FWF Data Management Plan (DMP)

DMP: Traffic Accident Severity Prediction in Austria

Author: Md. Abdul Quaiyum | Student ID: 12449471 | TU Wien, 2026

Course: FAIR Data Science (DaSt 2026) | Template: FWF DMP (01/2022)

Data Officer

Who is responsible for the data management and the DMP of the project?

Md. Abdul Quaiyum, Student, TU Wien (matriculation number: 12449471). As the sole researcher on this educational project, the author is responsible for all data management activities, including data collection, processing, storage, and publication throughout the project lifecycle.

I Data Characteristics

I.1 Description of the Data

What kinds of data / source code will be generated or reused (type, format, volume)?

Reused input data: The primary dataset is the *Strassenverkehrsunfalle mit Personenschaden ab 2013* (Austrian Road Traffic Accidents with Personal Injury from 2013), published by Statistik Austria under CC BY 4.0 on data.statistik.gv.at. The dataset is available in CSV format and contains approximately 500,000 records covering all police-recorded accidents on Austrian public roads from 2013 onwards. Features include: accident location (coordinates, federal state), road category, speed limit, weather conditions, lighting conditions, day of week, time of day, number of vehicles involved, vehicle types, and injury severity label (slight / serious / fatal).

Generated data and code: The experiment generates (1) preprocessed CSV files (train/validation/test splits), (2) PNG image files (EDA histograms, confusion matrix, feature-importance chart, classification report visualisation), (3) a serialised Random Forest model in joblib format (~5–50 MB), and (4) Python source code (~200–500 lines). Total estimated output volume: under 200 MB.

How will the research data be generated and which methods will be used?

Input data is downloaded directly from the Statistik Austria open data portal. Data is processed using Python 3 with pandas (cleaning, encoding, normalisation) and scikit-learn (stratified train/validation/test splitting at 70/15/15 %). A Random Forest Classifier is trained using scikit-learn. Visualisations are produced with matplotlib and seaborn.

How will you structure the data and handle versioning?

The project follows a standard ML directory structure: /data/raw, /data/processed, /models, /outputs, /src. Code versioning is handled via Git. Dataset versions are tracked by download date and Statistik Austria dataset version number. The DMP itself is treated as a living document and versioned via the TU Wien Research Data Repository.

Who is the target audience?

The primary audience is the course instructors and peer reviewers of the DaSt 2026 FAIR Data Science course at TU Wien. Secondary audiences include other students reviewing DMPs during Part 5 of the exercise, and any interested parties accessing the openly published data and code.

II Documentation and Metadata

II.1 Metadata Standards

What metadata standards (if any) will be in use and why?

The dataset deposit will follow the **Dublin Core Metadata Initiative (DCMI)** standard, as supported by the TU Wien Research Data Repository (InvenioRDM). This includes: title, creator, description, subject, date, type, format, identifier (DOI), source, language, and rights/licence. For the ML model and code, schema.org and CodeMeta metadata vocabularies will be applied where possible to ensure machine-readability and interoperability.

II.2 Documentation of Data

What information is needed for the data to be findable, accessible, interoperable, and reusable (FAIR)?

Findable: All outputs (code, model, processed data) will be deposited in the TU Wien Research Data Repository with a persistent DOI. Descriptive metadata (title, author, keywords, abstract) will be provided. Keywords include: road safety, traffic accidents, machine learning, classification, Austria, Random Forest, scikit-learn. **Accessible:** All outputs will be published under an open licence (CC BY 4.0 for data, MIT for code), freely downloadable without authentication. **Interoperable:** Data is stored in open formats (CSV, PNG, joblib). Metadata uses Dublin Core. Code is documented with inline comments and a README. **Reusable:** A detailed README.md will describe the directory structure, how to install dependencies (requirements.txt), how to reproduce results, and the provenance of the input data.

Is the data machine-readable?

Yes. Input and processed data are in CSV format. Model outputs are PNG images (human-readable) and a joblib file (machine-readable). Source code is plain-text Python. Repository metadata is exposed via the InvenioRDM REST API, making it machine-harvestable.

How are you planning to document this information?

Documentation will be provided via: (1) a README.md file in the code repository describing all files, reproduction steps, and data provenance; (2) inline code comments and docstrings; (3) this DMP, stored alongside the deposit; (4) metadata fields in the TU Wien Research Data Repository deposit form.

II.3 Data Quality Control

What quality assurance processes will you adopt?

The following quality assurance steps are applied: (1) EDA phase: inspection of missing values, outliers, and class imbalance using pandas profiling; (2) Data validation: checking that column types and value ranges match the Statistik Austria data dictionary; (3) Reproducibility: a fixed random seed (random_state=42) is used for all train/test splits and model training; (4) Stratified splitting: ensures that class proportions (slight/serious/fatal) are preserved across splits; (5) Model evaluation: performance is reported on the held-out test set only, never on training data.

How will the consistency and quality of data collection be controlled and documented?

The input dataset is sourced exclusively from Statistik Austria's official open data portal and is not modified at source. All preprocessing steps are documented in the Python code with comments, and a data processing log (CSV) records the number of rows before and after each cleaning step. The exact download date and dataset version are recorded in the README.

III Data Availability and Storage

III.1 Data Sharing Strategy

How and when will the data be shared and made accessible?

All project outputs (processed data, trained model, source code, output figures) will be made publicly accessible upon completion of the experiment (Part 3 of the course). The input dataset (Statistik Austria) is already publicly available and will be referenced by URL and DOI rather than re-uploaded, in compliance with its CC BY 4.0 licence terms.

What repository will you be using?

The TU Wien Research Data Repository (https://researchdata.tuwien.ac.at), which holds the CoreTrustSeal certification and is listed in re3data. This satisfies the course requirement for a certified, trusted repository. For the test/exercise phase, the test instance at https://test.researchdata.tuwien.ac.at is used.

What persistent identifier will be used?

A DOI (Digital Object Identifier) will be assigned automatically by the TU Wien Research Data Repository upon deposit. The DOI will be recorded in this DMP and in the README of the code repository. The 'Cite all versions' DOI will be used to ensure the identifier always resolves to the latest version.

III.2 Data Storage Strategy

What data are to be preserved for the long-term, and what data will not be stored?

Preserved: Processed datasets (train/val/test CSV splits), trained model (joblib), all output figures (PNG), source code (Python), README, requirements.txt, and this DMP. **Not preserved long-term:** Intermediate temporary files generated during preprocessing (e.g., uncleaned intermediate CSVs) and IDE/cache files.

How and where will the data be stored and backed up during the research?

During active research, data and code are stored locally on the researcher's workstation and backed up continuously via Git to a private GitHub repository. The raw input dataset is re-downloadable at any time from data.statistik.gv.at.

How and where will the data be stored after the project ends?

After project completion, all final outputs will be deposited in the TU Wien Research Data Repository under an open licence, ensuring long-term preservation and accessibility beyond the course duration.

For how long will the data be stored?

The TU Wien Research Data Repository guarantees data preservation for a minimum of 10 years. No earlier deletion is planned.

Are there any costs that need to be covered for storage?

No. The TU Wien Research Data Repository is provided free of charge to TU Wien students and researchers. No additional storage costs are anticipated.

Are there any technical barriers to making the research data fully or partially accessible?

No technical barriers exist. All data formats (CSV, PNG, joblib, Python) are open and widely supported. The repository provides open HTTP access without login requirements.

IV Legal and Ethical Aspects

IV.1 Legal Aspects

Are there any legal barriers to making the research data fully or partially accessible?

No. The input dataset (Statistik Austria) is published under CC BY 4.0, which explicitly permits reuse, redistribution, and adaptation with attribution. All generated outputs (code, model, figures) are original works of the author and will be published under open licences.

Who owns the data?

The input data is owned by Statistik Austria (Republic of Austria). The processed datasets, trained model, source code, and output figures are created by Md. Abdul Quaiyum as part of a TU Wien course exercise.

What licence for reuse are you planning to attach to the data?

Input dataset: CC BY 4.0 (as per Statistik Austria — attribution required). **Processed data and outputs:** Creative Commons Attribution 4.0 International (CC BY 4.0). **Source code:** MIT Licence, allowing free use, modification, and distribution with attribution.

Are there any restrictions on the re-use of the data? If so, why?

No restrictions beyond the attribution requirement of CC BY 4.0. Users must credit Statistik Austria as the original data source and Md. Abdul Quaiyum for derived outputs.

IV.2 Ethical Aspects

Are there any ethical barriers to making the research data fully or partially accessible?

No significant ethical barriers exist. The Statistik Austria accident dataset is fully anonymised at the point of publication — no personal identifiers (names, addresses, vehicle registration plates) are included. The data describes aggregate accident circumstances, not individuals. No sensitive personal data is processed in this experiment.

If applicable, how are you planning to deal with sensitive data during and after the project?

No sensitive or personal data is involved. The dataset contains only statistical records of accident circumstances as collected and anonymised by the Austrian police and Statistik Austria. No additional anonymisation steps are required by the researcher.

Summary

Item	Detail
Project title	Traffic Accident Severity Prediction in Austria
Data officer	Md. Abdul Quaiyum (12449471)
Input dataset	Statistik Austria – Road Traffic Accidents 2013+
Input licence	CC BY 4.0
Source URL	data.statistik.gv.at
Output formats	CSV, PNG, joblib, Python (.py)
Output licence	CC BY 4.0 (data/figures), MIT (code)
Repository	TU Wien Research Data Repository (CoreTrustSeal)

Persistent ID	DOI (assigned upon deposit)
Metadata standard	Dublin Core / CodeMeta
Storage duration	Minimum 10 years
Personal data	None — dataset is fully anonymised
DMP version	v1 — initial DMP (Part 2)

This DMP was prepared for the FAIR Data Science course (DaSt 2026) at TU Wien following the FWF DMP template (01/2022). It is a living document and will be updated in Part 4 of the exercise.