r/Feminism

Horizon Europe
Data Management Plan

23 January 2023

Data Management Plan created in Data Stewardship Wizard

History of changes

There are no named versions.

Contributors

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

- Weena Moulder
  w.moulder@student.rug.nl
  Roles: Data Collector, Researcher
- Jeli Haagsma
  j.s.haagsma@student.rug.nl
  Roles: Data Collector, Data Manager, Researcher
- Jing Li
  j.li.88@student.rug.nl
  Roles: Data Collector, Researcher
- Shanshan Liu
  s.liu.48@student.rug.nl
  Roles: Researcher
- Xi Yang
  x.yang.34@student.rug.nl
  Roles: Data Collector, Researcher
- Federico Pianzola
  f.pianzola@rug.nl, https://orcid.org/0000-0001-6634-121X
  Roles: Supervisor

Projects

We will be working on the following project. For those anyone who is interested in the project are the data and work described in this DMP.

Collecting data from r/Feminism

Acronym
N/A

Start date
2022-12-16

End date
2023-01-24

Funding
Did not apply for any funding yet.

This project’s aim is to collect data from the r/Feminism subreddit by using the Python Reddit Api Wrapper (PRAW). By doing this it will allow us to practice and develop our coding skills, how to construct a data management plan as well as using platforms such as Flourish to explore our dataset and practice data visualization. Furthermore, the dataset from this project will allow us to analyse in another project for the Tools and Methods course at the University of Groningen.

1. Data Summary

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 expect to have 0.000495 GB of data in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **feminism-reddit dataset.csv** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/Jeli-sh/MA-Digital-Humanities-group-project-Collecting-Data.git**

  We will distribute the dataset using:
  - *Domain-specific repository*: GitHub. We don't need to contact the repository because it is a routine for us.
    A persistent identifier will be assigned by the repository. The repository will make sure that the persistent identifier can be resolved to a digital object. The assigned persistent identifier is specified: https://github.com/Jeli-sh/MA-Digital-Humanities-group-project-Collecting-Data.git.

  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.

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 without 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).

All data will be owned by the institute.

For our produced data, conditions are as follows:

- **feminism-reddit dataset.csv** (published)
  The distributions will be accessible through:
  - *Domain-specific repository*: GitHub. We don't need to contact the repository because it is a routine for us. The distribution will be available under the following license:
    - Starting 2023-01-24: 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 as soon as possible after collecting it.

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):

-

2.4. Increase data re-use

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

- **feminism-reddit dataset.csv** (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 after the project but also already during the project. The minimum lifetime of the archive is 5 years. Data formats of data in cold storage will be upgraded if they become obsolete. Archived data will be migrated regularly to more modern storage media (e.g. newer tapes).

None of the used repositories charge for their services.

W. Moulder is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Weena Moulder, Jeli Haagsma , Jing Li , and Xi Yang are responsible for finding, gathering, and collecting data.

Jeli Haagsma is responsible for maintaining the finished resource, creating tutorials for others and develop active learning exercises.

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

We require the following hardware or software in addition to what is usually available in the institute: Python Reddit Api Wrapper (PRAW) to scrape data from r/Feminism.

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 risk of information loss in the project or organization is acceptably low. The risk of information leak in the project or organization is acceptably low. The risk of information vandalism in the project or organization is acceptably low.

We are not using any personal information.

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:

- **feminism-reddit dataset.csv**
  - 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". The usernames are left out of our data collection.

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/8549b232-0a30-4869-9858-978939efcedf.

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