Historical Spanish Newspaper

Horizon Europe
Data Management Plan

13 January 2023

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.4.4 (dsw:root:2.4.4).

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
There are no named versions		

Contributors

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

- **Maria Pilar Uribe**
  m.p.uribe@student.rug.nl, 0000-0002-5614-6215
  Roles: *Contact Person*
  Affiliation: *University of Groningen*

Projects

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

Historical Spanish Newspaper

Acronym: *Spanish Newspaper*

Start date: *2022-12-05*

End date: *2023-01-24*

Funding: *: grant number not yet given (planned)*

Newspapers are the tangible record of the history lived by a society, and their importance lies in the way they observe, describe and record the facts. Our project wants to provide a dataset of historical Spanish newspaper editions. This dataset should be suitable for tackling our underlying question: What influences newspaper style? The project focuses on the years 1890 to 1940. This is a subjective selection, but those years were eventful in Spain: the country moved from a monarchy to a monarchy with a dictatorship to a republic, had a civil war, a revolution, and then became a dictatorship.

1. Data Summary

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values**

  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

- **Spanish newspaper dataset** (published)
  The dataset has the following identifiers:

  We will distribute the dataset using:
  - *Domain-specific repository*: 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.

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. We will use the following format for: dates: yyyyy-mm-dd date and file name: yyyyy-mm-dd_namefile.format folders will be sorted: original texts, clean texts and files by type of program to be used for further analysis. as: gephi, stylo, flourish, etc. 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 immediately.

Limited embargo will not be used as all data will be opened immediately.

Metadata will be openly available without 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 Principal Investigator.

For our produced data, conditions are as follows:

- **Spanish newspaper dataset** (published)
  The distributions will be accessible through:
  - *Domain-specific repository*: 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 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**

  It is a standardized format.

2.4. Increase data re-use

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

- **Spanish newspaper dataset** (published) – This data set will be kept available as long as technically possible.

As stated already in Section 2.2, all of our data can become completely open immediately.

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.

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

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.

6. Ethics

For the data we produce, the ethical aspects are as follows:

- **Spanish newspaper dataset**
  - It does not contain personal data.
  - It does not contain sensitive data.

Data we collect

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

7. Other issues

We use the Data Stewardship Wizard with its *Common DSW Knowledge Model* (ID: dsw:root:2.4.4) knowledge model to make our DMP. More specifically, we use the https://researchers.ds-wizard.org DSW instance where the project has direct URL: https://researchers.ds-wizard.org/projects/a0f2af08-20b7-470b-b64e-ff05f73ce35d.

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