TECHNISCHE UNIVERSITÄT WIEN

Data management plan (DMP)

Analyzing the Connection Between Financial Earnings and Life Satisfaction in EU countries

EU-EARN-SATIS

Version	Effective date	Description of document/changes
1.0	28/05/2026	First version of the DMP – created for the start of the project

Level of distribution		 It is publicly available under (DOI to be assigned upon deposit)

Project details

Project Coordinator Principal Investigator	
Contact person (responsible for data management and DMP)	Ameer Hamza, e12448315@student.tuwien.ac.at , TU Wien, ROR: ror.org/04d836q62
Contributors	Ameer Hamza, e12448315@student.tuwien.ac.at , TU Wien, ROR: ror.org/04d836q62, Data Manager Haseeb Akhtar, e12510777@student.tuwien.ac.at, TU Wien, ROR: ror.org/04d836q62, Data Collector Matus Revay, e12544037@student.tuwien.ac.at, TU Wien, ROR: ror.org/04d836q62, Project Manager Gregor Vandák, e12128763@student.tuwien.ac.at, TU Wien, ROR: ror.org/04d836q62, Researcher
Start date	2026-04-30
End date	2026-06-30
Funder	
Funding programme, grant number	
Internal project number	

List of acronyms

DMP	data management plan
RDM	research data management
FAIR	Findable, Accessible, Interoperable, Reusable
ML	Machine Learning
API	Application Programming Interface
ROR	Research Organization Registry

Content

Introduction

Science Europe practical guide, FAIR data

A DMP is a structured document that keeps record of what research data is created and what happens to that data during and after a project. It helps with planning the research process and defining responsibilities in a research project involving several researchers or institutions.

For writing this DMP, we followed the recommendations of Science Europe as they reflect the guidelines agreed upon by the major funders in Europe.

To make our data FAIR, they generally will be treated according to the following criteria:

- We will make our data findable, by uploading it to a data repository that provides a persistent identifier and adding relevant metadata.
- We will make our data accessible by providing open access to data, wherever possible. In cases, where open access is not possible, we will provide meaningful metadata plus contact information for access requests.

- We will make our data interoperable by providing and describing data in a way that is common within our domain by using the same file formats, schemas and vocabularies. We will provide good documentation for all our datasets.
- We will make our data reusable by adding metadata and comprehensive Readme files to all published datasets. The descriptions include details on the methodology used, analytical and procedural information. In case of publication, licenses for code and data will always be assigned and clearly marked.

Relevant Policies and Guidelines

- European Commission's document on Ethics and Data Protection: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf
- - TU Wien Research Data Management Policy: https://www.tuwien.at/en/research/rtisupport/research-data
- Data description

1a Lists of datasets that will be reused or produced

Produced datasets

dataset ID	title	type	format	estimated volume	contains sensitive data
P1	EU Earnings and Job Satisfaction Input Dataset	Structured text	csv	100 - 1000 MB	no
P2	Trained Machine Learning Models	Software applications	pkl	100 - 1000 MB	no
P3	Generated Output Data, Figures and Evaluation Results	Structured text, Images	PNG , CSV	100 - 1000 MB	no

Description for "EU Earnings and Job Satisfaction Input Dataset":

Tabular dataset containing measurements of earnings and job satisfaction indicators across EU countries. Reused from a public source and ingested into DBRepo via REST API. Contains numeric attributes with units mapped to ontological concepts.

Description for "Trained Machine Learning Models":

Trained machine learning models produced from the EU earnings and job satisfaction input data. Deposited in the TU Wien Research Data Repository. Includes all hyperparameters, evaluation metrics, and FAIR4ML metadata.

Description for "Generated Output Data, Figures and Evaluation Results":

Output data produced by the ML pipeline including predictions, classification results, evaluation figures, and confusion matrices. Deposited separately in the TU Wien Research Data Repository as a Dataset resource type.

Technical resources for "EU Earnings and Job Satisfaction Input Dataset":

DBRepo REST APITechnical resources for "Trained Machine Learning Models":

TU Wien Research Data Repository (TUWRD)Technical resources for "Generated Output Data, Figures and Evaluation Results":

TU Wien Research Data Repository (TUWRD)

Reused datasets

1b Data generation and reuse

Methods and software used for data generation and reuse

The input dataset on EU earnings and job satisfaction is reused from an existing public source. Data is loaded into DBRepo using the REST API via a Jupyter notebook (Python). Machine learning models are trained using Python libraries including scikit-learn and pandas. Output data including predictions, evaluation metrics, confusion matrices, and figures are generated programmatically. All data access in the final experiment is done exclusively via the DBRepo REST API with no local file reads. To ensure reproducibility, data is retrieved automatically from DBRepo using a Python script (fetch_data.py) included in our code repository.

1. Documentation and data quality

2a Data organisation, metadata and documentation

The GitHub repository is structured with top-level folders: data/, src/, outputs/, docs/. A consistent file-naming convention is documented in the README covering input datasets, output files, scripts, and configuration files. Versioning is handled via GitHub releases. The project uses semantic versioning (v0.1.0, v0.2.0, etc.) with release notes documenting all changes between versions.

Metadata is provided using multiple standards: RO-Crate for the overall experiment package, CodeMeta for software metadata, FAIR4ML for trained machine learning models, Croissant for input datasets, and Model Cards for each trained model. All datasets are deposited in the TU Wien Research Data Repository with extensive metadata including ORCIDs, licences, and related identifiers. Semantic annotations are added to DBRepo using domain-specific ontologies. This will help others to identify, discover and reuse our data. We use the CodeMeta standard (codemeta.json) to describe our software and environment setup. Additionally, any machine learning models developed utilize Fair4ML metadata to ensure they remain FAIR.

Additionally, we will provide common metadata such as title, description or keywords when publishing data in open access repositories. In such a case, we will follow the default template provided by the repository, such as Data Cite Metadata or Dublin Core.

A far as possible, we will use controlled vocabularies for our data to allow inter-disciplinary interoperability and machine-actionability.

The README contains step-by-step reproduction instructions. All code is provided as Jupyter notebooks in the GitHub repository. The experiment is fully reimplemented to retrieve data via the DBRepo REST API, ensuring reproducibility. RO-Crate validation output is included under docs/validation/.

2b Data quality control

The following data quality checks will be done: standardised data capture, data entry validation, peer review of data and representation with controlled vocabularies.

2. Storage and backup during research process

3a Storage and backup facilities

For the duration of the project, storage and backup of data will be ensured by Ameer Hamza (acting as the person responsible for data management and DMP) in cooperation with the system operator. The data will be stored on the servers of TU Wien.

P1 (EU Earnings and Job Satisfaction Input Dataset), P2 (Trained Machine Learning Models), P3 (Generated Output Data, Figures and Evaluation Results) will be stored on TUgitLab: TUgitLab is an application for managing repositories based on Git provided and managed by Campus IT. Our institute's administrators will manage GitLab groups, assign project permissions, and appoint external project partners as additional GitLab users. This service is highly available and scalable on the Kubernetes platform.

3b Data security and protection of sensitive data

We pay strict attention to compliance with the relevant institutional and national data protection policies listed in the introduction of this document. At this stage, it is not foreseen to process any sensitive data in the project. If this changes, advice will be sought from the data protection specialist at TU Wien, and the DMP will be updated.

Access to data during research:

dataset ID	selected project members	all other project members	the public
P1	writing	reading only	no access
P2	writing	reading only	no access
P3	writing	reading only	no access

All incidents will be handled individually by an incident response team that is maintaining the affected service.

3. Legal and ethical requirements

4a Personal data

At this stage, it is not foreseen to process any personal data in the project. If this changes, advice will be sought from the data protection specialist at TU Wien, and the DMP will be updated.

4b Intellectual property rights and rights of use

The following individual(s) hold rights and control access to the project data: The input dataset is subject to its original licence from the public source. All four group members (Haseeb Akhtar, Matus Revay, Gregor Vandák, Ameer Hamza) have write access during the project. After publication, all artefacts are publicly accessible under their respective licences: input data under its original licence, software under MIT licence, and output data under CC BY 4.0.

4c Ethical issues

No particular ethical issue is foreseen with the data to be used or produced by the project. This section will be updated if issues arise.

4. Data sharing and long-term preservation

5a Data publication and access conditions

As far as possible, obtained datasets will be published in repositories. Details on access conditions, reuse licenses, reasons for restrictions, etc. are collected in the table below.

dataset ID	access conditions	estimated publication date	location for publication (repository)	PID	license
P1	Open	2026-06-01	TU Wien Research Data	DOI	CC-BY-4.0
P2	Open	2026-06-01	TU Wien Research Data	DOI	MIT
P3	Open	2026-06-01	TU Wien Research Data	DOI	CC-BY-4.0

Repository description:

TU Wien Research Data is an institutional repository of TU Wien to enable storing, sharing and publishing of digital objects, in particular research data. It facilitates the funders' requirements for open access to research data and the FAIR principles by making research output findable, accessible, interoperable, and reusable. A DOI is assigned to each dataset published in TU Wien Research Data. This service is developed by the TU Wien Center for Research Data Management and hosted by TU.it. https://researchdata.tuwien.ac.at/

Methods or software needed to access and use data: The input data is provided in CSV format, accessible without specialised software. Data can also be accessed via the DBRepo REST API using standard HTTP clients. The trained models are serialised in pickle format and require Python with

scikitlearn to load and run. The Jupyter notebooks require Python 3 with standard data science libraries (pandas, scikit-learn, matplotlib). All dependencies are documented in the repository with version pins

5b Long-term preservation and deletion of data

dataset ID	location for long-term storage	minimum retention period (≥ 10 years)	foreseeable research uses and/or users
P1	TU Wien Research Data	10 years	The primary audience is researchers and students in data science, economics, and social sciences interested in EU labour market trends. The input dataset on earnings and job satisfaction is relevant for economists and policy analysts studying quality of work across EU countries. The trained models and evaluation results are useful for data science educators and researchers benchmarking machine learning approaches on socioeconomic data. All artefacts are published openly to maximise reuse potential.
P2	TU Wien Research Data	10 years	
P3	TU Wien Research Data	10 years	

5. RDM responsibilities and resources

6a RDM-roles and responsibilities

Ameer Hamza (Data Manager) will direct the data management process overall, with the research assistants responsible for ensuring metadata production, day-to-day cross-checks, back-up and other quality control activities are maintained.

6b Resources

There are no costs dedicated to data management and ensuring that data will be FAIR.