|  |  |
| --- | --- |
| Project | |
| Project Number | [1] |
| Project Acronym | [MOSD-Exam-DW01] |
| Project Name | [Investigating the correlation of positivity rate and number of tests in Germany during Covid] |

|  |  |
| --- | --- |
| Data Management Plan | |
| Date | [01.08.2024] |
| Version | [1] |

## Data Summary

This project reuses existing data provided by the European Centre for Disease Prevention and Control Data and can be found under <https://www.ecdc.europa.eu/en/publications-data/archive-historical-data-testing-volume-covid-19>. The dataset is provided in the excel-format and contains the columns country, country code, year week, level, region, region name, new cases, tests done, population, testing rate, positivity rate and testing data source. It takes up 1.2 MB of memory.

The project includes statistical analyses of the dataset and important for that were the columns country, year week, level, tests done and positivity rate. The columns country and level were only used to filter out data from countries other than “Germany” and data that is not at a national level. The column year week was reshaped to 3 columns: “time”, “week” and “year”. “Time” contains a number derived from the week and year information that allows for comparison of dates, “week” contains just the week and “year” contains just the year. Afterwards, all unnecessary columns, the filtered-out rows and rows with NaN-values were omitted to create a new dataset.

The resulting dataset ranges from 2020, week 10 to 2022, week 33 and does not miss a single week within that timeframe. It is saved in a csv format and only takes up 7 KB of memory.

## FAIR data

### **Making data findable, including provisions for metadata**

Both the preprocessed and original dataset are provided on GitHub (https://github.com/NeWildeSache/MoSD-Exam), giving them a unique identifier.

Metadata of the original dataset is provided by the publisher and can be found in the repository as well as machine-readable metadata. For the preprocessed dataset, machine-readable and human-readable metadata files will be created and will improve the findability by providing important keywords.

Additionally, the repository will be synced to a research data repository (Zenodo) to further improve findability. Zenodo also provides a unique and persistent DOI for the repository.

### Making data accessible

GitHub.com is a website used by millions of users and therefore a trusted destination for people looking for data. The repository there will be publicly available for everyone and can be downloaded using ssh, https or just with a regular browser. These methods are free and universally implementable. The same goes for the synced repository on Zenodo.

Also, a recognisable file name will be chosen that allows people to know the general contents of the data file without having to open it.

### Making data interoperable

The data is provided in a standard csv-format which is very commonly used in the scientific community and supported in many statistical analysis tools and packages.

All metadata will be provided in human-readable (md/pdf) and machine-readable (xml) formats and the machine-readable metadata will follow the standard provided by the “dublincore-generator”.

### Increase data re-use

Preprocessing procedures and the statistical analyses are saved in jupyter notebooks using python code and can be repeated using information provided in a README file. This will include environment specifications and installation and running instructions.

An MIT non-copyleft license for the code files and a CC BY license for the data will be provided. Therefore, anyone can re-use the code and data however they want.

Provenance information about the data will be provided in a markdown file and the quality of the data is examined in a separate jupyter notebook. Metadata will describe the dataset in high level of detail.

The repository will also be archived to further convey that the dataset only contains historical data and will not be updated in the future. Archiving a repository means that it will be read-only.

## Other research outputs

Plots derived from the statistical analyses will be saved in the common png format. The treatment of code files was already explained in the chapter reusability.

## Allocation of resources

The original dataset could be downloaded for free and GitHub repositories of this size are free. Therefore, no costs will be tied to publicly providing the data.

As long as the owner of the repository – Dominic Wild – does not delete his account, the repository should be accessible. As of today, GitHub adds public repositories to their GitHub Archive Program, meaning that they will not be deleted. Furthermore, successor settings can be made to migrate the repository to another account in case of account deletion.

## Data security

GitHub provides solutions regarding the backups of repositories.

## **Ethics**

As there was no personal or sensitive data within the original dataset, no measures regarding ethics need to be taken.

## Other issues

/