Collecting Data: Political vocabulary in The Guardian

Horizon Europe
Data Management Plan

13 January 2024

*Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.6.3 (dsw:root:2.6.3).*

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
4.0 (M. Gao)	13-01-2024	Add descriptions in Ethics
3.0 (W.D. Noordink)	13-01-2024	Completely revised final version in Horizon Europe - template
2.0 (M. Gao)	12-01-2024	Add CSV as data types; Add provisions for data sharing; Add UG research Data Policy
1.0 (W.D. Noordink)	05-01-2024	Creating DMP with the basic information of the project
There are no other versions		

Contributors

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

- **Ming Gao S5702402**
  m.gao.6@student.rug.nl
  Roles: *Data Collector*, *Data Manager*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Haozhe Bai S5171326**
  h.bai.3@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Willem Dirk Noordink S2946076**
  w.d.noordink@student.rug.nl
  Roles: *Contact Person*, *Data Collector*, *Data Manager*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Shuangjun Zhang S5332451**
  s.zhang.47@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Zhuoli Lu S4719867**
  z.lu.12@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Aarti Balaji S5555248**
  a.balaji.4@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Jennifer Dijkstra S4117840**
  j.dijkstra.56@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Project Member*
  Affiliation:
  University of Groningen type Education

- **Wengyi Sun S5716683**
  w.sun.10@student.rug.nl
  Roles: *Data Collector*, *Data Curator*, *Project Member*
  Affiliation:
  University of Groninge type Education

Projects

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

Political vocabulary in The Guardian

Acronym:

PVG

Start date:

2023-11-13

End date:

2024-02-01

Funding:

Does not apply

The aim of this project is to study the use of left-wing and right-wing vocabulary in the news language used to cover the World News section of The Guardian. The project will explore the language it employs in reporting world news in light of the ideological inclination claimed by the newspaper.

1. Data Summary

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values** (CSV) type model and format

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  It is a standardized format. This is a suitable format for long-term archiving.

  We will use UTF-8 encoding standard, it is a variable-length character encoding standard.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **The Guardian: articles (dataset)** (not published)

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. Guardian_corpus_text_[name_article]. We will be keeping the relationships between data clear in the file names.

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.

Data will be released only as soon as restrictions are falling away.

Metadata will be openly available including instructions how to get access to the data. Metadata will not be available in a form that can be harvested and indexed.

We have made the following arrangements regarding the data ownership: idem: We use articles from 'The guardian'. These articles are free to open and view. Nevertheless, there is copyright, and there is a limited possibility of making this data available again. An obvious reference to the location of the articles is of course possible.

For our produced data, conditions are as follows:

- **The Guardian: articles (dataset)** (not published)

2.3. Making data interoperable

We will be using the following data formats and types:

- **Comma-separated Values** (CSV) type model and format

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- **Unicode Transformation Formatcv 8-bit** (UTF-8) type model and format

  UTF-8, which stands for Unicode Transformation Format 8-bit, is a character encoding standard that is capable of encoding all possible characters (code points) in Unicode. Unicode is a universal character set that aims to represent every character from every language and script in a standardized way.

  UTF-8 is a variable-width encoding, meaning that it uses different numbers of bytes to represent different characters. It can encode characters using one to four bytes.It is a standardized format.

2.4. Increase data re-use

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

- **The Guardian: articles (dataset)** (not 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 compute infrastructures.
- We will be instrumenting the tools into pipelines and workflows using automated tools.

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 already during the project.

None of the used repositories charge for their services.

Haozhe Bai S5171326, Shuangjun Zhang S5332451, Zhuoli Lu S4719867, Aarti Balaji S5555248, Jennifer Dijkstra S4117840, and Wengyi Sun S5716683 are responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Willem Dirk Noordink S2946076 , Ming Gao S5702402, Haozhe Bai S5171326, Shuangjun Zhang S5332451, Zhuoli Lu S4719867, Aarti Balaji S5555248, Jennifer Dijkstra S4117840, and Wengyi Sun S5716683 are responsible for finding, gathering, and collecting data.

Willem Dirk Noordink S2946076 and Ming Gao S5702402 are responsible for maintaining the finished resource.

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: Jupyter Notebook;

Note: For the course ‘Tools and Methods’ we use also: Voyant; Stylo.

5. Data security

Project members 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.

All personal data will be collected anonymously.

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.
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:

- It does not contain personal data or contain sensitive data.
- ChatGPT will assist in the processing of the data. The processed data will be used for further analysis in an aggregated format only.

Data we collect

We will not collect any data connected to a person, i.e. "personal data". The data collection is not subject to ethical legislation.

The data collected from The Guardian will not be published on GitHub to respect the copyright of the original text; only the code for collection and processing will be made available.

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 direct URL: https://researchers.ds-wizard.org/wizard/projects/e1f603d6-1e84-4db2-a4e3-93004cae47c8.

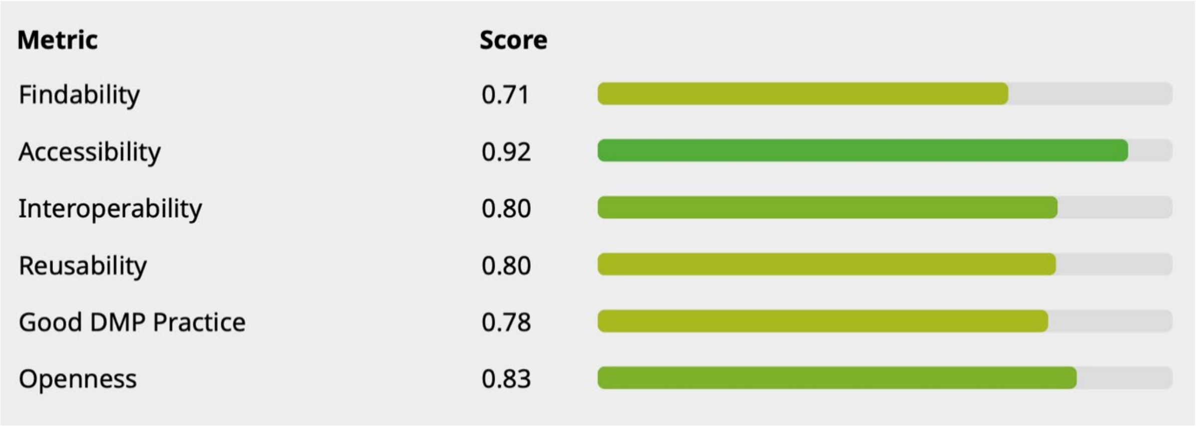
We will be using the following policies and procedures for data management:

- **This project is a (master) Learning Project - University of Groningen, no funders**

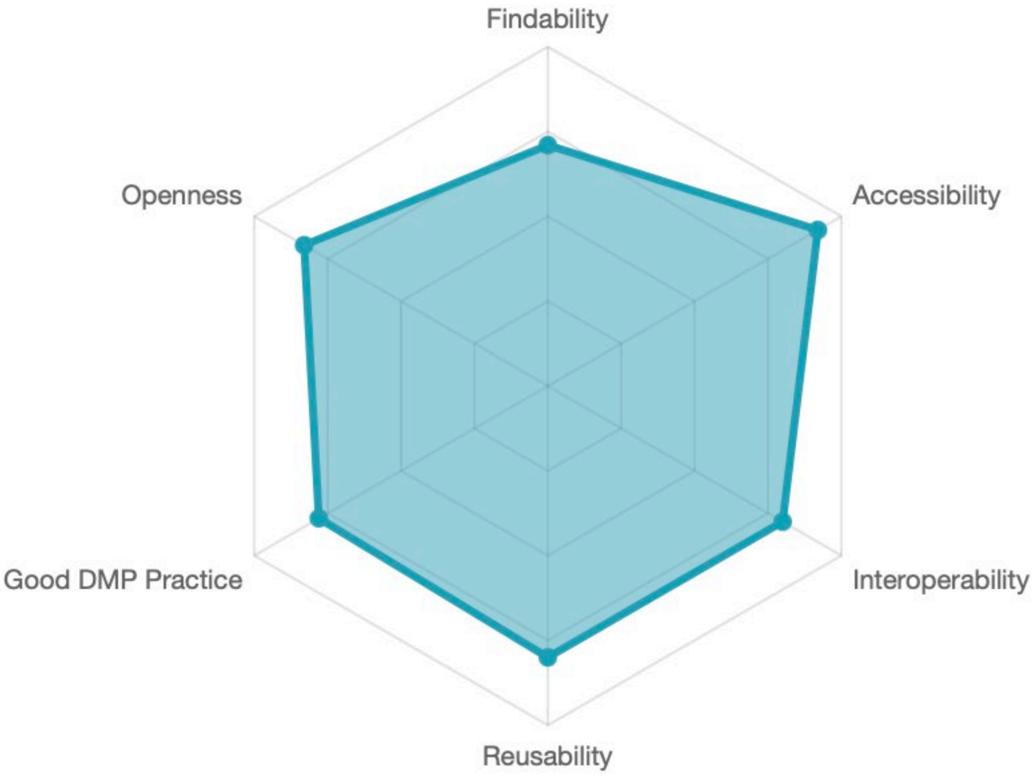
- **UG Research Data Policy**
  https://www.rug.nl/digital-competence-centre/ug-research-data-policy-2021.pdf
  To store my data in university while doing research and after research

- **The Turing Way**
  https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-storage.html?highlight=storage
  To prevent data loss and sort my files orderly

Appendix 1: Short report on findability, reusability and openness (by DS Wizard)



Metric	Score
Findability	0.71
Accessibility	0.92
Interoperability	0.80
Reusability	0.80
Good DMP Practice	0.78
Openness	0.83



Findability
Accessibility
Interoperability
Reusability
Good DMP Practice
Openness