A5.1

Horizon Europe
Data Management Plan

22 February 2023

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.4.4 (dsw:root:2.4.4).

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
There are no named versions		

Contributors

The following contributors are related to the project of this DMP:

- **Ziyi Wang**
  Role: *Researcher*

Projects

We will be working on the following project and for those are the data and work described in this DMP.

Explore Top-500-cost films

Acronym: *N/A*

Start date: *2023-02-15*

End date: *2023-02-22*

Funding: *: grant number not yet given (planned)*

This dataset is Top-500-cost films. In this project, I try to find out the relation between production cost and genre. Is there a specific type of cost higher than others?

1. Data Summary

Re-used datasets

We have found the following reference datasets that we have considered for re-use:

- **Top-500-movies**

  The dataset can be used in the provided format without any conversion needed.

  The original dataset will be available both from the provider and from us together with our results for the reproducibility.

  We will use the dataset as follows: explore.

We have found the following non-reference datasets that we have considered for re-use:

- **Top-500-movies**

  The dataset can be used in the provided format without any conversion needed.

  We will download or get a copy.

  It is a fixed dataset, changes will not influence reproducibility of our results.

  We will use the complete dataset.

  We will use the dataset as follows: for analyzing.

There is no need to harmonize different sources of existing data in our case.

Data formats and types

We will be using the following data formats and types:

- **CSV**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **Top-500-movies** (not published)

There are no 'Minimal Metadata About ...' (MIA...) standards for our experiments. However, we have a good idea of what metadata is needed to make it possible for others to read and interpret our data in the future.

We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open immediately.

Limited embargo will not be used as all data will be opened immediately.

Metadata will be openly available including instructions how to get access to the data. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

For our produced data, conditions are as follows:

- **Top-500-movies** (not published)

2.3. Making data interoperable

We will be using the following data formats and types:

- **CSV**

   It is a standardized format.

2.4. Increase data re-use

The metadata for our produced data will be kept as follows:

- **Top-500-movies** (not published)

As stated already in Section 2.2, all of our data can become completely open immediately.

We do not plan to be archiving data (using so-called *cold storage*) for long term preservation already during the project.

To validate the integrity of the results, the following will be done:

- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. The management and planning of other research outputs is done separately and is included as appendix to this DMP. Still, we benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

To execute the DMP, no additional specialist expertise is required.

We do not require any hardware or software in addition to what is usually available in the institute.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

6. Ethics

For the data we produce, the ethical aspects are as follows:

- **Top-500-movies**
  - It does not contain personal data.

For reused non-reference datasets, the consent for privacy sensitive data will be solved as follows:

- **Top-500-movies**

  None of the data from this dataset is personal data.

7. Other issues

We use the [Data Stewardship Wizard](#) with its *Common DSW Knowledge Model* (ID: dsw:root:2.4.4) knowledge model to make our DMP. More specifically, we use the https://researchers.ds-wizard.org DSW instance where the project has direct URL: https://researchers.ds-wizard.org/projects/9f195f3f-223a-4e97-9d6c-7d1f53e56edb.

We will be using the following policies and procedures for data management:

- **Github**
  https://github.com/Gaddafish/A5.1-new
  I will use github to present and reserve my research.