Lyrics Scraping and Sentiment Analysis

# **Horizon Europe** **Data Management Plan** 19 January 2023

## History of changes

*There are no named versions.*

## Contributors

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

* Zhao Yuanjie  
  [y.zhao.26@student.rug.nl](mailto:y.zhao.26@student.rug.nl)  
  Roles: Contact Person, Data Collector, Data Curator, Data Manager

## Projects

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

### Lyrics Scraping and Sentiment Analysis

Acronym

LySSA

Start date

2023-01-01

End date

2023-01-20

Funding

* : grant number not yet given

This project contains two parts:

a. scraping lyrics data;

b. lyrics sentiment analysis of two chosen albums from two different singers.

## 1. Data Summary

#### Re-used datasets

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

* **taylor\_swift\_lyrics** (<https://data.world/promptcloud/taylor-swift-song-data-from-all-the-albums/workspace/file?filename=taylor_swift_lyrics.csv>) ✔️
* Owner of this dataset: https://www.promptcloud.com.
* The dataset can be used in the provided format without any conversion needed.
* We already have a copy of this dataset.
* The dataset may get updated in the future; therefore, we need to make a snapshot.
* We will make sure the selected subset will be available together with our results.
* We will use the dataset as follows: I will use this dataset as the substitution data source of my project.

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**](https://fairsharing.org/10.25504/FAIRsharing.1943d4)
* 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

* **A data set contains all the lyrics collected in this project** (not published)

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

The provenance will be captured using W3C PROV.

We made a SOP (Standard Operating Procedure) for file naming. The file will be named after the following format: lyrics - album\_name.csv. since the project is simple and the data volum would be fairly low, so there will be no folder structure needed. 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 not be available in a form that can be harvested and indexed.

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

* **taylor\_swift\_lyrics** – freely available with obligation to quote the source (e.g. CC-BY).

For our produced data, conditions are as follows:

* **A data set contains all the lyrics collected in this project** (not published)

### 2.3. Making data interoperable

We will be using the following data formats and types:

* [**Comma-separated Values**](https://fairsharing.org/10.25504/FAIRsharing.1943d4)
* It is a standardized format.

### 2.4. Increase data re-use

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

* **A data set contains all the lyrics collected in this project** (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 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.

Yuanjie Zhao is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Zhao Yuanjie is responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Zhao Yuanjie is responsible for finding, gathering, and collecting data.

Zhao Yuanjie is responsible for maintaining the finished resource.

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:

* **A data set contains all the lyrics collected in this project**
  + 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".

For reused non-reference datasets, the consent for privacy sensitive data will be solved as follows:

* **taylor\_swift\_lyrics**
* None of the data from this dataset is personal data.

## 7. Other issues

We use the [Data Stewardship Wizard](https://ds-wizard.org) 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/bd1d13c8-a628-469f-8da8-284a00795358>.

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

* **RUG Policy**  
  <https://www.rug.nl/research/research-data-management/policy/ug-rdm/>  
  Its purpose is to ensure innovative research and research integrity.