Royal Christmas Speeches

Horizon Europe
Data Management Plan

dmp4 - 13/01/2024

13 January 2024

History of changes

Version	Publication date	Changes
dmp4 - 13/01/2024	13 Jan 2024	the fourth version of the data management plan
dmp3 - 08/01/2024	08 Jan 2024	the third version of the data management plan
dmp2 - 03/01/2024	03 Jan 2024	the second version of the data management plan
dmp1 - 30/12/2023	03 Jan 2024	the first version of the data management plan

Contributors

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

- Mengying Xu
  m.xu.8@student.rug.nl
  Roles: Data Curator, Data Manager, Producer, Project Member
  Affiliation:

  University of Groningen

- Miriam Weigand
  m.c.weigand@student.rug.nl
  Roles: Data Collector, Data Curator, Project Member
  Affiliation:

  University of Groningen

- Shiyan Jiang
  s.jiang.13@student.rug.nl
  Roles: Producer, Project Member, Researcher, Quality Assurance
  Affiliation:

  University of Groningen

- Mathilde Contreras Latorre
  m.m.g.contreras.latorre@student.rug.nl
  Roles: Producer, Project Member, Researcher
  Affiliation:

  University of Groningen

- Yunchi Liu
  y.liu.145@student.rug.nl
  Roles: Producer, Project Member
  Affiliation:

  University of Groningen

- Bente van Ingen
  b.n.van.ingen@student.rug.nl
  Roles: Producer, Project Member, Researcher
  Affiliation:

  University of Groningen

- Réka Jurth
  r.k.jurth@student.rug.nl
  Roles: Data Steward, Producer, 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.

Royal Christmas Speeches

Acronym
RCS

Start date
2023-12-01

End date
2024-01-18

Funding

- : grant number not yet given

In this project we are collecting the transcripts of the Christmas speeches given by Queen Elizabeth II between 1952 and 2021 and King Charles III in 2022 and 2023. We are gathering these transcripts through web scraping from the official website of the British Royal Family (https://www.royal.uk/the-christmas-broadcast). We aim to create a collection of these royal Christmas speeches enriched with metadata on the speeches (e.g. information about the type of media that broadcast the speech) and with annotations of the main text section of the speeches (word tokens, sentence tokens, lemmas). Our objective is to create datasets that would enable us to carry out further analyses of the speeches, answer our research questions, provide tutorials and create active learning exercises.

Research questions:

Collecting Data:

1.) Catsiapis (2005) mentions ‘Family’, ’Commonwealth’ and ’Christmas’ as overarching themes in her close reading of the Queen’s Christmas speeches. Can we identify the same themes with computational methods?

2.) When did the term 'empire' change into 'commonwealth'? And what are the things most associated with both?

Tools and Methods:

1.) Does the style of the Queen’s Christmas speeches evolve over time or does it stay the same?

2.) Are King Charles III's speeches stylistically similar or dissimilar to Queen Elizabeth's Christmas speeches?

3. a) What are the most used words by the Queen in 1952, 1953 and by the King in 2022, 2023? (a comparison of the Queen’s and the King’s first two speeches)

b) What are the most used words by the Queen in 2020, 2021 and by the King in 2022, 2023? (a comparison of the last two speeches of the Queen and the first two speeches of the King, from the perspective of continuity)

Reference:

Catsiapis, H. 2005. The Queen's Christmas Messages. In Vernon, P. (Ed.), Seeing Things: literature and the visual : Papers from the Fifth International British Council Symposium, September 2001. Presses universitaires François-Rabelais. doi :10.4000/books.pufr.4223

1. Data Summary

Non-equipment datasets

The non-equipment datasets are:

- **data.csv** – cleaned speeches dataset - the cleaned transcripts of the Christmas speeches given by Queen Elizabeth II between 1952 and 2021 and King Charles III in 2022 and 2023 scraped from the website https://www.royal.uk/the-christmas-broadcast
- **metadata.csv** – metadata on the speeches dataset
- **enriched_corpus.csv** – dataset of the cleaned speeches enriched with metadata on the speeches and with tokenized and lemmatized versions of the main text section of the speeches

Re-used datasets

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

- transcripts of the Christmas Speeches given by Queen Elizabeth II between 1952 and 2021 and King Charles III in 2022 and 2023
  It is available via: https://www.royal.uk/the-christmas-broadcast. It is used in the project.

  Owner of this dataset: The Royal Household © Crown Copyright https://www.royal.uk/contact .

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

  We will download or get a copy.

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

  We will use the complete dataset.

  We will use the dataset as follows: We aim to create a collection of these royal Christmas speeches enriched with metadata on the speeches (e.g. information about the type of media that broadcast the speech) and with annotations of the main text section of the speeches (word tokens, sentence tokens, lemmas). Our objective is to create datasets that would enable us to carry out further analyses of the speeches, answer our research questions, provide tutorials and create active learning exercises.

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

- **data.csv - cleaned speeches dataset** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/kin0330/final-project-H/blob/main/data.csv**

  We will distribute the dataset using:
  - *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide download-only service.
    A persistent identifier will be assigned by the repository. The repository will make sure that the persistent identifier can be resolved to a digital object. The assigned persistent identifier is specified: https://github.com/kin0330/final-project-H/blob/main/data.csv.

  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.
- **metadata.csv - metadata on the speeches dataset** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/kin0330/final-project-H/blob/main/metadata.csv**

  We will distribute the dataset using:
  - *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide download-only service.
    A persistent identifier will be assigned by the repository. The repository will make sure that the persistent identifier can be resolved to a digital object. The assigned persistent identifier is specified: https://github.com/kin0330/final-project-H/blob/main/metadata.csv.

  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.
- **enriched_corpus.csv - dataset of the cleaned speeches enriched with metadata on the speeches and with tokenized and lemmatized version of the main text section of the speeches** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/kin0330/final-project-H/blob/main/enriched_corpus.csv**

We will distribute the dataset using:

- *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide download-only service.
  A persistent identifier will be assigned by the repository. The repository will make sure that the persistent identifier can be resolved to a digital object. The assigned persistent identifier is specified: https://github.com/kin0330/final-project-H/blob/main/enriched_corpus.csv.

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.

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 made a SOP (Standard Operating Procedure) for file naming. The files are going to be named in the following way: 1. queen_crawler.ipynb - first upload for the web crawler 2. 'labeled_text_files' folder - containing all the scraped speeches in a .txt format this folder is required to clean the speeches. the plain text files containing the christmas speeches are going to be named according to the year in which the speeches were given, e.g. 1952.txt. 3. corpus_processing.ipynb - cleaning and pre-processing the corpus 4. data.csv - cleaned speeches dataset 5. metadata.csv - metadata on the speeches dataset 6. enriched_corpus.csv - dataset of the cleaned speeches enriched with metadata on the speeches and with tokenized and lemmatized versions of the main text section of the speeches 7. processing_ale.ipynb - processing, cleaning and active learning activities 8. python_analysis.ipynb - python_analysis uploaded 9. report.pdf - report on the stylometric analysis of the speeches 10. dmp.pdf - data management plan. 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).

We have made the following arrangements regarding the data ownership: The project members will be the rightsholders.

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

- transcripts of the Christmas Speeches given by Queen Elizabeth II between 1952 and 2021 and King Charles III in 2022 and 2023
  It is freely available with obligation to quote the source (e.g. CC-BY).

For our produced data, conditions are as follows:

- **data.csv - cleaned speeches dataset** (published)
  The distributions will be accessible through:
  - *Special-purpose repository for the project*. It will be *Open* (shared with anyone). We will be able to support this repository for a sufficiently long time. The repository will provide download-only service. The distribution will be available under the following license:
    - Starting 2024-01-18: Freely available with obligation to quote the source (e.g. CC-BY).

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.
- **metadata.csv - metadata on the speeches dataset** (published)
  The distributions will be accessible through:
  - *Special-purpose repository for the project*. It will be *Open* (shared with anyone). We will be able to support this repository for a sufficiently long time. The repository will provide download-only service. The distribution will be available under the following license:
    - Starting 2024-01-18: Freely available with obligation to quote the source (e.g. CC-BY).

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.
- **enriched_corpus.csv - dataset of the cleaned speeches enriched with metadata on the speeches and with tokenized and lemmatized version of the main text section of the speeches** (published)
  The distributions will be accessible through:
  - *Special-purpose repository for the project*. It will be *Open* (shared with anyone). We will be able to support this repository for a sufficiently long time. The repository will provide download-only service. The distribution will be available under the following license:
    - Starting 2024-01-18: Freely available with obligation to quote the source (e.g. CC-BY).

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.

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

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

- **data.csv - cleaned speeches dataset** (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.
- **metadata.csv - metadata on the speeches dataset** (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.
- **enriched_corpus.csv - dataset of the cleaned speeches enriched with metadata on the speeches and with tokenized and lemmatized version of the main text section of the speeches** (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.
- 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.

We are going to store our data in the following repository:
https://github.com/kin0330/final-project-H

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 after the project but also already during the project. The minimum lifetime of the archive is specially arranged – We are going to archive the data in the github repository of the project: https://github.com/kin0330/final-project-H. Data formats of data in cold storage will be upgraded if they become obsolete. Archived data will be migrated regularly to more modern storage media (e.g. newer tapes).

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication. For making data or other research outputs FAIR, we budgeted: 1 month.

Mengying Xu and Miriam Weigand 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.

Miriam Weigand is responsible for finding, gathering, and collecting data.

Mengying Xu is responsible for maintaining the finished resource.

Réka Jurth is responsible for the management and proficiency of data including data processing, data policies, data guidelines, and data availability.

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 can carry data with them on encrypted data carriers and password-protected laptops. 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 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:

- **data.csv - cleaned speeches dataset**
  - It contains personal data.
  - It does not contain sensitive data.
- **metadata.csv - metadata on the speeches dataset**
  - It contains personal data.
  - It does not contain sensitive data.
- **enriched_corpus.csv - dataset of the cleaned speeches enriched with metadata on the speeches and with tokenized and lemmatized version of the main text section of the speeches**
  - It contains personal data.
  - It does not contain sensitive data.

All the data that we produce contains personal data of exclusively public figures.

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: We aim to answer our research questions based on the transcripts of the Christmas speeches which contain personal data about Queen Elizabeth II, the British royal family and other public figures.

The data collection is not subject to ethical legislation.

All the data that we collect contains personal data of exclusively public figures.

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/06eea45d-1120-4799-b152-14a824ea5cec.

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
  We are using this policy to ensure an ethical management of research data (data collection, data storage, data curation, and data access).
- **Links Policy of the Royal Household**
  https://www.royal.uk/about-site
  We are using this policy to ensure an ethical management of data collection and data access.