European Research Council
Executive Agency

Established by the European Commission

European Research Council (ERC)

ERC Data Management Plan

European Research Council
Established by the European Commission

ERC OPEN RESEARCH DATA MANAGEMENT PLAN (DMP)

Project Acronym	Project Number
CoDeFeL	101096251

Template for the ERC Open Research Data Management Plan (DMP). The following sections should describe how you plan to make the project data Findable, Accessible, Interoperable and Reusable (FAIR). Each of the following five issues should be addressed with a level of detail appropriate to the project.

SUMMARY *(dataset[1] reference and name; origin and expected size of the data generated/collected; data types and formats)*

The ERC-CoDeFeL project works with two main types of data. First, publicly available datasets will be collected from open repositories such as kaggle.comand similar platforms. Second, synthetic datasets will be programmatically generated by the project team for benchmarking and validation purposes.

No sensitive, personal, clinical, or confidential data will be collected or processed at any stage of the project.

All processed datasets, source code, and experimental outputs will be stored and version-controlled in a public GitHub repository: https://github.com/DCN-FAU-AvH/.

The processed data will primarily consist of structured tabular datasets in **CSV** and **Parquet** formats, image datasets stored in Python array-like formats (e.g., **NumPy**), and synthetic datasets generated for numerical validation. The expected total size of all generated and processed datasets combined is approximately **10–50 GB**, depending on the number of experiment replications.

[1] *Several datasets may be included into a single DMP.*

ERC OPEN RESEARCH DATA MANAGEMENT PLAN (DMP)

1. MAKING DATA FINDABLE *(dataset description: metadata, persistent and unique identifiers e.g., DOI)*

All datasets used or produced in the project will be documented in accordance with the FAIR principles and ERC guidelines to ensure they are easily findable and reusable. Each dataset will include a clear description of its variables, detailed information on data provenance (including repository URLs and applicable licenses), and a record of any pre-processing steps applied. Additionally, all scripts used for data transformation will be openly available in the project's GitHub repository, ensuring full transparency and reproducibility.

Metadata will follow widely recognized standards such as **Dublin Core**, as recommended by the ERC and the Digital Curation Centre (DCC). Publicly sourced datasets will retain their original identifiers and URLs, while processed datasets hosted on GitHub will be versioned using Git commit hashes, providing persistent and traceable references for all data releases.

To further enhance findability, datasets will be assigned **Digital Object Identifiers (DOIs)** through Zenodo upon publication. Metadata will include relevant **keywords**, project identifiers, and links to associated publications to facilitate indexing in open repositories and search engines. This approach ensures that all data outputs are discoverable, citable, and interoperable within the broader research ecosystem.

2. MAKING DATA OPENLY ACCESSIBLE *(which data will be made openly available and if some datasets remain closed, the reasons for not giving access; where the data and associated metadata, documentation and code are deposited (repository?); how the data can be accessed (are relevant software tools/methods provided?)*

The project will make all non-sensitive materials openly accessible in line with the FAIR principles and ERC requirements. This includes all processed datasets generated from publicly available sources, all synthetic datasets created for benchmarking purposes, and all code, scripts, and documentation developed during the project.

As for eventual sensitive material that may be used in our research, this will be kept undisclosed unless explicit written authorization of the interested parties. This fact would activate a revision of the Data Management Plan for potential adjustments to guarantee the compliance with ERC Guidance.

These resources will be stored and version-controlled in the public GitHub repository: https://github.com/DCN-FAU-AvH/, which is fully open and does not impose any access restrictions.

Datasets obtained from external public repositories will remain accessible under their original terms of use, typically requiring only a free account. All software tools used in the project (e.g., Python, PyTorch, MATLAB) are either open-source or documented to ensure reproducibility.

To guarantee long-term accessibility and citability, major releases of datasets and code will be archived in **Zenodo**, where they will receive **Digital Object Identifiers (DOIs)**. This ensures persistent access and compliance with open science best practices.

Sensitive data are not expected in this project. Should any sensitive material be required, it will remain

undisclosed unless explicit written authorization is obtained from the relevant parties.

3. MAKING DATA INTEROPERABLE *(which standard or field-specific data and metadata vocabularies and methods will be used)*

Interoperability will be ensured by adhering to widely accepted standards and practices as recommended by the ERC guidelines. All datasets will be provided in universal, non-proprietary formats such as **CSV**, **JSON**, **TXT**, and **PNG**, which are broadly supported across scientific and machine learning workflows. Metadata will follow recognized standards, primarily **Dublin Core**, complemented by dataset-specific metadata aligned with the schema of the chosen repository when applicable.

To facilitate integration and reuse, data will be organized in clean tabular structures with clearly defined variable names and consistent naming conventions. All pre-processing steps will be documented in **Markdown files** and **Jupyter notebooks**, ensuring transparency and reproducibility. Furthermore, the project will provide code snippets and scripts to load and manipulate the data using standard Python libraries such as **pandas** and **NumPy**, enabling seamless integration into computational pipelines.

By using open formats, standardized metadata, and well-documented workflows, the project guarantees that its outputs can be easily combined with other datasets and incorporated into diverse analytical environments, including machine learning and scientific research platforms.

4. INCREASE DATA RE-USE *(what data will remain re-usable and for how long, is embargo foreseen; how the data is licensed; data quality assurance procedures)*

The project is committed to ensuring that all datasets and related outputs remain reusable well beyond its lifetime. To achieve this, clear **licensing**, transparent documentation, and robust **quality assurance** procedures will be implemented.

All synthetic datasets and processed derivatives will be released under MIT or CC-BY-4.0 licenses, enabling broad reuse without restrictions. Datasets obtained from public repositories will retain their original licenses, and appropriate reuse instructions and citation guidelines will be provided. No **embargo period** is foreseen; all non-restricted data and code will be made openly available on a continuous basis as results become available.

Quality assurance will be guaranteed through version control using GitHub commits, automated scripts for data cleaning and feature processing, and reproducible training and evaluation pipelines. Comprehensive documentation will accompany all datasets and code, following best practices recommended in the ERC DMP template, to ensure replicability and transparency.

For long-term preservation and reusability, major releases of data and code will be archived in Zenodo, where they will receive Digital Object Identifiers (DOIs). This guarantees persistent access and citability. In addition, the project commits to maintaining accessibility for at least **10 years**, in line with ERC recommendations on **long-term preservation**.

5. ALLOCATION OF RESOURCES and DATA SECURITY *(estimated costs for making the project data open access and potential value of long-term data preservation; procedures for data backup and recovery; transfer of sensitive data and secure storage in repositories for long term preservation and curation)*

Resources and Costs

No additional costs are anticipated for accessing datasets, as all data sources are publicly available through open repositories. Personnel time required for data curation, documentation, and maintenance is fully covered by the project's funding allocation. Since the project does not involve sensitive or personal data, no specialized security infrastructure or compliance costs are required.

Data Security and Backup

Although the datasets handled in this project are public and non-sensitive, appropriate measures will be implemented to ensure data integrity and availability. All code, processed datasets, and experimental outputs will be version-controlled and stored in the project's GitHub repository, which benefits from automatic replication across distributed servers. In addition, local institutional backups will be maintained to provide redundancy and safeguard against accidental data loss. No transfer or storage of personal, clinical, or confidential data will occur at any stage of the project.

Ethical and Legal Considerations

The project exclusively uses publicly accessible datasets and does not collect or process any personal, medical, or sensitive information. Consequently, no ethical restrictions apply, and the research fully complies with legal and institutional requirements. All reused datasets will respect their original licenses, and proper attribution will be provided in accordance with open science best practices.

DISCLAIMER. Please note that the ERC Data Management Plan is not a part of the Ethics Review. It is the responsibility of the Principal Investigator to inform the ERCEA Ethics Team of any ethics issues/concerns regarding the collection, processing, sharing and storage of data in relation to the project.