Collecting Data Group project

Horizon Europe
Data Management Plan

15 January 2023

History of changes

There are no named versions.

Contributors

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

- Jiang Jane
  jiangyuejane@outlook.com
  Roles: Data Collector, Data Curator, Data Manager, Project Member, Researcher
- Lei Pan
  panleipanleipanlei@gmail.com
  Roles: Contact Person, Data Collector, Data Curator, Data Manager, Project leader, Project Manager ,Researcher
- Li Yang
  y.li.171@student.rug.nl
  Roles: Data Collector, Data Curator, Data Manager, Project Member, Researcher Researcher
- Zhou Marina
  fance1235252@gmail.com
  Roles: Data Collector, Data Curator, Data Manager, Project Member, Researcher

Projects

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

Analysis of Lyrics during the Covid-19 Pandemic

Acronym
ALCP

Start date
2022-12-15

End date
2023-01-24

Funding
Did not apply for any funding yet.

We decided to work on the lyrics of some popular songs to see whether there are any differences between the songs released between 2018 and 2022.Does COVID-19 have any effect on lyrics? If it does, it is good or bad?

1. Data Summary

Re-used datasets

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

- **The Billboard Hot 100** (https://www.billboard.com/charts/year-end/hot-100-songs/) ✓

  Owner of this dataset: Name: Billboard Media, LLC. Contact details: PMC@wrightsmedia.com, Research@billboard.com.

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

  The original dataset will be available both from the provider and from us together with our results for the reproducibility.

  We will use the dataset as follows: We use this dataset to get the titles and artists of the top 50 songs in the US for each year from 2018 to 2022.

- **The Official Charts** (https://www.officialcharts.com/charts/) ✓

  Owner of this dataset: Name: Rob Poole Contact details: robpoole@officialcharts.com / commercial@officialcharts.com.

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

  The original dataset will be available both from the provider and from us together with our results for the reproducibility.

  We will use the dataset as follows: We use this dataset to get the titles and artists of the top 50 songs in the UK for each year from 2018 to 2022.

- **Top Mandarin Yearly Singles Chart** (https://kma.kkbox.com/charts/yearly/newrelease?lang=en&terr=my) ✓

  Owner of this dataset: Name: KKBOX Contact details: legal@kkbox.com .

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

  The original dataset will be available both from the provider and from us together with our results for the reproducibility.

  We will use the dataset as follows: We use this dataset to get the titles and artists of the top 50 songs in the Taiwan for each year from 2018 to 2022.

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

- **genius** (https://genius.com) ✓

Owner of this dataset: Name: ML Genius Holdings, LLC Contact details: terms@genius.com. We will need to request access to the dataset from its owners.

The dataset can be used in the provided format without any conversion needed.

We will download or get a copy.

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

We will make sure the selected subset will be available together with our results.

We will use the dataset as follows: We use this dataset to get the lyrics of the top 50 songs in the UK and the US for each year from 2018 to 2022.

- **KKBOX** (https://www.kkbox.com/tw/tc/) ✓

  Owner of this dataset: Name: KKBOX Contact details: bd@kkbox.com. We will need to request access to the dataset from its owners.

  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 make sure the selected subset will be available together with our results.

  We will use the dataset as follows: We use this dataset to get the lyrics of the top 50 songs in Taiwan for each year from 2018 to 2022.

We will need to harmonize different sources of existing data.

Data formats and types

We will be using the following data formats and types:

- **CSV**

  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

- **Analysis of Lyrics during the Covid-19 Pandemic** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/jiangyuejane1/Collecting-Data-Group-Project.git**

  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: b68e899.

  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 have 1 notebook file named 'Collecting_data_group_project'. In addition to this, we will create 3 folders named 'csv', 'Lyrics_By_Region' and 'Lyrics_By_Year' to include all the data scraped and processed for our research.

'csv' folder contains 3 subfolders named by music charts, the naming format is Top50_Charts_csv, and each subfolder contains 5 csv files with metadata named by Top50_Charts_Year.

'Lyrics_By_Region' folder contains 3 subfolders also named by music charts, the naming format is Top50_Charts_Lyrics. Each music chart subfolder has 5 lyrics folders named by year, naming format is Top50_Charts_Lyrics_Year, which includes txt files of lyrics for each year respectively.

'Lyrics_By_Year' folder contains 5 txt files named by Top50_Lyrics_Year. Each txt files include all lyrics text from 3 music charts each year.

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.

Data that is not legally restrained will be released after a fixed time period (According to China's copyright law, 50 years after the death of the copyright author of the lyrics of the song, the copyright has expired and entered the public domain. Then the lyrics are free to use thereafter.), unconditionally.

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).

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

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

- **The Billboard Hot 100** – freely available with obligation to quote the source (e.g. CC-BY).
- **The Official Charts** – freely available with obligation to quote the source (e.g. CC-BY).
- **Top Mandarin Yearly Singles Chart** – freely available with obligation to quote the source (e.g. CC-BY).
- **genius** – freely available with obligation to quote the source (e.g. CC-BY).
- **KKBOX** – freely available with obligation to quote the source (e.g. CC-BY).

For our produced data, conditions are as follows:

- **Analysis of Lyrics during the Covid-19 Pandemic** (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.

  A user of this data need specific software to be able to use it:
  - Python (https://www.python.org/downloads/)

  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:

- **CSV**

  It is a standardized format.

2.4. Increase data re-use

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

- **Analysis of Lyrics during the Covid-19 Pandemic** (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 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.

Jiang Jane is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Jiang Jane, Lei Pan, Zhou Marina and Li Yang 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.

Jiang Jane, Lei Pan, Li Yang, and Zhou Marina are responsible for finding, gathering, and collecting data.

Jiang Jane, Lei Pan, Zhou Marina and Li Yang are responsible for maintaining the finished resource.

To execute the DMP, additional specialist expertise is required and we have such trained support staff available.

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 risk of information loss in the project or organization is acceptably low. 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.

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:

- **Analysis of Lyrics during the Covid-19 Pandemic**
  - It contains personal data.
  - It does not contain sensitive data.

Data we collect

We will collect data connected to a person, i.e. "personal data". We have a legitimate interest: data subjects all expect us to do this data processing because of who we are.

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

- **genius**

  The existing consent already covers our reuse.

- **KKBOX**

  We will use another legal base for the processing.

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/11c5ad0f-48fe-48b0-8165-aaf41488c66b.

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