MISSIONE 4
ISTRUZIONE
RICERCA

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Progetto NFFA-DI – PNRR 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" – CUP B53C22004310006.

PILOT TRAINING COURSE IN DATA MANAGEMENT AND CURATION

APRIL 2024

General release

nffa-di
Piano Nazionale di Ripresa e Resilienza

Consiglio Nazionale delle Ricerche | AREA SCIENCE PARK | POLITECNICO MILANO 1863 | UNIVERSITÀ DEGLI STUDI DI MILANO

Table of Contents

1. INTRODUCTION

In the digital and data-driven paradigm promoted by Open Science, data is at the core of the scientific process and its production grows at ever increasing rates. The skills and knowledge of Scientific Data Management and Curation techniques are nowadays essential to ensure responsible and reproducible research.

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 PILOT TRAINING COURSE in Data Management and Curation, that in the future is set to become a proper professional **Master in Data Management and Curation** (MDMC), consists of a pilot course with a limited number of participants, who will be identified by the various Operating Units of the supporting projects.

2. OBJECTIVE

The involved projects have defined specific training objectives regarding FAIR data management. Specifically, WP8 of NFFA-DI – Training of a new generation of Research Infrastructure (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 project laboratories and locations involved (called hereafter "Operating Units", or OU). 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 OU involved.

To pursue the projects training objectives, the pilot course in Data Management and Curation is designed aiming at the following results:

- to make young researchers fully aware of the advantages of using a FAIR approach on Research Infrastructures as a unique resource for their research and scientific carrier, looking forward to growing a new class of professionals for the upcoming scientific challenges;

- to strengthen the staff competences of each Operational Unit of the PNRR supporting projects to operate the new upgraded instrumentation according to FAIR principles, to be part of the FAIR-data system and become compliant with EOSC.

2.1. Learning goals

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

- Open Science principles and methodologies, within the context of Horizon Europe Framework programme and EOSC;
- FAIR principles: data FAIR-by-design approach and FAIRification of data processes;
- Tools and software for data acquisition and metadata enrichment;
- Tools and methods for preliminary 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 Operating Unit to which they belong.

Given the heterogeneity of participants, specific additional learning goals will be determined with participants based on their level of experience, and the needs of their laboratory.

3. STRUCTURE OF THE COURSE

MDMC will be a 9-month training program with the initial 6 weeks of intensive, in presence lessons, in Trieste and the following 7 months of internship in a laboratory of their Operating Unit, involved in 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 (~ 160h)	~ 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, 6 weeks of intensive (6-8 hours per day, ~ 160 hours total) in presence lessons, described in detail in section 5.
- **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 supervisor of the selected laboratory and a supervisor in Trieste (AREA SCIENCE PARK – CNR).
- **Part III:** (From November 2024 to May 2025) in the selected laboratory of the Operating Unit of origin, 7 months of thesis work to implement the FAIR-by-design project tailored to the needs of the specific laboratory; Students will be followed with periodic follow-ups and tutoring by the teachers involved in Part I of the course to verify the progress of the practical work.

- **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 the laboratory; Each student will have at disposal 20+10 minutes (discussion + questions) to present their thesis in front of a board composed by selected lecturers and invited experts.

3.1. Course requirements and set-up

- An individual laptop (provided by course participants);
- each participant's laptop should have a working Linux environment (with Linux OS pre-installed/Ubuntu Linux on a Windows machine/Linux virtual machine);

No programming competency is required as a prerequisite for participation but only familiarity with the use of computers to perform very simple scientific tasks (plotting/fitting data).

4. PARTICIPANTS

The course is addressed to the following participants:

1. Students that hold at least bachelor's degree ("laurea triennale" or equivalent);
2. Students still enrolled in a university master's course ("laurea magistrale" or equivalent), in science, engineering, or informatics;
3. Participants to the supporting PNRR projects (researchers, technologists, PhD students, research fellows);
4. other personnel who will participate only (or partially) in the intensive lessons of the first 6 weeks.

Participants belonging to profiles 1) and 2) will receive a lump sum reimbursement of 15.000 EUR for participating;

The participants with profile 3) and 4) will be designated by their respective OUs based on their operational needs.

4.1. Recruitment and notice of participants

Student should be recruited or selected by each partner project following its internal regulation. As regards the participants of categories 1) and 2) who will be expressed by the CNR, they will be recruited following their expression of interest accompanied by an adequate CV demonstrating possession of the requirements. CVs will be evaluated by a commission and candidates may also be subjected to an in-depth interview where deemed necessary.

The recruited participants will be enrolled as SISSA students and, for the whole period of the course, they will have a SISSA account and email.

4.2. Certificate of attendance

Reimbursed and staff 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 whole course with a statement of the topics covered and the acquired skills.

Lesson participants 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.

5. SCIENTIFIC PROGRAMME

The training modules of intensive lessons 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 7-month internship will be carried out.

The list of the 7 modules provided is given here below:

- Introduction to Open Science
- Scientific Programming Environment
- Cloud Data Environment
- Programming: Python for data management (comprising three parts)
- Data Infrastructures
- Tools for Data Management and Curation
- Introduction to Statistical Data Analysis and Machine learning

5.1. Introduction to Open Science – OS (20 hours)

MAIN TOPICS:

- Concepts and principles of Open-Science
- Open Research Data and FAIR principles
- The European context: Horizon Europe and EOSC
- Research Data Management concepts and techniques
- Open Access to Publications
- Open Science policies
- Open Licensing and File Format
- Practical exercises (e.g., Board game "Super-Open Researcher, DOAJ, Sherpa Romeo, Re3Data, Zenodo, etc…)

5.2. Scientific Programming Environment - SPE (20 hours)

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
- Packet manager (dnf, yum, zipper, pacman)
- Cgroups, Resource monitoring
- Editors (Vim, nano) and file managers (ranger or broot, vim)
- Command line environment, org.freedesktop standard
- Access control (permission, groups, home), selinux
- Collaborative source code management: Git
- File system

- Network configuration
- Integrated development environments
- Visualization tools
- Debugging tools

5.3. Cloud Data Environment – CDE (20 hours)

MAIN TOPICS:

- Introduction to Virtual Machine and Containers
- Docker files and Podman
- Kubernetes
  - pod, services,
  - LoadBalancer, Ingress, gateway
  - CNI, (Flannel, Calico)
  - Storage: PV, PVC, CSI (nfs, ceph)
- Helm
- Git Ops

5.4. Programming: Python for data management – PY (30 hours)

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

MAIN TOPICS - Second part: Jupyter notebooks (install, configure, etc...)

- Git
- Release code on github/gitlab
- Sphinx or mkdocs

- How to document your code

MAIN TOPICS - 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

5.5. Data Infrastructures – DI (20 hours)

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

5.6. Tools for Data Management and Curation – TDMC (24 hours)

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.

5.7. Introduction to Statistical Data Analysis and Machine learning – SDA (24 hours)

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

6. OTHER INFORMATION

All the information in the present document and other information will be available by the end of April 2024 at a dedicated web page:

master-data-management-and-curation | Scuola Internazionale Superiore di Studi Avanzati (sissa.it)

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:

- **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)
- reachable by bus number 6 or 36 from the Central Train Station