Anime Recommendation System and Cultural Analysis

Horizon Europe
Data Management Plan

20 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:

- **Yingyue Jiang**
  y.jiang.34@student.rug.nl
  Roles: *Contact Person*, *Data Collector*, *Editor*, *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.

Anime Recommendation System and Cultural Analysis

Acronym:

ARS

Start date:

2023-11-20

End date:

2024-01-20

Funding:

Rijksuniversiteit Groningen *(The Netherlands)*

: grant number not yet given (planned)

This Anime Project aims to create a comprehensive recommendation system for anime based on user preferences on myanimelist.net. Apart from personalized recommendations, this project delves into the broader impact of anime on global cultures and societies. We aim to investigate evolving themes, genres, and storytelling trends within the anime industry over time through data-driven analysis.

1. Data Summary

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.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **Anime Recommendations Database** (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.

The provenance will be captured using W3C PROV.

We made a SOP (Standard Operating Procedure) for file naming. By adopting a systematic approach to file and folder naming, we will divide into 5 types of names, including anime, code, report, deliverables, and community engagement. The examples are showing as follows:
(1)anime_data: "naruto_crunchyroll_20230101.csv"
(2)code: "datacleaning_python_v1.0.py"
(3)reports: "weeklyprogress_animetrends_20230107.pdf"
(4)deliverables: "paper_culturalimpactofanime_20230215.docx"
(5)community_engagement: "survey_twitter_20230310.pdf"
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 over time.

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 be available in a form that can be harvested and indexed

(managed by the used repository / repositories). For our produced data,

conditions are as follows:

- **Anime Recommendations Database** (not published)
liscense: https://creativecommons.org/licenses/by/4.0/

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.

2.4. Increase data re-use

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

- **Anime Recommendations Database** (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 over time.

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 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.

Yingyue Jiang is responsible for finding, gathering, and collecting data.

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

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 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.

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:

- **Anime Recommendations Database**
  - It does not contain 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".

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/6d18041d-071d-45ae-b784-00dad92331c7.

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/
  Its purpose is to ensure innovative research and research integrity.