MDMC
Master in Data Management and Curation

STRUCTURE AND SCIENTIFIC PROGRAM OF MDMC

PROGRAM OVERVIEW

The ***Master in Data Management and Curation – MDMC* - is a postgraduate specialization course** designed to equip students with advanced skills in managing the full research data lifecycle, with a FAIR-by-design approach, from acquisition, cleaning, and analysis, to storage, preservation, publication and beyond

The MDMC responds to the growing demand for professionals who can ensure that data are managed according to the FAIR principles (Findable, Accessible, Interoperable, and Reusable).

These principles are essential for fostering open science, enhancing data quality and integrity, and maximizing the reuse of research results across domains.

Uniquely focused on the complexities of research data, MDMC emphasizes real-world application through a hands-on **internship and a thesis project**, enabling students to design and implement automated FAIR-by-design data workflows in active scientific environments.

The skills and tools developed during the MDMC are also highly transferable to industrial settings. The methodologies used to make research data FAIR-by-design closely align with the techniques adopted in industries for designing robust ETL/ELT (Extract, Transform, Load) pipelines. Both rely on structured metadata, quality control, traceability, and semantic interoperability to ensure data can be efficiently integrated, transformed, and reused.

This cross-sector applicability opens career opportunities not only in academia and research infrastructures but also in industry roles related to data governance, data engineering, business intelligence, and beyond.

Examples of **where FAIR Expertise Meets Industry Needs**

MDMC graduates are well-prepared to contribute to roles in sectors such as:

- **Pharmaceutical & Life Sciences**: Clinical data standardization, regulatory compliance
- **Manufacturing & IoT**: Sensor data ingestion, real-time monitoring pipelines
- **Energy & Utilities**: Environmental data curation
- **AI/Data Platforms**: Enabling machine-actionable datasets

SCIENTIFIC CURRICULUM

The 2025–2026 edition introduces new modules on **Artificial Intelligence and Machine Learning**, tailored to data challenges in scientific research.

The MDMC is a full-time 10-months program, which consists of 1500 hours of training activities, corresponding to 60 ECTS credits, and includes:

Coursework (10 ECTS – 250 hours)

8 weeks of intensive classes from mid Sepetmber to mid December 2025, delivered in person in Trieste (Area Science Park and SISSA campuses):

Training Modules	Hours
Foundation of Open Science	2
Open Access and Scholarly publishing	6
Data Stewardship and the Research Data Management Lifecycle	6
FAIR Data Principles for Responsible Research	8
Scientific Programming Environment	18
Scientific Programming Methods	18
Legal and Licensing Frameworks for Data and AI	8
Data Quality	6
Python for Data Management (Part I and II)	36
Data Infrastructure	24
Tools for Data Management and Curation: a FAIR-by-design approach	24
Introduction to AI	30
AI for Scientific Applications	30
External Speaker Seminars	24
Specialized Workshops	10
Total	**250**

Proposed Scientific Modules for the MDMC Program

1. Foundation of Open Science (2 hours):

Understand the philosophical and practical foundations of the Open Science movement. Explore their relevance in the evolving European research landscape, including major initiatives such as EOSC and Horizon Europe.

Key topics:

- Open Science frameworks and initiatives
- Policy context and funders' mandates

Learning Outcomes:

- Understand the core values, motivations, and historical evolution of the Open Science movement
- Identify and analyse key European and international Open Science initiatives (e.g., EOSC, UNESCO Recommendations)

2. Open Access and Scholarly publishing (6 hours):

Explore the evolving landscape of scholarly communication and publishing in the context of Open Science. This module delves into the principles, policies, and practices of Open Access, and their implications for researchers, institutions, and funders. Understand how data and publication workflows converge in the drive toward transparent, equitable, and responsible scientific dissemination.

Key topics:

- Models of Open Access (Gold, Green, Diamond, Hybrid)
- Preprints, repositories, and institutional archiving
- Tools and (AI) services supporting open publishing

Learning Outcomes:

- Understand the principles and models of Open Access publishing
- Integrate data sharing and publication practices to enhance openness and reproducibility

- Critically evaluate scholarly publishing options

---

3. **Data Stewardship and the Research Data Management Lifecycle (6 hours):** Explore the roles and responsibilities of a Data Steward, and learn how to manage research data across its entire lifecycle—from planning to preservation:

Key topics:

- Data stewardship roles and profiles
- The Research Data Management (RDM) lifecycle
- Best practices for RDM

Learning outcomes:

- Describe each phase of the research data lifecycle and its implications for data quality, integrity, and reuse
- Apply best practices for data management throughout the lifecycle, including naming conventions, file formats, version control, and secure storage
- Recognize common data challenges (e.g., data loss, poor documentation, ethical risks) and implement strategies to mitigate them

---

4. **FAIR Data Principles for Responsible Research (8 hours)**

Dive deep into the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—and learn how to interpret and apply them across different disciplines and data types. This module addresses both the theoretical foundations and the practical challenges of implementing FAIR, extending the discussion to new frontiers such as software and AI models.

Key topics:

- In-depth analysis of FAIR principles
- Structure of Data Management Plans
- Integration of FAIR in data management plans and research workflows

Learning Outcomes:

- Critically interpret each FAIR principle and understand their technical and conceptual implications
- Integrate FAIR principles into planning, execution, and publication of research outputs

---

5. Legal and Licensing Frameworks for Data and AI (8 hours)

Understand the evolving legal and ethical landscape surrounding data, software, and AI. This module provides a comprehensive overview of European regulations affecting data management and reuse, including the Data Act, AI Act, and Digital Markets Act. It also introduces participants to data licensing models and legal considerations when working with sensitive, personal, or proprietary data.

Key topics:

- The EU Data Act, AI Act, Digital Services and Markets Acts
- Data protection regulations (GDPR, ethical compliance)
- Data licensing: Creative Commons, Open Data Commons, software licenses
- Legal aspects of data reuse and sharing
- Intellectual property rights and copyright in data and software
- AI transparency and accountability principles

Learning Outcomes:

- Understand the core provisions and implications of the EU Data Act, AI Act, and Digital Markets Act for research and data management
- Identify legal obligations and ethical considerations when working with personal or sensitive data (e.g., under GDPR)
- Apply appropriate data and software licenses to ensure compliance and promote reuse
- Evaluate the legal risks and responsibilities related to data sharing, AI model deployment, and cross-border data flows

---

6. Data Quality (6 hours)

Principles and practices of data quality, with a focus on how to assess, monitor, and improve the quality of research data. Understanding and maintaining high data quality is essential for reproducibility, compliance, interoperability and data reuse.

Key topics:

- Dimensions of data quality (accuracy, completeness, consistency, timeliness, etc.)
- Techniques for assessing and improving data quality
- Tools and methods for data validation, cleaning, and error correction
- The impact of poor data quality on research outcomes

Learning Outcomes:

- Define the key dimensions of data quality and understand their importance in research
- Apply techniques for data quality assessment, including validation and cleaning methods
- Implement best practices to maintain data quality throughout the data lifecycle
- Identify common data quality issues and implement strategies for their resolution
- Evaluate the impact of poor data quality on research integrity, reproducibility, and publication

7. Scientific Programming Environment (18 hours)

This course provides an in-depth introduction to Unix-like operating systems and the tools that underpin modern scientific programming environments. It combines theoretical background with practical exercises aimed at developing core technical skills for data professionals and researchers working in computational environments.

Key topics:

- *Unix architecture: kernel vs. userspace, processes, file system semantics*
- *Command-line tools and shell scripting (Bourne shell)*
- *File system and access control (permissions, groups, SELinux)*
- *Text editors (Vim, Nano) and file managers (Ranger, broot)*

Learning Outcomes:

- Work confidently in a Unix/Linux CLI environment
- Automate tasks with shell scripts
- Manage software and user permissions.

8. Scientific Programming Methods (18 hours)

This module focuses on the essential tools and practices for developing scientific software in collaborative environments. Students will explore how to version, document, and share code according to FAIR principles for software, and how to integrate these practices into reproducible research workflows.

Key topics:

- *Version control with Git and collaborative workflows (GitHub/GitLab)*
- *Software documentation, licensing, and metadata (CodeMeta, software citation)*
- *FAIR principles for research software*
- *Packaging and sharing, basic CI/CD concepts*

Learning Outcomes:

- Use Git for collaborative coding in research projects
- Apply FAIR principles to software development and sharing
- Document, license, and publish research software with proper metadata

---

9. Python for Data Management – Part I (18 hours)

This module introduces Python programming fundamentals with a focus on data management workflows and basic AI model handling. Students will learn to write clean code, use key libraries for data manipulation and visualization, and get started with PyTorch to explore simple AI models.

Key topics:

- *Python syntax and programming logic, debugging and code structure*
- *Working with standard libraries and virtual environments*

- *Data import/export (CSV, JSON, Excel, etc.)*
- *Data manipulation with pandas, numerical computing with numpy*
- *Basic data visualization*
- *Introduction to AI tools with PyTorch*

Learning Outcomes:

- Write and debug Python scripts
- Work with core data libraries
- Visualize and analyze tabular datasets
- Understand and run basic AI models using PyTorch

10. Python for Data Management – Part II (18 hours)

This advanced module builds on Python fundamentals to introduce scalable, modular data applications. Students will learn object-oriented programming (OOP), work with databases using Object-Relational Mapping (ORM), and interact with external data services through APIs.

Key topics:

- *Object-Oriented Programming (OOP): classes, inheritance, encapsulation*
- *Database design and interaction via SQL and SQLAlchemy (ORM)*
- *REST APIs: data exchange with requests*
- *Structuring reusable and scalable Python modules*

Learning Outcomes:

- Design modular applications
- Manage and query databases through ORM
- Build Python scripts that interact with external data systems

11. Data Infrastructure (24 hours)

This module introduces the foundational components for building robust and FAIR-compliant research data infrastructures. It combines theoretical principles with practical sessions focused on relational databases, web frameworks, and system integration.

Key topics:

- *Software architecture concepts for FAIR research digital structures*
- *Basic Relational Databases schema design principles*
- *Web Framework fundamentals: Django*
- *Content Management System and external interfaces*

Learning Outcomes:

- Create and manage relational databases
- Develop and deploy web-based data access systems using Django
- Integrate CMS platforms with external services to enhance usability and data sharing

12. Tools for Data Management and Curation: a FAIR-by-design approach (24 hours)

This module covers key concepts, tools and methods to implement FAIR-by-design pipelines to manage, curate, and preserve research data in compliance with FAIR principles.

Key topics:

- *Ontology, data model and metadata schema*
- *Data ingestion from different sources, Data harmonization and data formats*
- *Interacting with data repository*
- *'Real world' data interoperability: main challenges and examples*
- *Exploring FAIR data as Fully AI Ready*
- *Brief introduction to Data Policy and Data Governance*

Learning Outcomes:

- Apply metadata standards to ensure interoperability

- Use curation tools to prepare datasets for sharing and preservation
- Automate routine data management tasks to improve efficiency

13. Introduction to AI (30 hours)

This module introduces fundamental AI concepts with a focus on probabilistic inference, unsupervised learning, and neural networks. It combines theoretical insights with practical understanding relevant to data science and research.

Key topics:

- *Basics of Bayesian Inference*
- *Unsupervised Learning and Dimensionality Reduction*
- *Neural Networks*

Learning Outcomes:

- Understand and apply Bayesian inference concepts
- Perform unsupervised learning and dimensionality reduction
- Explain neural network fundamentals and learning dynamics

14. AI for scientific applications (30 hours)

Applying Artificial Intelligence Techniques to Real-World Scientific Problems. This module focuses on practical AI methods and their application in scientific research. Students will learn how to implement, customize, and evaluate AI models to solve complex problems in various scientific domains.

Courses will be delivered by an international faculty composed by national and international well-renowned experts in the field.

- **Internship (20 ECTS – ~500 hours)**

A six-month internship (from January to June 2026) in partner laboratories, where students will work on experimental data collection and develop their own FAIR-by-design thesis project. Internships are supervised and provide hands-on experience in real scientific contexts.

- **Project Work (5 ECTS – 125 hours)**

Each student will deliver a final project (thesis) detailing the design and implementation of a FAIR-by-design data pipeline developed during their internship.

- **Individual Study (25 ECTS – 625 hours)**

Time allocated for self-study, assignment preparation, and thesis writing.

- **Duration and Modality**

  - Total Duration: 10 months, from September 15, 2025 to June 30, 2026
  - Teaching Language: English
  - Teaching Mode: In-person, with some remote sessions during the internship period
  - Location of Coursework: Trieste – Area Science Park & SISSA

- **Internship Opportunities**

A distinctive feature of the MDMC program is its **extensive six-month internship**, during which students gain hands-on experience by working directly within research laboratories and data-intensive environments.

- **Real-world Projects Across Italy**

In the 2024–2025 edition, MDMC students completed internships at prestigious institutions across Italy, including:

  - Several laboratories of the **National Research Council (CNR)** nationwide
  - **University of Milan (UniMi)**
  - **Politecnico di Milano (PoliMi)**
  - **University of Salento**
  - **University of Salerno**
  - The **LAboratoty of Data Engineering (LADE)** at **Area Science Park** in Trieste

These experiences allowed students to apply FAIR-by-design data management practices in a wide range of scientific and organizational contexts, from materials science to life sciences and beyond.

- **Ongoing and New Collaborations**

For the 2025–2026 edition, the MDMC will consolidate collaborations with many of these hosts and also establish **new partnerships**, including:

- **Neuroscience laboratories at SISSA**, where students will support data curation for high-impact experimental research.
- **SISSA Medialab**, a cutting-edge company developing digital infrastructures for scientific publishing and open science.
- **New industrial partners**, eager to support MDMC’s training mission by hosting internships aligned with data governance and FAIR principles.
- **CERIC-ERIC**, a multidisciplinary Research Infrastructure open for basic and applied users in the fields of Materials, Biomaterials and Nanotechnology.

📌*Internships are carefully supervised and matched to each student's interests and background, with the goal of developing a personalized thesis project applying FAIR-by-design principles to real data challenges.*

For further information, contact us sending an email to **MDMC**.

AREA
SCIENCE PARK

PER SEGUIR VIRTUTE E CANOSCENZA
SISSA