Analysis on the Rating Distribution in Netflix

Horizon Europe
Data Management Plan

26 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**
Rating Distribution by Film and TV shows in Netflix0226	26 Feb 2023	Update the tools

Contributors

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

- **Haozhe Bai**
  h.bai.3@student.rug.nl
  Roles: *Contact Person*, *Data Collector*, *Data Manager*, *Project Leader*, *Project Manager*
  Affiliation: *University of Groningen*

Projects

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

Analysis on the Rating Distribution in Netflix

Acronym: *RDN*

Start date: *2022-11-24*

End date: *2023-03-03*

Funding: *University of Groningen: grant number not yet given (planned)*

I wanted to study the number of releases of different rated films in Netflix to get a deeper understanding of the market focus of the streaming video site. It will compare the video output of Netflix TV series as well as movies in different ratings.This study can help us learn more about the development of adult-oriented video content in this area.The database was sourced from the Kaggle website.

1. Data Summary

Non-equipment datasets

We also collect data from electronic patient records. The non-equipment datasets are:

- **netflix_titles.csv** – This data set was created to list all shows available on Netflix streaming, and analyze the data to find interesting facts. This data was acquired in July 2022 containing data available in the United States.

Re-used datasets

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

- **Kaggle Datasets** (https://www.kaggle.com/datasets/shivamb/netflix-shows) ✔

  Owner of this dataset: SHIVAM BANSAL https://www.shivambansal.com.

  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: I will use this reference dataset to make some adjustments and additions to the dataset to improve its original missing parts.

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

- **Kaggle Datasets** (https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney) ✔

  Owner of this dataset: Ruchi Bhatia https://twitter.com/ruchi798.

  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: It can help me to have a clearer understanding of the data on various platforms, so that I can determine the direction of my research and select the type of data.

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:

- **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. FAIR Data

2.1. Making data findable, including provisions for metadata

- **Netflix TV Shows and 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 use an electronic lab notebook to make sure that there is good provenance of the data analysis.

We made a SOP (Standard Operating Procedure) for file naming. I will name it based on the content which i want to search from this file. 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).

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

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **Kaggle Datasets** – freely available for any use (public domain or CC0).
- **Kaggle Datasets** – freely available for any use (public domain or CC0).

For our produced data, conditions are as follows:

- **Netflix TV Shows and Movies** (not published)

2.3. Making data interoperable

We will be using the following data formats and types:

- **Comma-separated Values**

  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- **The Data Use Ontology**

2.4. Increase data re-use

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

- **Netflix TV Shows and Movies** (not published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.

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

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

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

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will be instrumenting the tools into pipelines and workflows using automated tools.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- 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.

We will be archiving data (using so-called 'cold storage') for long term preservation already during the project.

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication. For making data or other research outputs FAIR, we budgeted: The budget for making data or other research outputs more fair is about 100euro.

Haozhe Bai is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Haozhe Bai is responsible for finding, gathering, and collecting data.

Haozhe Bai is responsible for maintaining the finished resource.

To execute the DMP, additional specialist expertise is required and we have such trained support staff available.

We require the following hardware or software in addition to what is usually available in the institute: Python Anaconda.

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 data centers where project data is stored carry sufficient certifications. 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.

All personal information will be processed in pseudonymized form only. We pseudonymize inside the project, only limited people can access the keys.

The archive will be stored in a remote location to protect the data against disasters. The archive need to be protected against loss or theft. It is clear who has physical access to the archives.

6. Ethics

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

- **Netflix TV Shows and Movies**
  - It contains personal data.
  - It contains sensitive data.

Data we collect

We will collect data connected to a person, i.e. "personal data". We are legally obliged to do this data processing (i.e. legal requirement). An ethical committee will make an ethical review on the project. We need to conduct a data protection impact assessment (DPIA).

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

- **Kaggle Datasets**

  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/8ea65ea7-2bce-48f1-9305-241e0d9beaa2.

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

- **RUG Policy**

https://www.rug.nl/research/research-data-management/policy/ug-rdm/
Because I should follow the policy of my university. The purpose of policy is to ensure innovative research and research integrity.