Plan Overview

A Data Management Plan created using DMPonline

Title: Resilience in the picture

Creator: Vera Heininga

Principal Investigator: Vera Heininga

Data Manager: Vera Heininga

Project Administrator: Vera Heininga

Affiliation: Rijksuniversiteit Groningen

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan NWO (September 2020)

ORCID iD: 0000-0003-0889-8524

Project abstract:

Up to 50% of Dutch youth (aged 16-30) experience psychological problems such as anxiety and depressive symptoms, and suicide is the leading cause of death among Dutch youth. In this longitudinal experience sampling study, we will investigate to what extent risk factors and protective factors for psychological problems can be derived from photos shared on social media. We will apply scraping techniques and Machine Learning to extract facial expressions from photos - as well as relevant contextual factors (e.g., being with friends) - and test whether these factors can be used to predict youth's resilience or vulnerability to psychological problems.

ID: 134315

Start date: 01-10-2023

End date: 30-09-2024

Last modified: 26-09-2023

Grant number / URL: 406.XS.04.051

Resilience in the picture

General Information

Name applicant and project number

Dr. Vera E. Heininga
406.XS.04.051

Name of data management support staff consulted during the preparation of this plan and date of consultation.

Michiel de Ree, june-sept 2023
Marieke Steeman, june-sept 2023
Marlon de Jong, june-sept 2023

1. What data will be collected or produced, and what existing data will be re-used?

1.1 Will you re-use existing data for this research?

If yes: explain which existing data you will re-use and under which terms of use.

- No

1.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

The project will collect from 100-300 participants numeric answers to one baseline questionnaire (xls), 2x90 photo images (png), and 90x numeric daily diary data (xls).
The prepared dataset will contain 1) characteristics deduced from the BeReal photos and its minimised metadata, 2) baseline and diary questionnaires data from Flycatcher + Random_ID variable. The prepared dataset does not contain photo materials.

1.3. How much data storage will your project require in total?

- 10 – 100 GB

2. What metadata and documentation will accompany the data?

2.1 Indicate what documentation will accompany the data.

After applying FAIR principles to the prepared data, the prepared dataset (i.e., dataset including the characteristics deduced from the BeReal photos, minimised metadata, and questionnaire data; no photos included) will be shared Open Access available through DataverseNL for reuse accompanied with all information on the methodology used to collect the data, analytical and procedural information, and codebooks. The BeReal photos will never be shared. In consultation with the embedded data steward, there will be a de-identification protocol developed.

For the reasons of replication of the results reported in the scientific article, the analysed dataset (i.e., only the variables used for analysis) will be shared Open Access via the journal of publication and the Open Science Framework (OSF) together with its

Rmarkdown code. Furthermore, all information on the methodology used to collect the data, analytical and procedural information, and codebooks will be published Open Access on Open Science Framework (OSF).

2.2 Indicate which metadata will be provided to help others identify and discover the data.

When uploading the data to DataverseNL, the DataCite standard will be used. So, at least: Identifier (with mandatory type sub-property); Creator (with optional given name, family name, name identifier and affiliation sub-properties); Title (with optional type sub-properties); Publisher; PublicationYear; ResourceType (with mandatory general type description subproperty)

3. How will data and metadata be stored and backed up during the research?

3.1 Describe where the data and metadata will be stored and backed up during the project.

- Institution networked research storage

During data collection the raw questionnaire data will be stored on the servers of Flycatcher. The raw (meta)data from BeReal is stored on the servers of the CIT. We will apply minimisation, and Vera Heininga will only receive part of the metadata that is relevant for her research purpose from the CIT.

During data preparation, the data will be stored and backed up on the UG Y drive in a folder dedicated to the research project. Only Vera Heininga, and possibly the postdoc, will have access to the project folder. There will be a separate project folder created for the photo materials to which only Vera Heininga will have access.

During data analysis, the data will be stored and backed up on the UG Y drive in a folder dedicated to the research project. Only Vera Heininga, and possibly the postdoc, will have access to this folder.

3.2 How will data security and protection of sensitive data be taken care of during the research?

- Default security measures of the institution networked research storage

During the data collection process, participants are recommended to use a BeReal user name that does not contain directly identifiable information such as name and/or birth date (see page 5 of “2023 09 24 selectievragenlijst concept 3 - Veerkracht in beeld”)

During the whole data life cycle, Vera Heininga never gets access to directly identifying personal data from Flycatcher (i.e., name, email, phone number). As soon as possible after all data has been collected and at the latest within 6 months, Vera Heininga will receive a data file with the answers to the questions in the Baseline and Daily questionnaires (see also data flow 1). In addition, the data file contains a RANDOM_ID. This is an ID that the respondents receive per survey. This ID remains the same for the different questionnaires within the survey. So in this case, this means that someone who has RANDOM_ID 1 in the selection survey (i.e., the survey sent out by Flycatcher in which they inform and ask participants for consent) will also have RANDOM_ID 1 in the baseline questionnaire and the daily questionnaires. Someone who has RANDOM_ID 2 in the selection study also has RANDOM_ID 2 in the baseline questionnaire and the daily questionnaires. Vera Heininga can therefore link the answers to the different questionnaires within this study based on the RANDOM_ID. This RANDOM_ID does change between different surveys, so the data cannot be linked to data from previous Flycatcher projects or future projects. Flycatcher will delete all personal data at the latest one year after the end of the project.

After the data collection, Vera Heininga will download the photos (i.e., Front facing photo, Back facing photo) and minimised metadata (i.e., BeReal username, lateInSeconds, takenAt, creationDate, Caption, number of smileys and number of reactions by BeReal friends) via unishare from the CIT. Vera Heininga will immediately replace the BeReal user name by the Flycatcher RANDOM_ID in all files and folders (to prevent them being potentially identifiable, even though we will instruct participants to choose a non-identifiable name). During all parts of the data life cycle, the photos will be stored in a separate project folder (i.e., separating the photos from the project folder with all other project information such as questionnaire data). All photo materials will be stored for a maximum of two years.

In the data preparation part of the project, photos will be scored using AI software on characteristics (i.e., happy, neutral, sad facial expressions; being in company of friends; being in nature; being active). The AI software will be run locally on the servers of the CIT. Only the characteristics assigned to the photos will be part of the pseudonymized dataset used for analysis. In addition, the age variable will be minimised into categories (e.g., from 16-20; 21-25 etc.).

Flycatcher staff have access to the questionnaire data and will be controlled by a data processing agreement. The Privacy and Security officers of the BSS faculty and ABJZ are still working on the data processing agreement with Flycatcher - this will be signed before the start of the project. Data Science staff member Michiel de Ree of the CIT has access to the BeReal data.

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?

4.1 Will you process and/or store personal data during your project?

If yes, how will compliance with legislation and (institutional) regulation on personal data be ensured?

- Yes

During the data collection process, participants are recommended to use a BeReal user name that does not contain directly identifiable information such as name and/or birth date (see page 5 of “2023 09 24 selectievragenlijst concept 3 - Veerkracht in beeld”)

During the whole data life cycle, Vera Heininga never gets access to directly identifying personal data from Flycatcher (i.e., name, email, phone number). As soon as possible after all data has been collected and at the latest within 6 months, Vera Heininga will receive a data file with the answers to the questions in the Baseline and Daily questionnaires (see also data flow 1). In addition, the data file contains a RANDOM_ID. This is an ID that the respondents receive per survey. This ID remains the same for the different questionnaires within the survey. So in this case, this means that someone who has RANDOM_ID 1 in the selection survey (i.e., the survey sent out by Flycatcher in which they inform and ask participants for consent) will also have RANDOM_ID 1 in the baseline questionnaire and the daily questionnaires. Someone who has RANDOM_ID 2 in the selection study also has RANDOM_ID 2 in the baseline questionnaire and the daily questionnaires. Vera Heininga can therefore link the answers to the different questionnaires within this study based on the RANDOM_ID. This RANDOM_ID does change between different surveys, so the data cannot be linked to data from previous Flycatcher projects or future projects. Flycatcher will delete all personal data at the latest one year after the end of the project. The Privacy and Security officers of the BSS faculty of the UG and ABJZ of the UG are still working on the data processing agreement between the UG and Flycatcher - this will be signed before the start of the project.

After the data collection, Vera Heininga will download the photos (i.e., Front facing photo, Back facing photo) and minimised metadata (i.e., BeReal username, lateInSeconds, takenAt, creationDate, Caption, number of smileys and number of reactions by BeReal friends) via unishare from the CIT. Vera Heininga will immediately replace the BeReal user name by the Flycatcher RANDOM_ID in all files and folders (to prevent them being potentially identifiable, even though we will instruct participants to choose a non-identifiable name). During all parts of the data life cycle, the photos will be stored in a separate project folder (i.e., separating the photos from the project folder with all other project information such as questionnaire data). All photo materials will be stored for a maximum of two years.

In the data preparation part of the project, photos will be scored using AI software on characteristics (i.e., happy, neutral, sad facial expressions; being in company of friends; being in nature; being active). The AI software will be run locally on the servers of the CIT. Only the characteristics assigned to the photos will be part of the pseudonymized dataset used for analysis. In addition, the age variable will be minimised into categories (e.g., from 16-20; 21-25 etc.).

Furthermore, in data preparation, several steps will be taken:

1. 1. CIT will score the BeReal photos using AI software (Open Computer Vision Library; OpenCV) on characteristics (i.e., happy, neutral, sad facial expressions; being in company of friends; being in nature; being active). OpenCV will be run locally on the servers of the CIT.
   2. Vera Heininga will receive the data from CIT (characteristics deduced from the BeReal photos and minimised metadata). Following the de-identification guide, Vera Heininga will replace the BeReal user name with the Random ID assigned by Flycatcher and minimise the age variable by grouping the variable into categories (e.g., from 16-20; 21-25 etc.).
   3. Next, the data from CIT will be coupled to the data from Flycatcher (Baseline and daily questionnaire) using the Random ID variable.
   4. Finally, Vera Heininga will use R and Rstudio to prepare the data for analysis. Please note that the prepared dataset will contain 1) characteristics deduced from the BeReal photos and its minimised metadata, 2) baseline and diary questionnaires data from Flycatcher + Random_ID variable. The prepared dataset does not contain photo materials.

During data preparation, the data will be stored and backed up on the UG Y drive in a folder dedicated to the research project. Only Vera Heininga, and possibly the postdoc, will have access to the project folder. There will be a separate project folder created for the photo materials to which only Vera Heininga will have access.

4.2 How will ownership of the data and intellectual property rights to the data be managed?

Vera Heininga will be the owner of the data, and will have the rights to control access.

Flycatcher will delete all questionnaire data at the latest one year after the end of the project.

5. How and when will data be shared and preserved for the long term?

5.1 How will data be selected for long-term preservation?

- All data resulting from the project will be preserved for at least 10 years

Data retention periods:

1. 1. raw data 10 years. Except for the photos, all photo materials will be stored for a maximum of two years.
   2. prepared data 10 years (i.e., characteristics deduced from the BeReal photos, its minimised metadata, and baseline and diary questionnaires data from Flycatcher + Random_ID variable. The prepared dataset does not contain photo materials.)
   3. analysed data 10 years (i.e, selection of variables needed to replicate the results submitted for publication)

5.2 Are there any (legal, IP, privacy related, security related) reasons to restrict access to the data once made publicly available, to limit which data will be made publicly available, or to not make part of the data publicly available?

If yes, please explain.

- No

5.3 What data will be made available for re-use?

- Other (please specify)

The prepared dataset will be made ready for re-use, meaning all data without the actual photos because they are directly identifying. The photos will be saved for two years in a folder separate from the rest of the data to which only Vera Heininga will have access.

5.4 When will the data be available for re-use, and for how long will the data be available?

- Data available as soon as article is published
- The analysed dataset will be published via the Open Access journal and/or the Open Science Framework (OSF) together with its R code to facilitate replication of the results reported in the scientific article.
- Vera Heininga is responsible and can be contacted via v.e.heininga@rug.nl
- FAIR principles will be applied to the data.

5.5 In which repository will the data be archived and made available for re-use, and under which license?

The Open Science Framework (OSF) using the repository in Germany under a CC BY-NC license.

5.6 Describe your strategy for publishing the analysis software that will be generated in this project.

The open source tool developed using the "BeFake" python package, is publicly available via Github.

6. Data management costs

6.1 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

I will dedicate my time to data management, as coordinator of the project, together with the student assistant (0.3fte on the budget in my NWO grant application).