Data Management Plan

Analysis of the male/female ratio of actors in the top 500 TMDb movies

Contact person: ***Ze Yu*** *(z.yu.9@student.rug.nl)*

Based on: *Common DSW Knowledge Model, 2.3.0 (dsw:root:2.3.0)*

Project phase: *After Finishing the Project*

Created by: ***Ze Yu*** *(z.yu.9.nl@gmail.com)*

Generated on: *14 Feb 2022*

Projects

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

How does the male/female ratio compare among the top 500 TMDB movies?

Start date: *25 Nov. 2021*

End date: *12 Jan. 2022*

Funding: *Rijksuniversiteit Groningen: grant number not yet given (granted)*

We are analysing the male to female ratio of actors in the top 500 films of all time in order to see patterns of change and compare our findings between gender, film genre, year, production countries. Our data will derive from a dataset created via the TMDB API for which we will create the code to run the API and retrieve the following data on the top 100 rated movies of TMDB: movie title, budget, revenue, vote average, vote count, popularity, genre, release date, production country, production company, number of male actors in top 6 billed roles, number of female actors in top 6 billed roles, non-binary actors in top 6 billed roles.

Section A: Data Collection

1. What data will you collect or create?

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values**

  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. How will the data be collected or created?

There will be no instrument dataset in this project.

Storage and file conventions

We will use a filesystem with files and folders. We have made appointments about naming the files.

We will not be storing data in an "object store" system.

We will not use a relational database system to store project data.

We will not use a graph database for data in the project.

We will not be storing data in a triple store.

Section B: Documentation and Meta-data

3. What documentation and meta-data will accompany the data?

List of data to be published is given in Section E, Question 9. This also includes information about catalogs where the data can be found. Information about data types used is given in Section A, Question 1.

Section C: Ethics and Legal Compliance

4. How will you manage any ethical issues?

Data we collect

Our work on personal data can be done without consent using another legal base – legitimate interests.

5. How will you manage copyright and Intellectual Property Rights (IPR) issues?

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

All of our data can become completely open immediately.

Our data is legally not copyrightable, there is no legal owner.

Section D: Storage and Backup

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

Storage needs will be the same during the whole project.

The work space provides sufficient guarantees in terms of preventing a total loss of data. All project data stored outside of the working area will be adequately backed up.

7. How will you manage access and 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 data centers where project data is stored carry sufficient certifications. All project web services 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.

Only all project members have read/write access to the data.

Section E: Selection and Preservation

8. Which data are of long-term value and should be retained, shared, and/or preserved?

We plan to produce the following datasets:

- **Top 500 TMDB Movies and The Gender Distribution of Actors** (published) – This data set will be kept available as long as technically possible.

9. What is the longterm preservation plan for the dataset?

- **Top 500 TMDB Movies and The Gender Distribution of Actors** (published)
  The distributions will be stored in:
  - Domain-specific repository: GitHub. We don't need to contact the repository because it is a routine for us.
  - Domain-specific repository: Open Science Framework. We don't need to contact the repository because it is a routine for us.

None of the used repositories charge for their services.

Section F: Data Sharing

10. How will you share the data?

- **Top 500 TMDB Movies and The Gender Distribution of Actors**
  The dataset has the following identifiers:
  - URL: **https://docs.google.com/spreadsheets/d/1GoCQxoMxycjpq36yPrtVB08**

  The distributions will be available as follows:
  - Open (shared with anyone) using a domain-specific repository: Open Science Framework. The distribution will be available under the following license:
    - Starting 14 February 2022: Freely available for any use (public domain or CC0).

Information about used repositories (i.e. where will potential users find out about the data) is provided in Section E, Question 9.

Embargo on the data is described in Section C, Question 5, and Section F, Question 11.

11. Are any restrictions on data sharing required?

Ethical and legal restrictions are documented under Section C. We have used the Data Stewardship Wizard, which made us aware of options to minimize the restrictions.

No data sharing agreement will be required.

Section G: Responsibilities and Resources

12. Who will be responsible for data management?

Ze Yu is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Raya Allawy is responsible for finding, gathering, and collecting data.

Lena Ermolaeva is responsible for the management and proficiency of data including data processing, data policies, data guidelines, and data availability.

13. What resources will you require to deliver your plan?

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

Charges applied by data repositories (if any) are mentioned already in Section E, Question 9.