TECHNISCHE UNIVERSITÄT WIEN

Data management plan (DMP)

Home Advantage in Professional Football – Research Data Management Project

Home Advantage in Football

Version	Effective date	Description of document/changes
1.0	22/11/2025	First version of the DMP – created for the start of the project
1.1	23/11/2025	Updated data sources section to include detailed information about reused datasets (R1, R2)
1.2	25/11/2025	Added produced datasets table and refined documentation

Level of distribution		This DMP is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). This DMP is not published in a public repository and does not have a DOI.

FWF Data Management Plan (DMP)

I General Information

I.1 Administrative information

PI: Bartosz Ciesielski, bartosz.ciesielski@student.tuwien.ac.at, TU Wien, ROR: ror.org/04d836q62, Project Leader

FWF project number:

Internal Project ID:

DMP version: 01, 22.11.2025

Contributors:

Tomasz Miksa, tomasz.miksa@tuwien.ac.at, ORCID: 0000-0002-4929-7875, TU Wien, ROR: ror.org/04d836q62, Hosting Institution

I.2 Data management responsibilities and resources

Person responsible for data management and DMP: Bartosz Ciesielski, bartosz.ciesielski@student.tuwien.ac.at, TU Wien, ROR: ror.org/04d836q62

Co-ordination of data management responsibilities across partners: Bartosz Ciesielski, bartosz.ciesielski@student.tuwien.ac.at, TU Wien, ROR: ror.org/04d836q62

Resources costed in for RDM: There are no costs dedicated to data management and ensuring that data will be FAIR.

II Data Characteristics

II.1 Data description and collection or re-use of existing data

Produced datasets:

dataset ID	title	type	format	estimated volume	contains sensitive data	description
P1	all_matches	Tabular data	CSV/XLSX	~50–100 KB	no	Full combined match dataset with basic preprocessing applied
P2	league_results_counts	Tabular data	CSV/XLSX	~50–100 KB	no	Match result counts grouped by league
P3	league_results_percent	Tabular data	CSV/XLSX	~50–100 KB	no	Match result percentages grouped by league

P4	overall_distribution	Tabular data	CSV/XLSX	~50–100 KB	no	Overall distribution of match outcomes across both leagues
P5	summary_all_matches	Tabular data	CSV/XLSX	~50–100 KB	no	Summary statistics for all matches combined
P6	summary_by_league	Tabular data	CSV/XLSX	~50–100 KB	no	Summary statistics grouped by league

Reused datasets:

dataset ID	title	source	rights (e.g. license)	contains sensitive data	description
R1	Professional football match results (Ekstraklasa)	https://www.football-data.co.uk/poland.php		no	…
R2	Professional football match results (Premier League)	https://github.com/jalapic/engsoccerdata/tree/master/data		no	…

Methods and software used for data generation and reuse

No new data will be generated. Existing football match datasets will be reused and processed using Python (pandas) and Jupyter Notebooks. Raw data will be stored in /data/raw/ and processed outputs in /data/processed/.

III Documentation and Data Quality

III.1 Metadata and documentation

Data organisation, metadata, and documentation:

The project follows a clear folder structure separating raw data (/data/raw/), processed data (/data/processed/), notebooks (/notebooks/), reports (/reports/), and source code (/src/). Raw data will remain unchanged, while all transformations and analyses will be stored as new processed files to ensure traceability. Versioning is handled through Git and GitHub. All changes to data processing scripts, notebooks, and documentation are tracked automatically via commits, providing a complete and transparent version history. File naming will follow a consistent structure that reflects content and processing steps. This setup ensures reproducibility, clear documentation of changes, and long-term maintainability of the research data.

The project will provide descriptive metadata to ensure that the datasets can be easily identified, understood, and reused. Metadata will include: dataset title, source URL, data origin, time period covered, variable descriptions (teams, goals, match results, home/away indicators), file formats, and licensing

<table>
<tr><td></td><td>information. As there are no specific domain metadata standards required for this type of sports statistics data, metadata will be documented at the project level through a detailed README file. Each dataset will include clear naming conventions and accompanying descriptions in the /data/processed/ directory. Information about data provenance, reuse conditions, and processing steps will be provided in the documentation so that users can trace how the raw data were transformed. This ensures transparency, discoverability, and reusability of all research outputs. This will help others to identify, discover and reuse our data.

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.

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

All analysis steps are documented in the Jupyter notebook (home_advantage_study.ipynb), including data loading, cleaning, transformations, and visualizations. The repository README describes the project structure, datasets, processing workflow, software used, and instructions for reproducing results. Processed data files and visual outputs are stored in /data/processed/ and /reports/figures/ to enable verification and reuse.</td></tr>
<tr><td>III.2 Data quality control</td><td>Data quality control:

The following data quality checks will be done: peer review of data.</td></tr>
<tr><td colspan="2">IV Data Storage, Sharing, and Long-Term Preservation</td></tr>
<tr><td>IV.1 Data storage and backup during the research process</td><td>Storage and backup facilities:

No storage options are specified at the moment.

Data security and protection of sensitive data:

We pay strict attention to compliance with the relevant institutional and national data protection policies. 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:

<table>
<tr><th>dataset ID</th><th>selected project members</th><th>all other project members</th><th>the public</th></tr>
<tr><td>R1</td><td>writing</td><td>reading only</td><td>no access</td></tr>
<tr><td>R2</td><td>writing</td><td>reading only</td><td>no access</td></tr>
</table>
</td></tr>
</table>

IV.2 Data sharing and long-term preservation

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
-	-	-	-	-	-

No new datasets will be published. All reused datasets remain under the licenses of their original providers.

Methods or software needed to access and use data: The data are provided in standard tabular formats (.xlsx and .csv), which can be opened with common tools such as Excel, Python (pandas), or R. No specialized software is required. The folder structure and documentation in the repository support straightforward reuse.

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
-	-	-	-

No new datasets will be published. The project uses only existing openly accessible datasets (R1 and R2), which remain available under their original licenses. No additional repository deposit is required.

V Legal and Ethical Aspects

V.1 Legal aspects

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.

Intellectual property rights and rights of use:

The following individual(s) hold rights and control access to the project data: Access to all datasets will be controlled solely by the project owner (Bartosz Ciesielski). No additional project members have access rights.

V.2 Ethical aspects	**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.