Exploring Joe Biden's Presidential Speeches: Uncovering Themes and Trends in Political Discourse

Horizon Europe
Data Management Plan

15 January 2024

Data Management Plan created in Data Stewardship Wizard

History of changes

Version	Publication date	Changes
final_version	14 Jan 2024	This is the final version for the project

Contributors

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

- Arani Aslama
  A.Aslama@student.rug.nl, 0009-0004-2267-8696
  Roles: Contact Person, Data Collector, Project Member
  Affiliation:

  University of Groningen

- Baidan Chen
  B.chen.10@student.rug.nl, 0009-0009-2365-422X
  Roles: Contact Person, Data Collector, Project Member
  Affiliation:

  University of Groningen

- Luotong Cheng
  l.cheng.2@student.rug.nl, 0000-0002-5614-6215
  Roles: Contact Person, Data Collector, Project Member
  Affiliation:

  University of Groningen

- Wenjing Cai
  w.cai.4@student.rug.nl, 0009-0000-5893-6045
  Roles: Contact Person, Data Collector, Project Member
  Affiliation:

  University of Groningen

- Wuhong Xu
  w.a.xu@student.rug.nl, 0009-0000-9312-6398
  Roles: Contact Person, Data Collector, Project Member
  Affiliation:

  University of Groningen

- Mingkai Xu
  m.xu.7@student.rug.nl, 0009-0002-7648-6695
  Roles: Contact Person, Data Collector, Project Member
  Affiliation:

  University of Groningen

- Xiaoyu Zhou
  x.zhou.35@student.rug.nl

Roles: Contact Person, Data Collector, Project Member
Affiliation:

University of Groningen

Projects

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

Exploring Joe Biden's Presidential Speeches: Uncovering Themes and Trends in Political Discourse

Acronym
Presidential Speeches

Start date
2023-12-01

End date
2024-01-15

Funding

- **Rijksuniversiteit Groningen** (The Netherlands)

  : grant number not yet given (planned)

Political speech plays a vital role in comprehending political discourse, as it provides valuable insights into power dynamics and conflicts within a society. The language utilized by politicians, known as political discourse, serves as a potent tool for constructing a favorable and widely accepted self-representation or public perception (Neshkovska, 2019). By examining the main topics in presidential speeches and how presidents adapt their speeches and rhetoric to address these topics, researchers can gain insights into the shifts in public opinion, power dynamics, and conflicts within society.

In this project, we delve into the speeches of Joe Biden, the current President of the United States, to gain a deeper understanding of the prominent topics discussed and their evolution from 2021 to 2023. This allows us to capture the recent developments and changes in political discourse.

Our goal is to collect and analyze 21 speeches delivered by Joe Biden to uncover the underlying themes and trends in his speeches. We will employ computational techniques like the natural language processing and machine learning algorithms. Through topic modeling, we will uncover the latent themes within Biden's speeches, providing a comprehensive overview of the issues he has addressed. By analyzing the parts of speech and named entities, we will gain a deeper understanding of the linguistic patterns and the entities that Biden frequently references. By doing so, we aim to shed light on the key issues that have captured Biden's attention and how they have potentially shifted over time.

1. Data Summary

Re-used datasets

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

- Joe Biden's presidential speeches
  It is available via: https://data.millercenter.org. It is used in the project.

  Owner of this dataset: Miller Center of Public Affairs, University of Virginia. .

  We will first need to convert the format before using it.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: The dataset will be used as a collection of 21 speeches delivered by Joe Biden. This dataset will serve as the input for various computational techniques, such as natural language processing and machine learning algorithms. The goal is to analyze the speeches to uncover themes, trends, linguistic patterns, and entities frequently referenced by Joe Biden. By employing topic modeling, parts of speech analysis, and named entity recognition, we aim to gain insights into the issues that Biden has addressed and how they may have evolved over time.

There is no need to harmonize different sources of existing data in our case.

Data formats and types

We will be using the following data formats and types:

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

  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 have only a small amount of data stored in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

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

We made a SOP (Standard Operating Procedure) for file naming. In order to establish an organized and efficient data management system, we adopted a standardized approach for naming files and folders. when naming folders, we used descriptive terms that accurately represent the content or purpose of the files they contain. for instance, if there is a folder containing annotated files, it can be named "annotated file." similarly, a folder that stores data collected from the miller center can be named "collecting data." by using informative folder names, it becomes easier to locate files based on their intended use or source. in terms of file naming, we include relevant keywords that provide insights into the content of the file. for example, each csv file related to biden's speeches can be named in a way that includes the speaker's name (joe biden) and a brief description of the file's content. for instance, a cleaned version of biden's speech transcripts can be named "joe_biden_speech_clean.csv." . 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 how to get access to the data. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

All data will be owned by the institute.

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

- Joe Biden's presidential speeches
  It is freely available for any use (public domain or CC0).

2.3. Making data interoperable

We will be using the following data formats and types:

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

  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.

2.4. Increase data re-use

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.
- 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 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.

Arani Aslama, Baidan Chen , Luotong Cheng, Wenjing Cai, Wuhong Xu, Mingkai Xu, and Xiaoyu Zhou are responsible for finding, gathering, and collecting data.

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 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 are 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.

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 collect

We will not collect any data connected to a person, i.e. "personal data".

The data collection is subject to ethical legislation. It is covered by ethical review. It involves human subjects.

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/1e19ef21-cbae-4992-9cc6-413a1a06c405.

We will not be using any extra national, funder, sectorial, nor departmental policies or procedures for data management.