Data Stewardship UE 2025S – Data Management Part 2

Version 1.0
19 April 2025

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	19.04.2025	▪ Initial version

Action Number: DS-2025-: Data_Stewardship

Action Acronym: CO2ML
Action title: CO_2 Emission Prediction Using Machine Learning

Date: 2025-04-19

DMP version: 1.0

The Horizon Europe Model Grant Agreement requires that a data management plan ('DMP') is established and regularly updated. The use of this template is recommended for Horizon Europe beneficiaries. In completing the sections of the template the requirements for research data management of Horizon Europe as described in article 17 and analysed in the Annotated Grant Agreement, article 17, must be addressed.

1. Data Summary

I reused publicly available data from the 'Our World in Data - CO_2 dataset', accessed through TU Wien's DBRepo infrastructure. Due to technical constraints, I was only able to upload a subset (~3,379 rows) of the global dataset to DBRepo. The model was trained on a further filtered subset of 1,000 rows.

The data includes country-level yearly indicators such as population, GDP, cement CO_2 emissions, and growth percentages, all in CSV tabular format.

The data was used to train a Random Forest regression model to predict CO_2 emissions. Though the subset size limited prediction performance, the workflow remained fully reproducible and FAIR-compliant.

The data could be reused in climate modeling, policy studies, machine learning, and teaching scenarios.

The full project repository is available at: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

2. FAIR data

2.1. Making data findable, including provisions for metadata

All datasets and outputs were uploaded to TUWRD and GitHub with metadata including title, creator, license, version, and keywords. Persistent identifiers (PIDs) are assigned by TUWRD and Zenodo.

Keywords include: "CO_2", "machine learning", "data stewardship", "FAIR data", "Random Forest", "climate data".
Metadata was structured using Dublin Core and extended with Codemeta and FAIR4ML (for model metadata).

GitHub Repository: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

2.2. Making data accessible

Data is openly accessible via TUWRD and GitHub. Model outputs (plots, predictions, trained model) were published under the MIT license. Data is downloadable through a free and open protocol (HTTPS).

An embargo is not required. No sensitive data is included, so access approval is unnecessary.

GitHub Repository: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

2.3. Making data interoperable

All data is in CSV format with UTF-8 encoding. Metadata uses standard vocabularies (Dublin Core, Codemeta).
The notebook and scripts follow Python and Jupyter conventions. Repositories include metadata files and structured folder layouts for reuse.

2.4. Increase data re-use

A README and code comments document methodology and assumptions. `requirements.txt` ensures environment reproducibility. The code and data outputs are licensed (MIT and CC-BY-4.0) and reusable.

The provenance of data and models is documented through versioning (GitHub) and persistent identifiers (TUWRD).

GitHub Repository: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

3. Other research outputs

Besides the data, this project produced:
- A trained Random Forest model (`rf_model.joblib`)

- A Jupyter notebook
- Output CSVs and plots
All are reusable and documented via TUWRD and GitHub.

GitHub Repository: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

4. Allocation of resources

No direct costs were incurred. All work was conducted using TU Wien infrastructure, GitHub, and Zenodo (free).
I handled all data management tasks personally.

GitHub Repository: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

5. Data security

No sensitive or personal data was used. Files are stored on trusted repositories: TUWRD, Zenodo, and GitHub.

GitHub Repository: https://github.com/KhasrurRahman/Data-Stewardship-UE-2025S---Data-Management-part--2-

6. Ethics

No ethical or legal risks were involved. The dataset contains only public macro-level indicators.

7. Other issues

No other national or departmental DMP procedures were used.