Collecting Data Group E Project

Horizon Europe
Data Management Plan

18 January 2024

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.6.3 (dsw:root:2.6.3).

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

Contributors

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

- **Huang Jiaqi**
  j.huang.30@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Zhang Hanyue**
  zhy19990312@outlook.com
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Zhang Qimeng**
  kikiknowz@gmail.com
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Che Zicheng**
  ZichengChe@outlook.com
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Wang Mengyun**
  m.wang.43@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Wang Jingyi**
  j.wang.113@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:

University of Groningen type Education

- **Cao Yaru**
  yarucao@gmail.com
  Roles: *Data Collector*, *Data Curator*, *Data Manager*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

Projects

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

REVIEW EVOLUTION: A COMPARATIVE ANALYSIS OF 'THE DARK KNIGHT' AND 'INFERNAL AFFAIRS' REVIEWS

Acronym:

REA-CA: TDK & IA REVIEWS

Start date:

2023-12-10

End date:

2024-01-18

Funding:

Did not apply for any funding yet.

The aim of this study is the thematic changes and emotional trends of film criticism in a certain period of time, so 2001-2010 is selected as the film release interval. Within this interval, this study considers combining the awards won by films and the ratings of platform viewers, which respectively represent the recognition of films in the industry and the views and preferences of general audiences on films. After screening, two films, The Dark Knight and Infernal Affairs, are selected as research objects. Among them, The Dark Knight was the film with the highest rating on imdb platform among all the films that won various awards of the Academy Awards during this period, and Infernal Affairs was the film with the highest rating on Douban platform among all the films that won various awards of the Hong Kong Film Awards.

Research Question Explore the evolving focus of film reviews over time.

1. Data Summary

Non-equipment datasets

The non-equipment datasets are:

- **oscars_awards_2001_2010** – Oscar-winning movie information data set from 2001 to 2010.
- **2001_2010_HongKongFilmAwards_WinningWorks** – Hong Kong Film Awards winning film information data set from 2001 to 2010
- **Douban_InfernalAffairs_reviews** – From 2005 to 2023, movie reviews on Douban about Infernal Affairs.
- **The_Dark_Knight** – From 2005 to 2023, movie reviews on Douban about The dark knight.
- **IMDb_thedarknight_reviews** – From 2001 to 2023, movie reviews on IMDB about The dark knight.
- **IMDb_InfernalAffairs_reviews** – From 2001 to 2023, movie reviews on IMDB about Infernal Affairs.

Re-used datasets

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

- The Oscar Award, 1927 - 2023
  It is available via: https://www.kaggle.com/datasets/unanimad/the-oscar-award. It is used in the project.

  Owner of this dataset: RAPHAEL FONTES & David Lu .

  We will first need to convert the format before using it.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: Mainly used to extract the list of Oscar winning films from 2001 to 2010.

- IMDB 5000 Movie Dataset
  It is available via: https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset. It is used in the project.

  Owner of this dataset: Yueming .

  We will first need to convert the format before using it.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: It mainly matches the table of Oscar-winning movies from 2001 to 2010, and extracts the ratings of the winning movies on IMDB.

- Hong Kong Film Award data from 2001 to 2010
  It is available via: https://zh.wikipedia.org/wiki/%E9%A6%99%E6%B8%AF%E9%9B%BB%E5%BD%B1%E9%87%91%E5%83%8F%E7%8D%8E.
  It is used in the project.

  Owner of this dataset: Wikipedia.

  We will first need to convert the format before using it.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: Extract the list of film winners from the Hong Kong Film Awards from 2001 to 2010.

- Hong Kong Film Awards
  It is available via: https://www.hkfaa.com/winnerlist.html. It is used in the project.

  Owner of this dataset: Hong Kong Film Awards hkfaa@hkfaa.com.

  We will first need to convert the format before using it.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: Manually cross calibrate and reference the award-winning movie information dataset obtained from Wikipedia to ensure the accuracy of the list of award-winning movies.

We will need to harmonize different sources of existing data.

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values** (CSV) type model and format

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  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.

- TXT
  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

- **filtered_oscars_awards_2001_2010 / oscars/ 2001_2010_HongKongFilmAwards_WinningWork/ Douban_InfernalAffairs_reviews/ The_Dark_Knight / IMDb_thedarknight_reviews /IMDb_InfernalAffairs_reviews** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/RiverCho/CD_FinalProject_GroupE.git**

  We will distribute the dataset using:
  - *Special-purpose repository for the project*. The repository will provide download-only service.

  There won't be different versions of this data over time.
  We will not be adding a reference to any data catalogue because the data will be stored in a repository that is the prime source of data for re-use in the field.

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. Filtered and cleaned dataset saved as 'filtered_oscars_awards_2001_2010.csv' the filtered and processed csv file "oscars.csv" . 2001_2010_hongkongfilmawards_winningworks.csv douban_infernalaffairs_reviews.csv the_dark_knight.csv imdb_thedarknight_reviews.csv imdb_infernalaffairs_reviews.csv. 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

The data cannot become completely open because of:

- legal reasons
- non-patent business reasons: Due to the fact that the core dataset is comment data on movie platforms, the copyright of comments belongs to individual publishers and we are unable to provide it. The copyright of other datasets related to award-winning movies' information does not belong to anyone, and we can provide them.

Limited embargo cannot be used because some restricted data will be embargoed indefinitely.

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:

- The Oscar Award, 1927 - 2023
  It is freely available for any use (public domain or CC0).
- IMDB 5000 Movie Dataset
  It is available under specific restrictions, which we will follow in our project: 2.0 RIGHTS GRANTED AND CONDITIONS OF USE 2.1 Rights granted. The Licensor grants to You a worldwide, royalty-free, non-exclusive, perpetual, irrevocable copyright license to do any act that is restricted by copyright over anything within the Contents, whether in the original medium or any other. These rights explicitly include commercial use, and do not exclude any field of endeavour. These rights include, without limitation, the right to sublicense the work. 2.2 Conditions of Use. You must comply with the ODbL. 2.3 Relationship to Databases and ODbL. This license does not cover any Database Rights, Database copyright, or contract over the Contents as part of the Database. Please see the ODbL covering the Database for more details about Your rights and obligations. 2.4 Non-assertion of copyright over facts. The Licensor takes the position that factual information is not covered by copyright. The DbCL grants you permission for any information having copyright contained in the Contents. 3.0 WARRANTIES, DISCLAIMER, AND LIMITATION OF LIABILITY 3.1 The Contents are licensed by the Licensor “as is” and without any warranty of any kind, either express or implied, whether of title, of accuracy, of the presence of absence of errors, of fitness for purpose, or otherwise. Some jurisdictions do not allow the exclusion of implied warranties, so this exclusion may not apply to You. 3.2 Subject to any liability that may not be excluded or limited by law, the Licensor is not liable for, and expressly excludes, all liability for loss or damage however and whenever caused to anyone by any use under this License, whether by You or by anyone else, and whether caused by any fault on the part of the Licensor or not. This exclusion of liability includes, but is not limited to, any special, incidental, consequential, punitive, or exemplary damages. This exclusion applies even if the Licensor has been advised of the possibility of such damages. 3.3 If liability may not be excluded by law, it is limited to actual and direct financial loss to the extent it is caused by proved negligence on the part of the Licensor.
- Hong Kong Film Award data from 2001 to 2010
  It is freely available for any use (public domain or CC0).
- Hong Kong Film Awards
  It is freely available for any use (public domain or CC0).

For our produced data, conditions are as follows:

- **filtered_oscars_awards_2001_2010 / oscars/ 2001_2010_HongKongFilmAwards_WinningWork/ Douban_InfernalAffairs_reviews/ The_Dark_Knight / IMDb_thedarknight_reviews /IMDb_InfernalAffairs_reviews** (published)
  The distributions will be accessible through:
  - *Special-purpose repository for the project*. It will be *Open* (shared with anyone). The repository will provide download-only service. The distribution will be available under the following license:
    - Freely available for any use (public domain or CC0).

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.

2.3. Making data interoperable

We will be using the following data formats and types:

- **Comma-separated Values** (CSV) type model and format

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  It is a standardized format.

- TXT
  It is a standardized format.

2.4. Increase data re-use

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

- **filtered_oscars_awards_2001_2010 / oscars/ 2001_2010_HongKongFilmAwards_WinningWork/ Douban_InfernalAffairs_reviews/ The_Dark_Knight / IMDb_thedarknight_reviews /IMDb_InfernalAffairs_reviews** (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 explained in Section 2.2, our data cannot become completely open.

There are no IP reasons why our data can not be open.

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 a subset of our jobs several times across the different compute infrastructures.
- 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.

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: 1 month.

Huang Jiaqi, Zhang Hanyue, Zhang Qimeng, Che Zicheng, Wang Mengyun, Wang Jingyi, and Cao Yaru are responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Huang Jiaqi, Zhang Hanyue, Zhang Qimeng, Che Zicheng, Wang Mengyun, Wang Jingyi, and Cao Yaru are responsible for finding, gathering, and collecting data.

Huang Jiaqi, Zhang Hanyue, Zhang Qimeng, Che Zicheng, Wang Mengyun, Wang Jingyi, and Cao Yaru are responsible for maintaining the finished resource.

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

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 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 data will be collected anonymously.

We are not running the project in a collaboration between different groups nor institutes. Therefore, no collaboration agreement related to data access is needed.

6. Ethics

Data we produce

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

- **filtered_oscars_awards_2001_2010 / oscars/ 2001_2010_HongKongFilmAwards_WinningWork/ Douban_InfernalAffairs_reviews/ The_Dark_Knight / IMDb_thedarknight_reviews /IMDb_InfernalAffairs_reviews**
  - It contains personal data.
  - It does not contain sensitive data.

Data we collect

We will not collect any data connected to a person, i.e. "personal data".

The data collection is subject to ethical legislation. It is covered by ethical review. It involves human subjects.

7. Other issues

We use the Data Stewardship Wizard with its *Common DSW Knowledge Model* (ID: dsw:root:2.6.3) knowledge model to make our DMP. More specifically, we use the https://researchers.ds-wizard.org/wizard DSW instance where the project has

direct URL: https://researchers.ds-wizard.org/wizard/projects/31f6fcbc-dfe4-4f99-a175-3d14f16845ee.

We will not be using any extra national, funder, sectorial, nor departmental policies or procedures for data management.