PILOT TRAINING COURSE

(MASTER) IN DATA MANAGEMENT AND CURATION

MDMC

STUDENT GUIDELINES

SEPTEMBER 2024

Table of Contents

1. INTRODUCTION

The "Master in Data Management and Curation" is a training course aimed at developing skills in the management, curation, cataloguing, and analysis of research data. This aligns with the requirements of the European Commission, other funding bodies and more generally with the Open Science framework, which ask that research data be handled according to the FAIR principles (Findable, Accessible, Interoperable, Reusable) to ensure responsible and reproducible research.

The structure and program of the master are entirely innovative on both national and international levels, with a specific focus on the implementation of FAIR-by-design processes during the laboratory internship. FAIR-by-design is a native management approach for research data: it involves managing research data from their creation, ensuring they are handled according to the FAIR principles until their publication. To achieve FAIR-by-design, it is essential to design and implement an automated chain of hardware and software connections including research instrumentation, data annotation software to obtain metadata (additional information about the data), and data analysis programs. The adoption of FAIR-by-design significantly reduces human intervention during all phases of the research data lifecycle, while ensuring high data quality and integrity. This approach requires considerable effort and resource investment in the initial design phase, but it represents a strength for the research infrastructures that adopt it, significantly improving the management and quality of data in experimental and computational laboratories.

In this context, also the possibilities offered by the European Open Science Cloud (EOSC) need to be disseminated. The EOSC provides the enabling framework to share, connect and upscale best practices and services by the communities to implement FAIR principles for (open, where possible) data sharing and management and it is essential that scientific infrastructures and their users are strictly linked to this initiative.

Having EOSC compliant Research Infrastructures and FAIR-by-design Research Data Management is among the objectives of the two supporting projects:

- NFFA-DI (Nano Foundries and Fine Analysis - Digital Infrastructure);
- PRP@CERIC (Pathogen Readiness Platform for CERIC-ERIC Upgrade);

funded by the National Recovery and Resilience Plan ("PNRR") within "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", with funds from the European Union – NextGenerationEU.

This first edition of the **Master in Data Management and Curation** (MDMC), consists of a pilot course with a limited number of participants, who have been identified by the various Operational Units of the supporting projects.

The organization of the proposed course is governed by a cooperation agreement pursuant to art. 15 of Law no. 241 of 7 August 1990 and art. 7, par. 4, of Legislative Decree no. 36 of 31 March 2023 among Area Science Park, CNR and SISSA (Scuola Internazionale Superiore di Studi Avanzati) containing all the details about the duration and the structure of the course.

Pursuant to Article 9 paragraph 1 of the aforementioned agreement, to ensure the comprehensive development of MDMC, it is established an equitable Scientific committee. This committee will be responsible for defining the scientific, educational, and organizational needs of the program. The scientific committee is composed by:

- prof. Giorgio Rossi (University of Milan and CNR-IOM);
- dr. Stefano Cozzini (Area Science Park);
- prof. Stefano Ruffo (SISSA)

Additionally, dr. Mariarita de Luca has been appointed by the Scientific committee as a coordinator to oversee and manage the activities, ensuring smooth organization and coordination.

2. STRUCTURE OF THE COURSE

MDMC is a nine-month professional training program with the initial six weeks of full-time live lectures in Trieste, and the following seven months of internship in a laboratory of one of the supporting projects (i.e., NFFA-DI and PRP@CERIC).

The course structure is outlined in Table 1 and described in detail below.

	Part I	Part II	Part III	Part IV
Duration	6 weeks (~ 166h)	~ 2-3 days	7 months	~ 2-3 days
Dates	September 16th - October 25th 2024	October 28th - 30th 2024	November 2024 - May 2025	end of May 2025
Topic	Introduction to Data Management and tools	Definition of FAIR-by-design approach in the labs	Implementation of FAIR-by-design approach in the labs	Thesis Discussions
Location	Training in Trieste	Presentations and meetings in Trieste	OU and labs	Presentations and meetings in Trieste

Table 1: Outline of the structure and organization of the course.

> ➢ **Part I:** (From September 16th to October 25th 2024) in Trieste, six weeks of intensive (six hours per day, ~ 166 hours total) live lectures, described in detail in section 5. The lectures and presentations will be in English.
> ➢ **Part II:** (From October 28th to October 30th 2024) in Trieste, two-three days of presentations by each participant to outline the FAIR-by-design thesis project agreed with the supervisors. All local supervisors (from the laboratories) are invited in Trieste for the FAIR-by-design project presentations.
> ➢ **Part III:** (From November 2024 to May 2025) in the selected laboratory of the Operational Unit of origin, seven months of thesis work to implement the FAIR-by-design project tailored to the needs of the specific laboratory;
> ➢ **Part IV:** (End of May, dates TBD) in Trieste, two-three days of presentations by each participant to describe the FAIR-by-design thesis project developed and carried out in laboratory;

3. OBJECTIVES

The involved projects have defined specific training objectives regarding FAIR data management to train qualified data managers, data collectors, data curators and data stewards.

In particular, WP8 of NFFA-DI – Training of a new generation of RI operators and of researchers for exploiting NFFA-DI and European analytical RIs, aims to carry out closed-number training in Data Management and Data Curation, aimed at preparing future Data Curators and/or Data Stewards who will operate according to FAIR principles at the various Operational Units of the project. Similarly, WP12 of PRP@CERIC is dedicated to training and education, aiming at answering the needs of hands-on knowledge from both the RI scientific staff's and academic or industrial user's perspectives, with particular attention to the implementation of FAIR methods of data production, management and curation at each Operational Unit.

The **intermediate objectives** of MDMC training program can summarized as:

a) First level: developing DMPs at single laboratory level to implement FAIR data concept in-in-house research.

b) Second level: FAIR-data services implemented and possible development of user interfaces for exploiting these services.
c) Third level: Engagement of participants for implementing DMPs in all laboratories; user training for guiding to FAIR-data production and for exploiting FAIR-data services.
d) User assistance by participants for generating FAIR data sets from their research at the NFFA-DI and PRP@CERIC facilities.

The achievement of the intermediate objectives of the Master in Data Management and Curation will lead to the following **outcomes**:

1. having a stable presence in all RI nodes of FAIR data experts both at the level of implementation, maintenance and constant upgrade of the FAIR-by-design technology, interoperability with the centralized headquarters, and stewardship with respect to the users;
2. creation of a new generation of experts, formed in FAIR data management and stewardship.

The constant supervisors commitment in all laboratories in supporting participants is essential to achieve such ambitious results.

4. PARTICIPANTS

In this pilot version, during the academic year 2024/2025, MDMC foresees the participation of 33 students selected by the various Operational Units of the supporting projects.

The course involves the participation of two categories of students:

- 19 full participants: students enrolled in the entire training program (six weeks of live lecturers and seven months of laboratory internship);
- 14 auditors: Students enrolled only in the six weeks of live lectures.

4.1. Participant learning goals

At the end of the course each participant will have acquired the skills and competences to operate with:

- Open Science principles and practices, within the context of Horizon Europe Framework programme and EOSC;
- FAIR principles: data FAIR-by-design approach and FAIR-ification of data processes;
- Basic knowledge of data infrastructure and cloud data infrastructure
- Tools and software for data acquisition and metadata enrichment;
- Tools and methods for data and metadata analysis.

The tangible results, obtained step-by-step, are a set of FAIR-by-design data acquisition procedures and automatic software methods for metadata enrichment. For instance, students will be able to build unsupervised procedures that retrieve, collect, and assign automatically useful information. They will test the implementation of workflows for a FAIR-by-design data acquisition in the Operational Unit to which they belong.

Given the heterogeneity of participants, **before the beginning of the live lectures an assessment survey** will be proposed to each participant to better define eventual additional learning goals that will be determined with participants themselves based on their level of experience, and the needs of their laboratory.

4.2. Calendar

The detailed calendar of the six weeks of live lectures is available as [Google calendar](#), below a detailed list of the dates and locations in Trieste:

- September 16th: Conference Hall, Building C, Area Science Park, Campus Padriciano
- September 17th - 20th: Room 005, SISSA, Campus Bonomea
- September 23rd - 27th: Room 128-129, SISSA, Campus Bonomea
- September 30th - October 4th: Room D, SISSA, Campus Beirut
- October 7th - 25th: Auditorium, Building C, Area Science Park, Campus Padriciano
- October 28th - 30th: Conference Hall, Building C, Area Science Park, Campus Padriciano

The first day of MDMC is scheduled for September 16th, 2024, starting at 10:00 AM. The morning session will focus on presenting the objectives and organization of MDMC, along with an overview of the PNRR projects that enabled this training initiative. In the afternoon, technical support will be provided for computer setup for all students.

4.3. Certificate of attendance

Participation is required in presence during the six weeks of live lectures.

Full participants that will attend at least 70% of all the training modules and will discuss a FAIR-by-design thesis project will receive a certificate of attendance of the Master in Data Management and Curation with a statement of the topics covered and the title of the project thesis.

Auditors will receive a certificate of attendance with a statement of the topics covered in the attended (for at least 70% of their duration) training modules.

We will request to all participant to sign the attendance sheet to demonstrate their presence in the classroom.

4.4. Laboratory internship and MDMC thesis

Full participants, enrolled in the entire training programme, will attend the six weeks of live lectures of Part I and the following seven months of internship in the selected laboratory of the Operational Unit of origin.

In Part II, the students must present with a brief talk the idea of the FAIR-by-design project that will be developed during the laboratory internship. The course lecturers and the laboratory staff are invited in Trieste to participate to the presentations (28th-29th October 2024).

During the internship period in the laboratory (Part III), participants will have **two supervisors**. A **data supervisor**, identified between the course lecturers, who will support the students with regular follow-up to ensure the successful implementation of the FAIR-by-design project; and a **local supervisor** identified in each laboratory, to support students with respect to the technical characteristics of the laboratory instruments.

Part III also foresees disseminating the FAIR knowledge acquired by the participants to the rest of the laboratory staff. The dissemination methodology will be established based on standard laboratory practices and overseen by the local supervisor.

At the end of the internship, during Part IV, students will discuss their thesis in Trieste (dates TBD) in front of a panel composed by (local and data) supervisors and invited external experts.

5. SCIENTIFIC PROGRAMME

The training modules in Part I have been designed to provide all the skills and competences necessary for the development and execution of the subsequent FAIR-by-design project in the laboratory where the following seven-month internship will be carried out. All the lectures and presentations will be in English.

The list of the seven modules provided is given here below:

- Introduction to Open Science (OS)
- Scientific Programming Environment (SPE)
- Cloud Data Environment (CDE)
- Python for data management (PY)
- Data Infrastructures (DI)
- Data Management Tools (DMT)
- Introduction to Statistical Data Analysis and Machine learning (SDA&ML)

5.1. Training materials

All the training materials (slides and additional materials) will be openly available on the Zenodo Community of MDMC : Master Data Management and Curation (MDMC) (zenodo.org)

5.2. Invited external experts

In addition to the planned training modules, the course calendar includes three seminars with invited external experts:

- September 26th 2024, h 11:15, "Legal framework of data stewardship: European and Italian perspectives", Ludovica Paseri, University of Turin,
- October 2nd 2024, h 11:15, "Introduction to intellectual property and FOSS: basic theory and praxis", Matija Šuklje, Liferay,
- October 16th 2024, h11:15, "Data Management and Curation for scientific research: relevant questions, case studies and future developments", Vittorio Iacovella, University of Trento.

5.3. Detailed description of training modules and objectives

Module 1	**Introduction to Open science (OS)**
Course coordinator	Mariarita de Luca (MdL)
Lecturers	Elena Giglia and MdL
Tutors	Gerardina Cargnelutti, Stefania Cantagalli and Barbara Corzani
Class duration	21 hours

Module description	**Principles:** Open Science is about increased transparency, re-use, participation, cooperation, accountability, and reproducibility for research. It aims to improve the quality and reliability of research through principles like inclusion, fairness, equity, and early sharing. Open Science can be viewed as research simply done properly, and it extends across the Life and Physical Sciences, Engineering, Mathematics, Social Sciences, and Humanities (Open and Responsible Research). **Practices:** Open Science includes changes to the way science is done, including opening access to research publications, data-sharing, open notebooks, transparency in research evaluation, ensuring the reproducibility of research (where possible), transparency in research methods, availability of open-source code, software and infrastructure, with the contribution of citizen science and the accessibility of open educational resources.
Main topics	- Concepts and principles of Open-Science - Policies on Open Science (European Commission, United Nations...) - The European context: Horizon Europe and OS practices - Open Science components: focus on publications (Open Access flavors, early sharing and open peer review - Open Science components: Open Research Data and FAIR principles, open formats, open licensing - FAIR Research Data Management concepts, general tools and techniques - FAIR principles and Data Management Plan structure - European OS platforms: EOSC (European Open Science Cloud) and Dataspaces
Objectives	On successful completion of this module students should have their own understanding on what is Open Science, how to adopt and adapt Open Science practices to different research fields, how to apply FAIR principles in Research Data Management. They will also be able to find documentation and tools to support researchers inOpen Science and FAIR practices.
References	- (OS Training Handbook Introduction · GitBook, 2018, no copyright, contents to be updated) - The Turing way handbook to reproducible, ethical and collaborative data science - Passport for Open Science, 2024 - UNESCO Open Science toolkit - FAIR Principles: Interpretations and Implementation Considerations, 2020

Module 2	**Scientific Programming Environment (SPE)**

Course coordinator	Ruggero Lot (RL)
Lecturers	Nicolò Tosato, Gianfranco Gallizia and RL
Tutors	Massimo Luin
Class duration	21 hours
Module description	Introduction to Linux operating system and fundamental concepts of containers and container orchestration. In the first part, students will learn the Linux basics, including the command-line interface, file system navigation, user and package management, and collaborative tools. Additionally, the course covers essential Linux utilities, scripting, and an introduction to system administration concepts.
Main topics	- Introduction to Unix-like operating systems - (kernel vs. userspace, processes/threads, file system semantics) - Shell scripting (bourne shell) - Configuring, compiling, linking software packages - principal commands, shell, tools, and scripting [1, 4] - packet manager (dnf, yum, zipper, pacman) [4] - Cgroups, Resource monitoring [TBD] - Editors (Vim, nano) and file managers (ranger or broot, vim) [1] - Command line environment [1, 4], org.freedesktop standard [3] - Access control (permission, groups, home) [4], selinux - Collaborative source code management: Git [1, 2] - File system [4] - Network configuration [5] - Integrated development environments - Visualization tools: - Debugging tools
Objectives	On successful completion of this module students should have their own software environment and tools prepared and configured for the rest of the activities of the master program. Students will learn the workflows of software development and collaborative work. These concepts are the basis on which efficient data application and workflow for data management will be developed.

References	- [1] the missing semester, mit https://missing.csail.mit.edu/), - [2] pro git (https://git-scm.com/book/en/v2) - [3] freedesktop specs and thereafter (https://specifications.freedesktop.org/basedir-spec/basedir-spec-latest.html) - [4] Learning Modern Linux (978-1-098-10894-6) - [5] Computer Networking (978-0-13-285620-1)

Module 3	**Cloud Data Environment (CDE)**
Course coordinator	Ruggero Lot (RL)
Lecturers	Gianfranco Gallizia and RL
Class duration	21 hours
Module Description	Introduction of cloud computing infrastructures in the framework of data analysis. Students will be introduced to the main concepts related to Cloud Computing, and they will be able to experience the installation, configuration and usage of a cloud infrastructure and discussing how to use them proficiently to deploy simple applications in data related applications.
Main topics	- Introduction to Virtual Machine and Containers. - docker / docker compose / docker file - podman - Networking container - AWS
Objectives	On successful completion of this module students should have the basic skills to operate in a cloud environment, to setup and configure their own scientific environment, and to use them proficiently to deploy simple applications in data related applications.
References	- Many resources online available

Module 4	**Python for data management (PY)**

Course coordinators	Jacopo Nespolo (JN) and Matteo Poggi (MP)
Lecturers	JN and MP
Class duration	30 hours
Module description	Introduction to Python language for data management. The course is in three different parts: the first one is a quick overview of the language structure and philosophy, the second on practical use of Jupiter notebooks and the third on specific Python libraries for data management.
Main topics	First part: introduction to Python - why it is so important in data science - basic features and good practices in Python - how to check which version of Python - create and env - different env manager - Python on different OS - IDEs - Datatypes - Context managers - Functions - Classes - Super quick overview on software design (SOLID principles) in Python - Unittesting with pytest Second part: Jupyter notebooks (install, configure, etc...) - Git - Release code on github/gitlab - sphinx or mkdocs - how to document your code Third part: data handling and visualization - numpy - intro - broadcast - big data in numpy - dask/zarr - alternative to numpy for big data - when and why use dask and zarr - pandas - series and data frames - missing data aggregation

	- pandas to sql - Matplotlib - data visualization
Objectives	On successful completion of this module students should be able to write simple Python programs and know the right tool to easily develop Python workflows on data managements.
References	Lot of online references

Module 5	**Data infrastructures (DI)**
Course coordinator	Marco Prenassi (MP)
Lecturers	MP
Class Duration	21 hours
Module Description	This module discusses the main component of data infrastructure both on hardware and software side. Data lake concepts and practical implementation will be presented, discussed and used.
Main Topics	- The lower level: Ansible, Ceph, bucket, modern Filesystems features for data managing and resilience. - Infrastructure scalability (horizontal vs vertical) - Relational databases and SQL - Optimizing DB schema - ORM, SQLAlchemy - Rest API in flask - NoSQL databases - Data lake concepts and implementation - Data access control
Objectives	On successful completion of this module students should be able to understand the main concepts behind data infrastructure, interact properly with data infrastructures available to them and define their own requirements to consider when a data infrastructure is planned.

Module 6	**Data Management Tools (DMT)**
Course coordinator	Stefano Cozzini (SC)
Lecturers	Federica Bazzocchi and SC
Tutors	Rodolfo Tolloi, Cecilia Zagni
Class duration	24 hours
Module Description	This module will focus on data management and curation knowledge and skills. It will provide students with an introduction to research data management and curation with practical examples and case studies. Students will understand their data management needs across the research data lifecycle, and they will be made familiar with best practices for working with FAIR-by-design data. After this course, students will be better equipped to manage data throughout their laboratory activities and to offer data stewardship support to the laboratory staff.
Main Topics	- **Data management and curation**: definition, meaning and their interplay. Data management as a combination of software, tools and best practices. Different approaches to data curation. Data and metadata concepts. - **Data management Plans**: Scientific data lifecycles. Data management plan structure and requirements. Useful tools for data management plan. FAIR assessment tools. - **Open Data Repository:** worldwide scenario, FAIR certification, solutions adopted in NFFA-DI. - **Metadata**: definition and their importance in data lifecycle. Collecting metadata. Electronic notebook. Ontology and vocabulary. Choosing a metadata schema. - **Formats**: Metadata formats. Open data formats. Hierarchical Data Format and Research Object Crate. Useful packages. - **Scientific Data Management System:** Architecture & Interfaces. Database system for scientific data. Relational and non-relational database systems. - **Workflows:** Planning workflow for scientific data curation. Workflows for open science. Workflow tools, libraries examples.

Objectives	On completion of this module students should have understood meanings and advantages of FAIR-by-design experimental data acquisition, be able to write a realistic and sustainable DMP, plan a workflow for data acquisition and use tools and scripts to realize it.

Module 7	**Introduction to Statistical Data Analysis and Machine Learning (SDA&ML)**
Course coordinator	Matteo Biagetti (MB)
Lecturers	Tommaso Rodani and MB
Class duration	24 hours
Module description	This module will focus on a variety of techniques and tools for the statistical analysis of data and meta-data for parameter inference, pattern recognition, model fitting and comparison. The module is split into two main parts: an introduction to probability theory and statistics, and an introduction to machine learning and neural networks. Each topic is introduced with a hands-on approach, with plenty of practical examples and exercises.
Main topics	**Introduction to probability theory and statistic:** - elements of probability theory, - sampling and sampling distributions, - point and interval estimation, - hypothesis testing, - chi-square test, - nonparametric methods, - elements of Bayesian inference - Model fitting and comparison **Introduction to Machine Learning and Neural Networks** - Linear methods for regression and classification - Kernel methods for regression and classification - Learning with imbalanced/missing data - Hyperparameter tuning, cross validation - Data intrinsic dimension and dimensionality reduction techniques - Clustering methods - Elements of deep learning

Objectives	On completion of this module students should have learnt how to use a basic set of algorithms and tools for analyzing data in their laboratory. They can devise several techniques to 1) extract information/infer parameters, 2) recognize unknown/known patterns, 3) fit theoretical or empirical models, 4) solve simple regression or classification tasks using machine learning algorithms, from a dataset and/or its metadata.
References	Books: - 'Pattern Recognition and Machine Learning', C. Bishop - 'An introduction to Statistical Learning' G. James, D. Witten, T. Hastie, R. Tibshirani -'The elements of statistical learning' T. Hastie, R. Tibshirani, J. Friedman -'Deep learning with Python', F. Chollet Other references: -https://www.biophys.mpg.de/2139320/statistical-data-analysis - https://ocw.mit.edu/courses/15-075j-statistical-thinking-and-data-analysis-fall-2011/pages/lecture-notes/ - https://people.sc.fsu.edu/~sshanbhag/Cowan_Statistics.pdf

6. PRACTICAL INFORMATION

6.1. Opening day of MDMC

The opening day of MDMC will be held on September 16th 2024. The morning session will start at 10:00 at the Conference Hall, Building C, Area Science Park, Campus Padriciano. This session is dedicated to the presentation of the whole training program "Master in Data Management and Curation" and the supporting projects NFFA-DI and PRP@CERIC. In Annex 1 the detailed program of the morning session. All the sessions will be in English.

The afternoon session will start on 14:00 until 16:00 and will be dedicated to student registration and technical support for computer set-up.

6.2. Student registration

It will be possible to register for MDMC on Monday September 16th during the afternoon session starting from 14:00 to 16:00. SISSA staff will be available to help participants with all the paperwork and documentation needed for student enrolment.

6.3. Course requirements and set-up

The specific requirements for the 6 weeks of live lessons are the following:

- each participant's laptop should have a working Linux environment (with Linux OS pre-installed/Ubuntu Linux on a Windows machine/Linux virtual machine);
- before the beginning of the live lessons (17.09.2024) each participant must request the access to Area Science Park computation facility ORFEO, following the instructions available at the web page: Getting access - ORFEO documentation (areasciencepark.it).

A specific session on Monday September 16th will be dedicated to offer technical support to participants in the computer set-up (from 14:00 to 16:00).

6.4. Support for MDMC

Participants who need support about administrative problems, certificate of attendance, information and/or any other issue can refer to:

- Laboratorio Interdisciplinare – SISSA
  Via Bonomea, 265 - 34136 TRIESTE
  Email : ilas@sissa.it, read by
  - Sabrina Morena: tel. +39 040 3787631
  - Claudia Parma: tel. +39 040 3787401
  - Mila Bottegal: tel. +39 040 3787549
- MDMC email : mdmc@sissa.it

For any issue regarding the account on ORFEO datacenter please send an email to:

- support@areasciencepark.it with [ORFEO] in the subject
- or visit the web page Contacts - ORFEO documentation (areasciencepark.it)

6.5. Access to services ARDiS for students enrolled in university level studies

All participants can have access to ARDiS services (regional agency for the right to education) of the Friuli Venezia Giulia. All available services are described on the website ARDIS > ardiss/index.

For catering services, which can be used in the canteens of Area Science Park, SISSA and ICTP, and to consult the discounted rates, please consult the page ARDIS > ardiss/contenuti.

To use the services, it is necessary to register online on the portal:

Sportello Studente (dirittoallostudio.it)

6.6. Living and moving in Trieste

Practical information about the students staying in Trieste can be found at the following link

Home | Welcome Office FVG.

Lectures will be held on the SISSA and AREA SCIENCE PARK campuses (see Figure 1):

- **AREA SCIENCE PARK** campus Padriciano (Località Padriciano 99)
  reachable by bus number 51 or 51/ from the Central Train Station,
- **SISSA campus Bonomea** (Via Bonomea 265)
  reachable by bus number 38 from Oberdan Square,
- **SISSA campus Beirut** (Via Beirut 2, near Castello di Miramare in Figure 1)
  reachable by bus number 6 from the Central Train Station.

Further information about transports in Trieste are available at:

Homepage | Trieste Trasporti

Home - TPL FVG

All the information in the present document and other information are available at a dedicated web page: master-data-management-and-curation | Scuola Internazionale Superiore di Studi Avanzati (sissa.it)
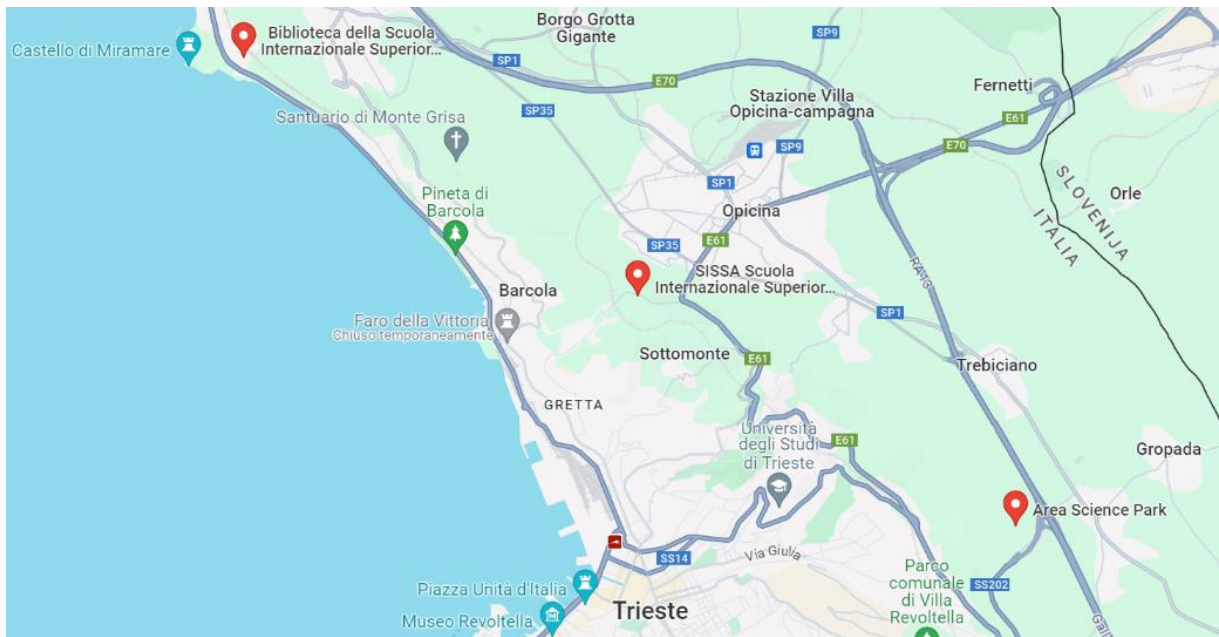

Figure 1: Maps of the live lesson locations.