“Must-Read Content” of Incels Forum

# **Horizon Europe** **Data Management Plan** 08 January 2024

## History of changes

There are no named versions.

## Contributors

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

* Eleni Paipeti  
  [E.Paipeti@student.rug.nl](mailto:E.Paipeti@student.rug.nl)  
  Roles: Researcher
* Harm Bredewold  
  [H.Bredewold@student.rug.nl](mailto:H.Bredewold@student.rug.nl)  
  Roles: Data Collector
* Jan Czechowicz  
  [J.J.Czechowicz@student.rug.nl](mailto:J.J.Czechowicz@student.rug.nl)  
  Roles: Researcher
* Sijie Qiu  
  [S.Qiu.6@student.rug.nl](mailto:S.Qiu.6@student.rug.nl)  
  Roles: Researcher
* Xinyi Song  
  [X.Song.6@student.rug.nl](mailto:X.Song.6@student.rug.nl)  
  Roles: Researcher
* Yuhan Si  
  [Y.Si.4@student.rug.nl](mailto:Y.Si.4@student.rug.nl)  
  Roles: Researcher
* Yuning Xie  
  [Y.Xie.19@student.rug.nl](mailto:Y.Xie.19@student.rug.nl)  
  Roles: Researcher

## Projects

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

### “Must-Read Content” of Incels Forum

Acronym

N/A

Start date

2023-12-14

End date

2024-01-

Funding

Did not apply for any funding yet.

Incels is a forum for men who struggle with or are unable to get into romantic relationships with women despite trying. Our project wants to provide a dataset about the “Must-Read Content” section of this forum. This dataset could be used to analyze the prevailing themes and sentiments within these discussions. We believe that studying manifestations of incel ideology is important, as it is gaining popularity and may lead to serious violence, putting public health and safety at risk.

## 1. Data Summary

#### Non-equipment datasets

We collect data from Incels.is (https://incels.is/). The non-equipment dataset is:

* **incels\_forum\_data** – This dataset contains the the Title, Tags, Poster, Date, number of Views and Replies, and Content of the post in the “Must-Read Content” section of the Incels forum.

#### Data formats and types

We will be using the following data formats and types:

* **Comma-separated Values** (CSV)
* 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

* **incels\_forum\_data.csv**
* 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.

The dataset has the following identifiers:

<https://github.com/hbredewold/GROUP_G_Collecting_Data/blob/main/incels_forum_dat> a.csv

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

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 lab notebooks to make sure that there is good provenance of the data analysis.

We made a SOP (Standard Operating Procedure) for file naming. 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).

### 2.3. Making data interoperable

We will be using the following data formats and types:

* **Comma-separated Values** (CSV)
* It is a standardized format.

### 2.4. Increase data re-use

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 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. Data formats of data in cold storage will not be upgraded over time. Archived data will not be migrated to other storage media over time.

None of the used repositories charge for their services.

Harm Bredewold 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

Project members can carry data with them on password-protected laptops. 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 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.

The archive will be stored in a remote location to protect the data against disasters.

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 collect

We will not collect any data connected to a person, i.e. "personal data". We will keep the poster’s name anonymous

## 7. Other issues

We use the [Data Stewardship Wizard](https://ds-wizard.org) 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/9228f677-2b30-4a80-8f24-69449db728c5>.

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

* **UG Research Data Policy**  
  <https://www.rug.nl/digital-competence-centre/ug-research-data-policy-2021.pdf>  
  This project is for the course Collecting Data in the University of Groningen. The Research Data policy of the University of Groningen is a general framework outlining basic principles and responsibilities for dealing with data that is usable for research that can be published or exploited.