“Must-Read Content” of Incels Forum

Horizon Europe

Data Management Plan

17 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

  Roles: Researcher
- Harm Bredewold

  H.Bredewold@student.rug.nl

  Roles: Data Collector, Data Manager
- Jan Czechowicz

  J.J.Czechowicz@student.rug.nl

  Roles: Researcher
- Sijie Qiu

  S.Qiu.6@student.rug.nl

  Roles: Researcher
- Xinyi Song

  X.Song.6@student.rug.nl

  Roles: Researcher
- Yuhan Si

  Y.Si.4@student.rug.nl

  Roles: Researcher
- Yuning Xie

  Y.Xie.19@student.rug.nl

  Roles: Researcher

Projects

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

“Must-Read Content” of Incels Forum

Start date

2023-12-14

End date

2024-01-18

This Data Management Plan (DMP) outlines the methodology and ethical considerations for the project focused on the "Must-Read Content" section of the Incels forum. The term 'Incels,' a portmanteau of 'involuntary celibates,' refers to an online subculture of individuals who express difficulty in finding romantic or sexual partners despite desiring one. This phenomenon, which has garnered increasing academic interest, sits at the intersection of digital sociology, psychology, and gender studies.

Recent research in digital sociology has highlighted the importance of understanding online communities and their impact on societal norms and individual behavior. The study by Moskalenko et al. (2022) in "Incel Ideology, Radicalization and Mental Health" offers a comprehensive examination of the community's sentiments and mental health challenges. This

research suggests a nuanced landscape within the community, where a spectrum of attitudes towards violence and radicalization is present. Similarly, the narrative review from Broyd et al. (2023), presents an in-depth analysis of the diverse behavioral patterns within the incel community. This review emphasizes the critical need for informed risk assessment and clinical intervention strategies, particularly considering the potential escalation of certain extreme ideologies. Moreover, Chan (2023) provides an essential perspective on the role of technology in the propagation of incel ideologies. This research links these ideologies to wider societal issues, highlighting the importance of understanding the digital dimensions of such movements.

Our project aims to contribute to this growing body of knowledge by analyzing the prevailing themes and sentiments within the Incels forum, specifically focusing on its "Must-Read Content" section. The project's objective is to create a comprehensive dataset that facilitates a deeper understanding of the Incels' ideological expressions and discourse patterns. The research questions guiding this project include: What are the central themes present in the "Must-Read Content" of the Incels forum? How do these discussions reflect broader societal attitudes toward gender and relationships? The answers to these questions will provide valuable insights into this digital subculture, contributing to a more nuanced understanding of online communities and their real-world implications.

Methodologically, the project will employ data scraping techniques to collect posts from the forum, followed by a content analysis. The dataset, primarily composed of titles, tags, dates, view

counts, replies, and post content, will be stored in a Comma-separated Values (CSV) format for accessibility and long-term preservation.

As we embark on this project, we are mindful of the delicate balance between rigorous academic inquiry and the ethical responsibilities inherent in dealing with such sensitive topics. We attempt to contribute meaningfully to the understanding of this community while maintaining the highest standards of research ethics and data privacy. The users' discourse often involves language that is misogynistic, aggressive, and potentially harmful. This presents distinct challenges in our research, especially regarding ethical considerations, data sensitivity, and the handling of triggering or offensive content. Navigating these challenges requires a careful and considered approach. Firstly, it is essential to address the potential exposure to sensitive personal data. Our adherence to the General Data Protection Regulation (GDPR) and other relevant data protection laws ensures that personal information is treated with the utmost confidentiality and respect. Secondly, the nature of the language used within these communities necessitates a robust method for handling and analyzing triggering words and hate speech. This involves not only technical solutions for data filtering and analysis but also a thoughtful consideration of the impact such language can have on both researchers and potential audiences.

In accordance with the University of Groningen's Research Data policy (University of Groningen, 2021), this DMP demonstrates our commitment to responsible data management, emphasizing the FAIR (Findable, Accessible, Interoperable, and Reusable) principles throughout the project lifecycle.

1. Data Summary

Non-equipment datasets

The non-equipment datasets are:

- **incels_forum_data.csv** – This dataset contains the Date, Title, Tags, number of Views and Replies, the link, sentiment score, and the length of each post in the "Must-Read Content" section of the Incels forum (https://incels.is/forums/must-read-content.23/).

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** (published)

  The dataset has the following identifiers:

  – URL:

  https://github.com/hbredewold/Collecting_Data_GroupG/blob/main/incels_forum_data.csv

  We will distribute the dataset using:

  – *Domain-specific repository*:

  GitHub (GitHub)

  We don't need to contact the repository because it is 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/hbredewold/Collecting_Data_GroupG/blob/main/incels_forum_data.csv

  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 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 over time.

Limited embargo will not be used as all data will be opened immediately.

Metadata will be openly available including instructions on 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:

- **incels_forum_data.csv** (published)

  The distributions will be accessible through:

  – *Domain-specific repository*:

  GitHub (GitHub)

  We don't need to contact the repository because it is routine for us. 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)

  It is a standardized format.

2.4. Increase data re-use

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

- **incels_forum_data.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 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 computing 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 are done separately and are included as an 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 in 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.

Yuhan Si and Xinyi Song are responsible for implementing the DMP, and ensuring it is reviewed and revised.

Harm Bredewold is responsible for maintaining the finished resource.

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 on the project or organization, if information is leaked, is small. The possible impact on 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 or institutes. Therefore, no collaboration agreement related to data access is needed.

6. Ethics

Data we collect

Our project collects data with reference to The General Data Protection Regulation (GDPR) in the EU, which provides cross-country information within the EU on the applicable regulations and guidelines regarding the participation of humans in research. The GDPR defines personal data as any information that relates to an individual who can be directly or indirectly identified. In our project, we will not collect any data connected to a person (e.g. user id, user name, gender, area). We will keep the poster's name anonymous and delete the post content when we present the research results and create the final data set.

Data we produce

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

- Sensitivity of Content:
  - The final dataset will be meticulously reviewed to ensure it does not contain directly identifiable personal data to maintain anonymity.
  - Our team examines the data to identify and appropriately handle content that, while not personal, may be considered sensitive due to its nature, including hate speech or misogynistic language.
- Handling Triggering Content:

- A clear and conspicuous warning is included at the beginning of the README file accompanying the dataset. The advisory will inform users the possibility of being triggered.
- Researchers involved in the study will give their informed consent, acknowledging their understanding of the sensitive nature of the content they will analyze. Our team members will be informed about the availability of the potential for encountering distressing content, and their right to stop from the project at any time they are feeling triggered.

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 a 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 at 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.