Data Management Plan

Stylometric Analysis on political speeches: where do Churchill's Speeches fall in the political spectrum of the British Parliament between 1939-1940?

Contact person: ***Bakema, J. T., Hoekstra, T., Agapitos, P. & Sygletou. M.*** *(j.bakema.1@student.rug.nl, t.hoekstra.15@student.rug.nl, p.agapitos@student.rug.nl & m.sygletou@student.rug.nl, 0000-0001-9001-7057)*
University of Groningen

Based on: *Common DSW Knowledge Model, 2.3.0 (dsw:root:2.3.0)*

Project phase: *After Finishing the Project*

Created by: ***Maria Sygletou*** *(m.sygletou@student.rug.nl)*

Generated on: *28 Jan 2022*

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»

Projects

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

Stylometric Analysis on political speeches: where do Churchill's Speeches fall in the political spectrum of the British Parliament between 1939-1940?

Start date: *2021-11-28*

End date: *2022-01-28*

Funding: *Not Applicable: Not Applicable*

This project aims to determine to which political party the speeches of Winston Churchill circulated from 1939 to 1940 come closer. Once the data were defined by focusing on two different archives, the British Political Speech, and the International Churchill Society, it used the scraping technique to extract data. Then, the three levels of stylometric analysis were used to provide answers to our research question, determining where exactly in the political spectrum of the British Parliament Churchill's Speeches fell during the relevant time.

Section A: Data Collection

1. What data will you collect or create?

Data formats and types

We will be using the following data formats and types:

- **Textual data and metadata**

  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. How will the data be collected or created?

There will be no instrument dataset in this project.

Storage and file conventions

We will use a filesystem with files and folders with the following folder conventions:

- There will be a **folder for each sample/subject**.

Moreover, we have made appointments about naming the files.

We will not be storing data in an "object store" system.

We will use a relational database system to store project data. Modifications will be made by *Expiring* the existing data and *Adding* updated data.

We will not use a graph database for data in the project.

We will not be storing data in a triple store.

Section B: Documentation and Meta-data

3. What documentation and meta-data will accompany the data?

List of data to be published is given in Section E, Question 9. This also includes information about catalogs where the data can be found. Information about data types used is given in Section A, Question 1.

Section C: Ethics and Legal Compliance

4. How will you manage any ethical issues?

Data we collect

We don't need any consent for collected data because those are not personal.

5. How will you manage copyright and Intellectual Property Rights (IPR) issues?

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open immediately.

Our data is legally not copyrightable, there is no legal owner.

Section D: Storage and Backup

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

Storage needs will be the same during the whole project.

All essential data is also stored elsewhere to prevent a total loss of data. We will make (automated) backups of all data stored outside of the working area.

7. How will you manage access and security?

Project members will not store data or software on computers in the lab or external hard drives connected to those computers.They 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 addressed via secure http (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. 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.

We are not using any personal information.

Only all project members have read/write access to the data.

Section E: Selection and Preservation

8. Which data are of long-term value and should be retained, shared, and/or preserved?

We plan to produce the following datasets:

- **The first dataset contains the so-called 'Leader's Speeches', which refers to speeches of politicians from three different parties, the Conservative, Liberal Democrat and Labour. The second data set consists of speeches of Winston Churchill exclusively.** (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.

9. What is the longterm preservation plan for the dataset?

- **The first dataset contains the so-called 'Leader's Speeches', which refers to speeches of politicians from three different parties, the Conservative, Liberal Democrat and Labour. The second data set consists of speeches of Winston Churchill exclusively.** (published)
  The distributions will be stored in:
  - Domain-specific repository: GitHub. We don't need to contact the repository because it is a routine for us.
  - Domain-specific repository: Open Science Framework. We don't need to contact the repository because it is a routine for us.

  We will be adding a reference to the published data to at least one data catalogue.

None of the used repositories charge for their services.

Section F: Data Sharing

10. How will you share the data?

- **The first dataset contains the so-called 'Leader's Speeches', which refers**

to speeches of politicians from three different parties, the Conservative, Liberal Democrat and Labour. The second data set consists of speeches of Winston Churchill exclusively.
The dataset has the following identifiers:

- DOI: **DOI 10.17605/OSF.IO/3KPB7**

The distributions will be available as follows:

- Open (shared with anyone) using a domain-specific repository: GitHub. The distribution will be available under the following license:
  - Starting The Public Domain, CC0-1.0. 2022-01-08: Freely available for any use (public domain or CC0).
- Open (shared with anyone) using a domain-specific repository: Open Science Framework. The distribution will be available under the following license:
  - Starting The Public Domain, CC0 1.0 Universal. 2022-01-05: Freely available for any use (public domain or CC0).

We will be adding a reference to the published data to at least one data catalogue.

Information about used repositories (i.e. where will potential users find out about the data) is provided in Section E, Question 9.

Embargo on the data is described in Section C, Question 5, and Section F, Question 11.

11. Are any restrictions on data sharing required?

Ethical and legal restrictions are documented under Section C. We have used the Data Stewardship Wizard, which made us aware of options to minimize the restrictions.

No data sharing agreement will be required.

Section G: Responsibilities and Resources

12. Who will be responsible for data management?

Maria Sygletou is responsible for implementing the DMP, and ensuring it is reviewed and revised.

13. What resources will you require to deliver your plan?

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: The Jupyter Notebook was utilized during most of the project's phases such as curation, visualization, analysis. The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.

Charges applied by data repositories (if any) are mentioned already in Section E, Question 9.