be informative and helpful, directing users toward solutions rather than simply reporting problems.

Ensuring that the converted data accurately represents the original measurements was another important challenge. Data validation is essential for maintaining the scientific integrity of the data through the conversion process. To address this, the configuration file mappings were carefully tested with various input files to verify the correctness of the conversion. The structure of the configuration file allows for precise control over how each field is mapped, ensuring that the data is transformed correctly.

The .vms format presented particular challenges due to its proprietary nature. Inconsistencies between files from the same instrument, such as slightly different field names or units, required flexible mappings in the configuration file to handle these variations. Fallback mechanisms were implemented to handle missing or unusual values, improving the robustness of the conversion process.

## 3.8. Future Improvements and Integration with Laboratory Workflows

While the current implementation successfully converts XPS data to a FAIR-compliant format, several improvements and integrations are planned for the future. A key development will be the integration of this conversion pipeline with the broader laboratory data management system at MODA.

The laboratory plans to implement eLabFTW, an electronic laboratory notebook system, to manage various types of experimental data, including those from the XPS instrument. This integration will allow for more automated and streamlined data management workflows. Instead of manually creating ELN YAML files through the converter script's prompts, the metadata will be directly exported from the eLabFTW database.

This planned pipeline will work as follows:

1. Researchers will enter experimental metadata into the eLabFTW system before or during experiments.
2. When XPS data needs to be converted, the system will automatically generate the necessary ELN YAML files from the eLabFTW database.
3. These files will then be used by the conversion script, potentially in an automated process, to create standardized NeXus files.

This integration will reduce the manual effort required for data conversion and ensure more consistent metadata collection. It will also create a more cohesive data management ecosystem within the laboratory, linking experimental records in the electronic notebook directly to the data files.

Other potential improvements to the current implementation include:

A web-based interface for the conversion tool would eliminate the need for command-line interaction, making it even more accessible to users with limited technical backgrounds. Such an interface could provide a more visual and intuitive way to select files and provide metadata, potentially increasing adoption among researchers.

Adding functionality to directly upload converted files to institutional or domain-specific repositories would streamline the process of data sharing and publication. This would make it easier for researchers to fulfill open data requirements imposed by journals or funding agencies.

More sophisticated validation of the converted data against the NXxps standard would help ensure compliance and data quality. This could involve checking for required fields, validating value ranges, and verifying relationships between different data elements.

Enhancing the script to handle batch conversion of multiple files with the same metadata would be beneficial for high-throughput experiments. Many XPS experiments involve multiple scans of different regions or samples under similar

conditions. Being able to apply the same metadata to multiple files would save time and ensure consistency across the dataset.

## 3.9. Conclusion

The implementation described in this chapter successfully addresses the challenge of making XPS data FAIR-compliant. By converting proprietary format files to standardized NeXus files with comprehensive metadata, the data becomes findable, accessible, interoperable, and reusable.

The Python-based conversion pipeline provides a user-friendly interface that guides researchers through the conversion process, ensuring that all necessary metadata is captured. The configuration file structure allows for flexible mapping between input and output formats while maintaining compliance with the NXxps application definition.

This implementation represents a significant step forward in applying FAIR principles to experimental data in the MODA laboratory environment, contributing to more open and reproducible scientific research in the field of X-ray photoemission spectroscopy. The planned integration with the eLabFTW system will further enhance the FAIR compliance of XPS data by streamlining metadata collection and ensuring consistency across datasets.

While the current solution has some limitations, it provides a solid foundation for improving data management practices in the laboratory. The tools developed are designed with practical considerations in mind, balancing the need for comprehensive metadata with the reality of laboratory workflows and researcher familiarity with data management concepts.

The long-term success of this implementation will depend on researcher adoption and continued refinement of the tools based on user feedback. However, the groundwork has been laid for a more FAIR-compliant approach to XPS data management at the MODA laboratory.

# Chapter 4

## Conclusions and Future Work

### 4.1. Summary of Achievements

This thesis has explored the implementation of FAIR principles in the context of X-ray Photoemission Spectroscopy (XPS) data collected at the MODA laboratory. The journey began with an examination of Open Science and FAIR principles, establishing the theoretical foundation for making scientific data more Findable, Accessible, Interoperable, and Reusable. We then explored the MODA laboratory environment and the specific challenges it presents for data management. A detailed discussion of X-ray Photoemission Spectroscopy provided the necessary technical background to understand the data being managed. Finally, we described the implementation of a conversion pipeline that transforms proprietary XPS data into standardized, FAIR-compliant NeXus files.

The primary achievement of this work has been the development of a practical solution for implementing FAIR principles in a real laboratory environment. The conversion pipeline successfully addresses the challenges posed by proprietary data formats and incomplete metadata. Sometimes labs just generate data with no plan for future use. By leveraging the `pynxtools` library with its XPS plugin and designing a user-friendly interface, we have created a tool that is accessible to researchers with varying levels of technical expertise. The configuration file structure provides the flexibility needed to accommodate the specific characteristics of the MODA laboratory's XPS instrument while ensuring compliance with community standards.

Another significant achievement has been the integration of metadata collection into the conversion process. The Electronic Laboratory Notebook (ELN) functionality allows researchers to capture essential contextual information about their experiments, enhancing the findability and reusability of the data. The interactive prompts guide users through this process, ensuring that all necessary metadata is collected without overwhelming them with excessive requirements. This balanced approach is crucial for practical adoption of FAIR principles in busy research environments where time is often limited and researchers are focused primarily on

data collection rather than documentation.

The work presented in this thesis demonstrates that FAIR principles can be implemented in specialized scientific domains with appropriate tools and methodologies. The conversion pipeline serves as a concrete example of how theoretical principles can be translated into practical solutions that enhance the value and impact of scientific data. The solution provides a bridge between the ideals of open science and the practical realities of laboratory research, showing that meaningful progress toward FAIR data management is achievable with targeted interventions.

Moreover, the implementation addresses the challenges of legacy data, allowing researchers to apply FAIR principles not only to newly collected data but also to existing datasets that might otherwise remain locked in proprietary formats. This retrospective application of FAIR principles maximizes the value of historical data and ensures that it can continue to contribute to scientific progress. The MODA lab has accumulated XPS data for years. Now it's all usable.

## 4.2. Impact on Research Practices

The implementation of FAIR principles for XPS data has several important implications for research practices at the MODA laboratory and beyond. Perhaps the most immediate impact is on data sharing and collaboration. By converting proprietary data to a standardized format with comprehensive metadata, researchers can more easily share their results with collaborators, regardless of the software or systems they use. This removes technical barriers to collaboration and broadens the potential network of research partners.

Long-term data preservation is another area significantly impacted by this work. Scientific data represents a substantial investment of time, resources, and expertise, yet much of it becomes inaccessible or unusable over time due to obsolete formats or missing contextual information. The conversion of XPS data to the standardized NeXus format with embedded metadata ensures that this valuable resource remains accessible and interpretable for years to come. This contributes to the preservation of scientific knowledge and protects the investment made in data acquisition.

The implementation also enhances research reproducibility, a cornerstone of scientific integrity. With comprehensive metadata and standardized data formats, other researchers can better understand and potentially reproduce the experimental conditions under which the data was collected. This transparency builds confidence in scientific results and facilitates the verification and extension of research findings. I've seen too many papers without proper methods. Data often sits forgotten on lab computers. Not anymore.

Educational opportunities are another positive outcome of this work. The conversion pipeline and its documentation serve as practical examples of FAIR data management for students and early-career researchers. By participating in the data

conversion process, they gain hands-on experience with data management practices that will be increasingly important throughout their careers. This educational aspect contributes to building a culture of open and responsible data management in the scientific community.

Furthermore, the implementation of FAIR principles can enhance the visibility and impact of research conducted at the MODA laboratory. Data that is more findable and accessible has a greater chance of being cited and reused in future studies, increasing the research impact of the original work. This aligns with institutional goals of maximizing research impact and can help justify the investment in data collection and instrumentation.

The standardized approach to data management also creates new opportunities for meta-analysis and data mining across multiple experiments or research groups. By having XPS data in a consistent format with reliable metadata, researchers can more easily identify patterns, trends, or correlations that might not be apparent when examining individual datasets. This potential for new insights represents a significant added value of FAIR data practices.

Additionally, the conversion pipeline can serve as a model for similar initiatives in other spectroscopic techniques or research domains. The approach developed for XPS data could be adapted to related techniques, contributing to a broader ecosystem of interoperable scientific data. This potential for expansion multiplies the impact of the current implementation beyond its immediate application to XPS data at the MODA laboratory.

## **4.3. Limitations of the Current Implementation**

Despite its achievements, the current implementation has several limitations that should be acknowledged. One significant limitation is the dependency on the specific structure of the .vms files produced by the MODA laboratory's XPS instrument. While the configuration file provides flexibility in mapping data fields, substantial changes in the input file structure might require reconfiguration or even code modifications. This limits the direct transferability of the solution to other instruments or laboratories without additional development work.

The metadata collection process, while comprehensive, relies on user input for information that is not automatically captured by the instrument. This introduces the possibility of incomplete or inaccurate metadata if users skip fields or provide incorrect information. Although the script includes prompts and validation for essential fields, it cannot enforce the quality or completeness of all the information provided.

Another limitation is the current lack of integration with data repositories or discovery platforms. While the converted data is in a FAIR-compliant format,

additional steps are required to make it truly findable and accessible to the wider scientific community. Without direct connections to repositories or catalogs, the data may remain isolated despite its FAIR-compliant format. The implementation focuses primarily on the interoperability and reusability aspects of FAIR principles, with less emphasis on the findability and accessibility dimensions.

The validation of converted data is also limited in the current implementation. While basic checks ensure that the conversion process completes successfully, more comprehensive validation against the NXxps standard or verification of data integrity would enhance confidence in the converted files. This is particularly important for scientific data, where small errors or omissions can significantly impact interpretation and analysis. The lack of comprehensive validation means that converted files might technically conform to the NeXus format while still containing errors or inconsistencies in the data or metadata.

The current implementation also lacks robust error handling for edge cases or unexpected data formats. If the input data deviates significantly from the expected structure, the conversion process might fail without providing clear guidance on how to resolve the issue. This can be frustrating for users and might discourage adoption of the tool, particularly among researchers who are less comfortable with technical troubleshooting.

Additionally, the implementation requires manual intervention for each file conversion, which could become cumbersome for large datasets or batch processing. The lack of automation options limits the scalability of the approach, particularly in research environments where large volumes of data are generated routinely. A more automated workflow would enhance the practicality of the solution for everyday research activities.

Finally, the current implementation focuses primarily on the technical aspects of FAIR compliance, with less emphasis on the social and organizational changes needed to fully embed FAIR principles in research practices. Technical solutions alone cannot ensure the adoption of good data management practices without corresponding changes in research culture, incentives, and education. This broader context represents a limitation of the current approach, which addresses only one component of the complex system that constitutes research data management.

## 4.4. Future Directions

Looking ahead, several promising directions could extend and enhance the work presented in this thesis. A natural next step would be to integrate the conversion pipeline with data repositories and discovery platforms. This could involve implementing APIs to automatically deposit converted data in institutional or domain-specific repositories, along with appropriate metadata mappings to ensure discovery through various search interfaces. Such integration would further enhance the findability and accessibility of the data, completing the implementation of all

four FAIR principles.

Extending the solution to other types of spectroscopic data would broaden its impact. The approach developed for XPS data could be adapted to similar techniques such as Ultraviolet Photoelectron Spectroscopy (UPS), Auger Electron Spectroscopy (AES), or even complementary techniques like X-ray Diffraction (XRD). This would create a more comprehensive ecosystem of FAIR-compliant data across multiple experimental methods, facilitating integrated analysis and interpretation of complementary measurements.

Enhanced data validation and quality assurance mechanisms would strengthen the reliability of the converted data. This could include automated checks against the NXxps standard, verification of physical consistency, and detection of potential errors or anomalies in the data. Such features would be particularly valuable for ensuring the scientific integrity of data intended for sharing or publication. Validation tools could range from simple syntax checkers to more sophisticated validators that examine the scientific credibility of the data and metadata.

Integration with data analysis workflows represents another promising direction. By ensuring that the converted data can be directly imported into common analysis tools, the practical utility of the FAIR-compliant format would be enhanced. This could involve developing plugins for popular analysis software or creating standardized analysis pipelines that work with the NeXus format. Such integration would create immediate practical benefits for researchers, encouraging adoption of the FAIR-compliant format as part of their regular workflow rather than as an additional step.

Automation of the conversion process for batch processing would enhance the scalability of the solution. This could include a watchfolder approach, where new data files are automatically detected and converted, or a scheduling system that processes files at regular intervals. Such automation would reduce the manual effort required to maintain FAIR data practices and would be particularly valuable in high-throughput research environments.

The development of a graphical user interface (GUI) for the conversion pipeline would make it more accessible to researchers with limited programming experience. A well-designed GUI could guide users through the conversion process, provide real-time feedback on metadata quality, and visualize the resulting FAIR-compliant data. This would lower the technical barriers to adoption and broaden the potential user base.

Collaboration with other research groups and standardization bodies could contribute to the refinement and broader adoption of the NXxps standard. By sharing experiences and best practices, the community could work toward a more robust and widely accepted standard for XPS data, enhancing interoperability across different research environments and instrumentation. This collaborative approach would align with the community-driven nature of successful standardization efforts in other scientific domains.

Finally, broader educational and outreach initiatives could help promote the adoption of FAIR principles beyond the technical solution. Workshops, tutorials,

and documentation could help researchers understand the benefits of FAIR data management and how to incorporate it into their daily practices. Collaboration with other laboratories and research groups could also facilitate the exchange of best practices and the establishment of community standards. Building a community of practice around FAIR data management for spectroscopy would sustain and extend the impact of the work presented in this thesis.

## 4.5. Concluding Remarks

The implementation of FAIR principles for XPS data represents a concrete step toward more open, transparent, and efficient scientific research. By addressing the practical challenges of data conversion and metadata management, this work contributes to the broader movement toward Open Science. The solutions developed here demonstrate that FAIR principles can be successfully applied to specialized scientific domains with appropriate tools and methodologies.

The journey toward fully FAIR scientific data is ongoing, requiring continuous adaptation to new technologies, standards, and research practices. However, each implementation like the one described in this thesis brings us closer to a future where scientific data is more easily shared, understood, and reused. This not only enhances the efficiency of individual research projects but also accelerates the collective advancement of scientific knowledge.

As data volumes continue to grow and research becomes increasingly collaborative and interdisciplinary, the importance of FAIR data management will only increase. The XPS data converter developed in this thesis addresses a specific need within the MODA laboratory but also serves as a model for similar initiatives in other fields. It shows how domain-specific challenges can be overcome with tailored solutions that respect the unique characteristics of particular experimental techniques while adhering to broader principles of data management.

The technical solution presented here is just one component of a larger socio-technical system that supports responsible data management. For FAIR principles to achieve their full potential, they must be embraced by researchers, supported by institutions, and incorporated into funding and publication requirements. This broader ecosystem of support and incentives will determine the long-term success of FAIR data initiatives.

Despite the limitations and challenges discussed, the implementation of FAIR principles for XPS data represents meaningful progress toward a more open and collaborative scientific enterprise. Each step toward standardization, interoperability, and thoughtful metadata management contributes to a foundation upon which future researchers can build. The value of these efforts may not always be immediately apparent, but they create the conditions for more robust, reproducible, and impactful scientific research.

In closing, this thesis demonstrates that the technical barriers to FAIR data

management can be overcome with appropriate tools and methodologies. The next challenge is to build a research culture that values and prioritizes good data management practices as an integral part of the scientific process. By providing a practical implementation for a specific domain, this work contributes to that larger goal and offers a template for similar efforts in other fields. The path toward fully FAIR scientific data is long and complex, but it is one worth pursuing for the benefit of science and society.

# Appendices

## Appendix A: Installation and Setup Guide

This appendix provides detailed instructions for setting up the conversion pipeline for XPS data in the MODA laboratory. The installation process requires several dependencies and proper configuration to ensure smooth operation.

Python Requirements must be installed first. The converter tool relies on Python 3.6 or higher. Several packages are essential for operation. The pynxtools library forms the core of the conversion process. It can be installed via pip using the command "pip install pynxtools". The XPS plugin for pynxtools is also required. YAML support is needed for metadata handling, installed via "pip install pyyaml". PyYAML handles the metadata files critical to the conversion process.

System requirements aren't demanding. Any modern computer should suffice. The script was tested on Windows 10 and Linux Ubuntu 20.04. Memory requirements depend on data size but typically 4GB RAM is adequate. Disk space requirements vary with the number and size of files being processed.

The configuration file placement is important. The MODA\_config.json file must be in the same directory as the converter script. This file contains the mapping information essential for proper conversion. All paths in the configuration file are relative to the script location.

Setting up directories follows a simple pattern. Create a working directory for the conversion process. Place the Python converter script in this directory. Copy the MODA\_config.json file to the same directory. Place VMS files to be converted in this directory. Optional YAML metadata files can be included as well.

Testing the installation is straightforward. Open a command prompt or terminal in the working directory. Run the script with the command "python nexus\_converter\_moda\_xps.py". If dependencies are properly installed, the script should display a list of available VMS files. This confirms the basic functionality is working.

Common installation issues might occur. If pynxtools installation fails, ensure Python development headers are installed on the system. If the script fails to find the configuration file, check that it's in the correct directory. Permission errors may occur if the script lacks write access to the directory. These can be resolved by modifying file permissions.

The dataconverter command must be accessible in the system path. This command is installed with pynxtools but sometimes requires additional path configuration.

Adding the pynxtools bin directory to the system PATH variable typically resolves this issue.

Updates to the converter might be released periodically. Check the repository for new versions. Configuration files should be kept up to date with instrument changes. The mapping structure may need adjustment if instrument output formats change.

## Appendix B: NeXus File Structure and Hierarchy

The NeXus file format follows a hierarchical structure based on HDF5. This structure is designed to organize scientific data in a standardized way. The NXxps application definition specifies the particular structure used for XPS data.

The top level of the hierarchy contains the entry group. This group includes basic information about the experiment. Metadata fields at this level provide context for the entire dataset. This includes experiment title, start and end times, and identifiers.

The instrument group contains information about the equipment used. The source\_probe group describes the X-ray source characteristics. The electron\_analyzer group contains details about the analyzer configuration. The detector group includes detection parameters and raw data.

The coordinate system group defines spatial relationships. This is particularly important for XPS where sample orientation affects results. Multiple coordinate transformations establish relative positions of components. The transformations use rotation and translation matrices for precise positioning.

The sample group contains information about the material being analyzed. This includes sample name, identifier, and any relevant properties. Sample orientation is defined through transformation matrices. These transformations position the sample within the overall coordinate system.

The data group holds the actual spectroscopic data. This includes both raw data from the detector and processed spectra. Energy axis information associates energy values with data points. Signal data arrays contain the measured intensity values. Error estimates may also be included for uncertainty quantification.

The NeXus structure allows for standardized access to data. Software can navigate this hierarchy using well-defined paths. The structure remains consistent across different experiments. This consistency enables automated processing and analysis tools to work with any compliant file.

Additional groups may be present for specific analysis needs. Calibration information can be stored in dedicated groups. User-defined annotations can be added in custom groups. The extensible nature of HDF5 allows for these additions while maintaining compatibility.

# Bibliography

- [1] B. Fecher and S. Friesike, "Open Science: One Term, Five Schools of Thought," in *Opening Science*, S. Bartling and S. Friesike, Eds., Cham: Springer, 2014, pp. 17-47.
- [2] S. Venkataraman and J. P. Hey, "Open Science: The new paradigm for knowledge creation and dissemination," *Harvard Data Science Review* **3**(4), (2021). <https://doi.org/10.1162/99608f92.76565a22>
- [3] H. Piwowar and T. Vision, "Data reuse and the open data citation advantage," *PeerJ* **1**, e175 (2013). <https://doi.org/10.7717/peerj.175>
- [4] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- [5] M. Stocker et al., "Persistent Identification of Instruments," *Data Science Journal* **19**(1), (2020), p. 18. <https://doi.org/10.5334/dsj-2020-018>
- [6] B. Mons et al., "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud," *Information Services & Use* **37**(1), (2017), pp. 49-56.
- [7] G. Grossi, M. Buscema, M. Giannocco, and F. Priami, "The NFFA-DI Project: Building a Data Infrastructure for Nanoscience in Italy," *Digital Infrastructures for Research* (2022), pp. 125-132.
- [8] P. Manghi, L. Candela, and D. Castelli, "A reference architecture for data infrastructure supporting open science," *Program: Electronic Library and Information Systems* **53**(3), (2019), pp. 298-310.
- [9] O. Peroni, A. Mannocci, and P. Manghi, "Research Data Management in the Time of the FAIR Principles," *International Journal of Digital Curation* **15**(1), (2020), pp. 1-12.
- [10] A. Pasquali, P. Pitsch, and S. De Carlo, "Metadata Standards for Scientific

Data: The NeXus Format and Its Application in X-ray Spectroscopy," *Journal of Synchrotron Radiation* **28**(5), (2021), pp. 1420-1429.

[11] C. L. Borgman, "Big Data, Little Data, No Data: Scholarship in the Networked World", MIT Press, 2015

[12] H. Stodden, P. Guo, and Z. Ma, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," *PLoS ONE* **8**(6), e67111 (2013).

[13] Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., & Robinson, G. E. "Big data: Astronomical or genetical?" *PLoS Biology*, **13**(7), e1002195 (2015) <https://doi.org/10.1371/journal.pbio.1002195>

[14] Weinstein, J., Collisson, E. *et al.* "The Cancer Genome Atlas Pan-Cancer analysis project". *Nat Genet* **45**, 1113–1120 (2013) <https://doi.org/10.1038/ng.2764>

[15] F. Rossi, M. Caldarola, and P. Rossi, "Challenges in Implementing Open Science Principles: The Case of Small Research Groups in Nanoscience," *Journal of Electronic Publishing* **24**(1), (2021), pp. 1-15.

[16] S. Aalbersberg et al., "Making Science Transparent By Default: Introducing the TOP Statement," *Proceedings of the National Academy of Sciences* **118**(41), e2103238118 (2021).

[17] C. Tenopir et al., "Data sharing by scientists: practices and perceptions," *PLoS ONE* **6**(6), e21101 (2011).

[18] R. Katz et al., "Cost models for running an electronic laboratory notebook service," *International Journal of Digital Curation* **12**(2), (2017), pp. 215-227.

[19] F. Ardito, G. Michelotti, and V. Caruso, "The FAIRmat initiative: Standardizing materials science data for interoperability," *Scientific Data* **8**, 273 (2021).

[20] F. Miletto Granozio, "TWEET: Towards ferroelectricity in two dimensions" <https://www.tweet.spin.cnr.it/moda/>  
And <https://www.spin.cnr.it/outreach-and-t-t/media>

[21] R. Eason, Ed., *Pulsed Laser Deposition of Thin Films: Applications-Led Growth of Functional Materials*, Wiley-Interscience, 2007.

[22] J. F. Watts and J. Wolstenholme, *An Introduction to Surface Analysis by XPS and AES*, 2nd ed., Wiley, 2019.

- [23] C. S. Fadley, "X-ray photoelectron spectroscopy: Progress and perspectives," *Journal of Electron Spectroscopy and Related Phenomena* **178-179**, (2010), pp. 2-32. <https://doi.org/10.1016/j.elspec.2010.01.006>
- [24] F. de Groot and A. Kotani, *Core Level Spectroscopy of Solids*, CRC Press, 2008.
- [25] T. L. Barr, "Modern ESCA: The Principles and Practice of X-Ray Photoelectron Spectroscopy," *CRC Press*, 1994.
- [26] G. Greczynski and L. Hultman, "X-ray photoelectron spectroscopy: Towards reliable binding energy referencing," *Progress in Materials Science* **107**, 100591 (2020). <https://doi.org/10.1016/j.pmatsci.2019.100591>
- [27] C. Willenberg, P. Willmott, and S. Ward, "Electronic Laboratory Notebooks in a Public–Private Partnership," *Journal of Laboratory Automation* **19**(2), (2014), pp. 152-160. <https://doi.org/10.1177/2211068213501725>
- [28] S. Köster et al., "NIAC: A Nexus International Advisory Committee Technical Report on NeXus Data Format," *International Journal of Digital Curation* **15**(1), (2020), pp. 1-25.
- [29] D. G. de Oteyza and C. Rogero, "The Role of Interoperability in Surface Science: Towards a Standard Data Format," *Applied Surface Science* **504**, 144313 (2020). <https://doi.org/10.1016/j.apsusc.2019.144313>