TECHNISCHE UNIVERSITÄT WIEN

Data management plan (DMP)

Correlation Between the High School Graduation Rates and the Personal Income Data

[acronym]

Version	Effective date	Description of document/changes
1.0	dd/mm/yyyy	First version of the DMP – created for the start of the project

Level of distribution		This DMP is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). It is publicly available under [fill DOI here]. [If you do not publish this DMP, remove the CC BY icon and change the text above accordingly.]

Project details

Project Coordinator Principal Investigator	
Contact person (responsible for data management and DMP)	
Contributors	Renata Hess, rehess2004@gmail.com, ORCID iD: 0009-0001-0669-6585, Technische Universität Wien
Start date	2024-10-31
End date	2025-01-09
Funder, funding programme, grant number	
Internal project number TU Wien	

List of acronyms

DMP	data management plan
RDM	research data management
…	…
…	…
…	…
…	…
…	…
…	…

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

- TU Wien Policy for Research Data Management: https://www.tuwien.at/index.php?eID=dms&s=4&path=Directives%20and%20Regulations%20of%20the%20Rectorate/Policy%20for%20Research%20Data%20Management.pdf
- TU Wien Code of Conduct – Rules to Ensure Good Scientific Practice: https://www.tuwien.at/index.php?eID=dms&s=4&path=Directives%20and%20Regulations%20of%20the%20Rectorate/Code%20of%20Conduct%20%E2%80%93%20Rules%20to%20Ensure%20Good%20Scientific%20Practice.pdf
- Directives and Regulations of the TU Wien Rectorate: https://www.tuwien.at/en/tu-wien/organisation/central-divisions/data-protection-and-document-management/directives-regulations/
- TU Wien Data Protection: https://www.tuwien.at/en/tu-wien/organisation/central-divisions/data-protection-and-document-management/data-protection-at-tu-wien
- Other (e.g. from a project partner)

1. Data description

1a Lists of datasets that will be reused or produced

Produced datasets

dataset ID	title	type	format	estimated volume	contains sensitive data

Reused datasets

dataset ID	title	PID (e.g. DOI) or source	rights (e.g. license)	contains sensitive data
R1	Personal Income Data	https://apps.bea.gov/regional/zip/SQINC.zip		no
R2	Public high school graduates, by region, state, and jurisdiction: Selected school years, 1980-81 through 2031-32	https://nces.ed.gov/programs/digest/d23/tables/xls/tabn219.20.xlsx		no

1b Data generation and reuse

Methods and software used for data generation and reuse

The data will be used to analyze the relationship between personal income and education rates across U.S. states. The dataset includes two main variables: personal income (from personalIncome.csv) and educational attainment rates (from educationalRate.csv). The data processing and aggregation steps include cleaning the data, handling missing values, and transforming the data into a format suitable for analysis. The goal is to create visualizations comparing these two factors. The software used for this project will be Python, with key libraries such as Pandas for data manipulation and Matplotlib for creating visualizations. Pandas will be employed to load, clean, and merge datasets, while Matplotlib will be used to generate various plots, such as bar charts for comparing educational rates and income across states.

2. Documentation and data quality

2a Data organisation, metadata and documentation

The filenames will follow the projects naming convention and include a timestamp of creation. Version control is automated.

As there are no domain specific metadata standards applicable, we will provide a README file with an explanation of all values and terms used at project level. 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.

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

To facilitate data validation and reuse, I will provide a detailed README file outlining the structure and transformation of the data, including the methodology and any assumptions made during analysis. Additionally, I will include scripts and explanations on how the analysis was performed, ensuring that the results are reproducible by others.

2b Data quality control

The following data quality checks will be done: standardised data capture, and data entry validation.

3. 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 the project manager in cooperation with the responsible representative of TU.it. The infrastructure of TU Wien will be used for this purpose.

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
R1	writing	writing	reading only
R2	writing	writing	reading only

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

4. 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 ownership

There are no legal restrictions on the processing and disclosure of our data.

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.

5. 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	restrictions / embargo reasons	estimated publication date	location for publication (repository)	PID	license

Methods or software needed to access and use data: The user will need a software and a data analysis tool such as Python.

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

6. RDM responsibilities and resources

6a RDM-roles and responsibilities

The [PI / data officer XY] 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.

Cost name	Cost type	Description	Unit	Value
Estimated total costs				**0**

Coverage of costs

...