



# **STOP THE STEAL: The Metaphorical Construction of Disinformation and Conspiracy Theories in Donald Trump's Tweets before and during the Capitol Riots**



28 June 2024

**Horizon Europe  
Data Management Plan**

*Data Management Plan created in Data Stewardship Wizard «[ds-wizard.org](https://ds-wizard.org)»  
using Common DSW Knowledge Model v2.6.5 (dsw:root:2.6.5).*

HISTORY OF CHANGES		
Version	Publication date	Changes
Version 2	26 Jun 2024	All produced datasets are edited in; VI. Preserving data is completed; Texts are edited for clearer description.
Version 1	26 Jun 2024	Answering as many of the questions as possible

# Contributors

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

- **Lotte Telnekes**

[l.telnekes@student.rug.nl](mailto:l.telnekes@student.rug.nl)

Roles: *Contact Person, Data Collector, Data Curator, Data Manager, Editor, Researcher*

## Projects

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

### **STOP THE STEAL: The Metaphorical Construction of Disinformation and Conspiracy Theories in Donald Trump's Tweets before and during the Capitol Riots**

Acronym:

*Stop the Steal*

Start date:

*2024-04-01*

End date:

*2024-07-01*

Funding:

***Rijksuniversiteit Groningen (The Netherlands)***

*: grant number not yet given*

# 1. Data Summary

## Re-used datasets

We have found the following non-reference datasets that we have considered for re-use:

- **Trump Twitter Archive**

It is available via: <https://www.thetrumparchive.com/>. It is used in the project.

Owner of this dataset: Tweets by the former president are part of the public domain, but this specific dataset is created by the Trump Twitter Archive.

Name: Brendan Contact: [trumptwitterarchive2@gmail.com](mailto:trumptwitterarchive2@gmail.com) More information on: <https://www.thetrumparchive.com/faq> .

The dataset can be used in the provided format without any conversion needed.

We already have a copy of this dataset.

It is a fixed dataset, changes will not influence reproducibility of our results.

We will make sure the selected subset will be available together with our results.

We will use the dataset as follows: this will be one of the datasets which my research will be based upon. I will combine this with another dataset to make sure it includes all of Donald Trump's tweets. These will be later filtered to the right timespan as well as be cleaned for textual analysis, such as keyword searching and textual exploring through Voyant. .

- **Kaggle Datasets** (Kaggle Datasets)

It is available via: <https://www.kaggle.com/datasets/codebreaker619/donald-trump-tweets-dataset>. It is used in the project.

Owner of this dataset: Tweets by the former president are part of the public domain, but this specific dataset is created by the following Kaggle user: User: codebreaker619 Contact information available here: <https://www.kaggle.com/datasets/codebreaker619/donald-trump-tweets-dataset>.

The dataset can be used in the provided format without any conversion needed.

We already have a copy of this dataset.

It is a fixed dataset, changes will not influence reproducibility of our results.

We will make sure the selected subset will be available together with our results.

We will use the dataset as follows: this will be one of the datasets which my research will be based upon. I will combine this with another dataset to make sure it includes all of Donald Trump's tweets. These will be later filtered to

the right timespan as well as be cleaned for textual analysis, such as keyword searching and textual exploring through Voyant. .

We will need to harmonize different sources of existing data.

## 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.

- **Text file**

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

- **1.merged\_trump\_tweets\_clean** (published)

The dataset has the following identifiers:

- URL: [https://github.com/LTelnekes/STS\\_MA\\_Thesis/blob/main/STS\\_Datasets/1.merged\\_trump\\_tweets\\_clean.csv](https://github.com/LTelnekes/STS_MA_Thesis/blob/main/STS_Datasets/1.merged_trump_tweets_clean.csv)

We will distribute the dataset using:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

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.filtered\_trump\_tweets** (published)

The dataset has the following identifiers:

- URL: [https://github.com/LTelnekes/STS\\_MA\\_Thesis/blob/main/STS\\_Datasets/2.filtered\\_trump\\_tweets.csv](https://github.com/LTelnekes/STS_MA_Thesis/blob/main/STS_Datasets/2.filtered_trump_tweets.csv)

We will distribute the dataset using:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

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.

- **3.filtered\_trump\_tweets\_keyword** (published)

The dataset has the following identifiers:

- URL: [https://github.com/LTelnekes/STS\\_MA\\_Thesis/blob/main/STS\\_Datasets/3.filtered\\_trump\\_tweets\\_keywords.csv](https://github.com/LTelnekes/STS_MA_Thesis/blob/main/STS_Datasets/3.filtered_trump_tweets_keywords.csv)

We will distribute the dataset using:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

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.

- **4.all\_candidate\_metaphors\_trump\_tweets** (published)

The dataset has the following identifiers:

- URL: [https://github.com/LTelnekes/STS\\_MA\\_Thesis/blob/main/STS\\_Datasets/4.all\\_candidate\\_metaphors\\_trump\\_tweets.csv](https://github.com/LTelnekes/STS_MA_Thesis/blob/main/STS_Datasets/4.all_candidate_metaphors_trump_tweets.csv)

We will distribute the dataset using:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

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.

- **5.all\_context\_candidate\_metaphors** (published)

The dataset has the following identifiers:

- URL: [https://github.com/LTelnekes/STS\\_MA\\_Thesis/blob/main/STS\\_Datasets/5.all\\_context\\_candidate\\_metaphors.csv](https://github.com/LTelnekes/STS_MA_Thesis/blob/main/STS_Datasets/5.all_context_candidate_metaphors.csv)

We will distribute the dataset using:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

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.

- **6.all\_metaphors\_trump\_tweets** (published)

The dataset has the following identifiers:

- URL: [https://github.com/LTelnekes/STS\\_MA\\_Thesis/blob/main/STS\\_Datasets/6.all\\_metaphors\\_trump\\_tweets.csv](https://github.com/LTelnekes/STS_MA_Thesis/blob/main/STS_Datasets/6.all_metaphors_trump_tweets.csv)

We will distribute the dataset using:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

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.

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

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. Metadata will be available in a form that can be harvested and indexed (managed by the used repository / repositories).

We have made the following arrangements regarding the data ownership: the data in this project consists of tweets by former President Trump. Trump's tweets are considered presidential records and as such are not subject to copyright. .

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **Trump Twitter Archive**

It is freely available with obligation to quote the source (e.g. CC-BY).

- **Kaggle Datasets** (Kaggle Datasets)

It is freely available for any use (public domain or CC0).

For our produced data, conditions are as follows:

- **1.merged\_trump\_tweets\_clean** (published)

The distributions will be accessible through:

- *Domain-specific repository:*

**GitHub** (GitHub)

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



A user of this data can use it without any specific software.  
The dataset will be published after all our processing has finished.

- **2.filtered\_trump\_tweets** (published)

The distributions will be accessible through:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

A user of this data can use it without any specific software.  
The dataset will be published after all our processing has finished.

- **3.filtered\_trump\_tweets\_keyword** (published)

The distributions will be accessible through:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

A user of this data can use it without any specific software.  
The dataset will be published after all our processing has finished.

- **4.all\_candidate\_metaphors\_trump\_tweets** (published)

The distributions will be accessible through:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

A user of this data can use it without any specific software.  
The dataset will be published after all our processing has finished.

- **5.all\_context\_candidate\_metaphors** (published)

The distributions will be accessible through:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

A user of this data can use it without any specific software.  
The dataset will be published after all our processing has finished.

- **6.all\_metaphors\_trump\_tweets** (published)

The distributions will be accessible through:

- *Domain-specific repository:*

**GitHub** (GitHub)

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

A user of this data can use it without any specific software.  
The dataset will be published after all our processing has finished.

## 2.3. Making data interoperable

We will be using the following data formats and types:

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

It is a standardized format.

- **Text file**

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:

- **1.merged\_trump\_tweets\_clean** (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.
- **2.filtered\_trump\_tweets** (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.
- **3.filtered\_trump\_tweets\_keyword** (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.
- **4.all\_candidate\_metaphors\_trump\_tweets** (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.
- **5.all\_context\_candidate\_metaphors** (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.
- **6.all\_metaphors\_trump\_tweets** (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 do not plan to be archiving data (using so-called *cold storage*) for long term preservation already during the project.

## 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.

None of the used repositories charge for their services.

Lotte Telnekes is responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Lotte Telnekes is responsible for finding, gathering, and collecting data.

Lotte Telnekes 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

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:

- **1.merged\_trump\_tweets\_clean**
  - It contains personal data.
  - It does not contain sensitive data.
- **2.filtered\_trump\_tweets**
  - It contains personal data.
  - It does not contain sensitive data.
- **3.filtered\_trump\_tweets\_keyword**
  - It contains personal data.

- It does not contain sensitive data.
- **4.all\_candidate\_metaphors\_trump\_tweets**
  - It contains personal data.
  - It does not contain sensitive data.
- **5.all\_context\_candidate\_metaphors**
  - It contains personal data.
  - It does not contain sensitive data.
- **6.all\_metaphors\_trump\_tweets**
  - It contains personal data.
  - It does not contain sensitive data.

## Data we collect

We will collect data connected to a person, i.e. "personal data". We explored General Data Protection Regulation (GDPR) considerations and relevant materials. We collect personal data for the benefit of society, and this is more important than the privacy of the subjects (i.e. public interest). The purpose of processing the personal data can be described as follows: the content of the tweet is important to interpreting the data for my research. These tweets may include names and information on people. I am also collecting Trump's Twitter account name. .

The data collection is not subject to ethical legislation.

## 7. Other issues

We use the [Data Stewardship Wizard](https://researchers.dsw.elixir-europe.org/wizard) with its *Common DSW Knowledge Model* (ID: dsw:root:2.6.5) knowledge model to make our DMP. More specifically, we use the <https://researchers.dsw.elixir-europe.org/wizard> DSW instance where the project has direct URL: <https://researchers.dsw.elixir-europe.org/wizard/projects/bb1dfc50-d4e3-4397-aadc-d0f9b8864d6e>.

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

- **UG Research Data Policy 2021**  
<https://www.rug.nl/digital-competence-centre/research-data/policies?lang=en>  
 This data is part of my MA Thesis for the Master's in Digital Humanities, which is a program offered by University of Groningen. I will be adhering to the university's research policy.