Music Popularity Across Platforms: United States' Spotify Charts vs. Billboard Hot 100

Horizon Europe
Data Management Plan

17 January 2024

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.6.3 (dsw:root:2.6.3).

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
2	16 Jan 2024	Artist names are considered as personal data; A dataset that was considered but not re-used is edited in; Texts are edited for clearer description.
1	14 Jan 2024	Initial Version

Contributors

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

- **Shuxiang Du**
  s.du.6@student.rug.nl
  Roles: *Contact Person*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Linchun Hou**
  l.hou@student.rug.nl
  Roles: *Data Collector*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Yingyue Jiang**
  y.jiang.34@student.rug.nl
  Roles: *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Spring Wan**
  s.shuo.wan@student.rug.nl
  Roles: *Data Collector*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Josie Wisowaty**
  s.j.wisowaty@student.rug.nl
  Roles: *Editor*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Lotte Telnekes**
  l.telnekes@student.rug.nl
  Roles: *Data Collector*, *Editor*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

- **Maria Kutepova**
  m.kutepova@student.rug.nl

Roles: *Editor*, *Project Member*, *Researcher*
Affiliation:
University of Groningen type Education

- **Tessa Eisen**
  t.i.eisen@student.rug.nl
  Roles: *Editor*, *Project Member*, *Researcher*
  Affiliation:
  University of Groningen type Education

Projects

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

Music Popularity Across Platforms: United States' Spotify Charts vs. Billboard Hot 100

Acronym:

MPAP

Start date:

2023-12-19

End date:

2024-01-18

Funding:

Rijksuniversiteit Groningen *(The Netherlands)*

: grant number not yet given (planned)

Our project aims to determine the differences between popular music on streaming platforms and mainstream charts and to answer ‘Does music popularity differ across platforms, and what factors can influence this phenomenon?’. Exploring the differences between artists famous on traditional music charts (like Billboard) and the streaming behavior of the public on Spotify can shed light on public listener behavior. This information is essential for musical artists, labels, and managers to understand audience preferences and platform engagement. To answer our research question, we need to collect data on the top songs shown on Spotify and the Billboard Top 100, including information surrounding the artist's name, song title, place in the top 100, and date in the top 100. While we have found datasets related to Billboard's top charts, we would need to scrape the Spotify website (or other platforms if we wish to expand the scope of our project) for that data.

1. Data Summary

Re-used datasets

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

- **Kaggle Datasets** (Kaggle Datasets) type repository

  Kaggle Datasets is a public open data platform which combines data, users, discussions, and software. Datasets can be searched, bookmarked and explored via Kernels and discussions. All types of data are allowed. Individual users and organizations can register. Datasets can be published either privately or shared publicly.

  It is available via: https://www.kaggle.com/datasets/dhruvildave/billboard-the-hot-100-songs/data. It is used in the project.

  Owner of this dataset: Dhruil Dave. Contact information available on https://www.kaggle.com/dhruvildave.

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

  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 will process this dataset for the Billboard Hot 100 chart and compare it with the Spotify dataset we produced.

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

- Spotify Charts
  It is available via: https://charts.spotify.com/charts/overview/global. It is used in the project.

  Owner of this dataset: Spotify AB Regeringsgatan 19 SE-111 53 Stockholm Sweden Reg no: 556703-7485 office@spotify.com.

  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 will record and process the relevant data to compare with the Billboard Hot 100 dataset.

- **Kaggle Datasets** (Kaggle Datasets) type repository

  Kaggle Datasets is a public open data platform which combines data, users, discussions, and software. Datasets can be searched, bookmarked and

explored via Kernels and discussions. All types of data are allowed. Individual users and organizations can register. Datasets can be published either privately or shared publicly.

It is available via: https://www.kaggle.com/datasets/dhruvildave/spotify-charts. It was considered but will not be used in the project.

We decided not use this non-reference dataset because of: this dataset included a lot of information, including the "Top 200" and "Viral 50" charts published globally by Spotify. Spotify publishes a new chart every 2-3 days, and this dataset includes its entire collection since January 1, 2017 until December 31, 2021. As we want to compare Spotify's weekly charts in 2021 to those of the Billboard Hot 100 in 2021, we decided to directly scrape this data from the official Spotify website instead, as this dataset only included these daily charts. .

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) type model and format

  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

- **Billboard Top 100 Weekly Charts in 2021** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/LTelnekes/Collecting_Data_GroupD**

  We will distribute the dataset using:
  - *Domain-specific repository*:
    GitHub (GitHub) type repository

    GitHub Inc. is a Git-based software repository that provides version control. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own

features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for each project. It provides a number of different access levels, from free through to paid services.

. 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.

- **Spotify Top 100 Weekly Charts in 2021** (published)
  The dataset has the following identifiers:
  - URL: **https://github.com/LTelnekes/Collecting_Data_GroupD**

  We will distribute the dataset using:
  - *Domain-specific repository*:
    GitHub (GitHub) type repository

    GitHub Inc. is a Git-based software repository that provides version control. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for each project. It provides a number of different access levels, from free through to paid services.

    . 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.

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 name the files according to the time created, the content within the relevant document, and the version number. . 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. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

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

- **Kaggle Datasets** (Kaggle Datasets) type repository

  Kaggle Datasets is a public open data platform which combines data, users, discussions, and software. Datasets can be searched, bookmarked and explored via Kernels and discussions. All types of data are allowed. Individual users and organizations can register. Datasets can be published either privately or shared publicly.

  It is freely available with obligation to quote the source (e.g. CC-BY).
- Spotify Charts
  It is available under specific restrictions, which we will follow in our project: The following is not permitted for any reason whatsoever in relation to the Services and the material or content made available through the Services, or any part thereof: 1. reverse-engineering, decompiling, disassembling, modifying, or creating derivative works, except where such restriction is expressly prohibited by applicable law. If applicable law allows you to decompile any part of the Services or Content where required in order to obtain the information necessary to create an independent program that can be operated with the Services or with another program, the information you obtain from such activities (a) may only be used for the foregoing objective, (b) may not be disclosed or communicated without Spotify's prior written consent to any third party to whom it is not necessary to disclose or communicate in order to achieve that objective, and (c) may not be used to create any software or service that is substantially similar in its expression to any part of the Services or the Content; 2. copying, reproducing, redistributing, "ripping," recording, transferring, performing, framing, linking to or displaying to the public, broadcasting, or making available to the public, or any other use which is not expressly permitted under the Agreements or applicable law, or which otherwise infringes intellectual property rights; 3. importing or copying any local files that you do not have the legal right to import or copy in this way; 4. transferring copies of cached Content from an authorized Device to any other Device via any means; 5. "crawling" or "scraping", whether manually or by automated means, or otherwise using any automated means (including bots, scrapers, and spiders), to view, access or collect information, or using any part of the Services or Content to train a machine learning or AI model or otherwise ingesting Spotify Content into a machine learning or AI model;.

For our produced data, conditions are as follows:

- **Billboard Top 100 Weekly Charts in 2021** (published)
  The distributions will be accessible through:
  - *Domain-specific repository*:
    GitHub (GitHub) type repository

    GitHub Inc. is a Git-based software repository that provides version control. It offers all of the distributed version control and source code

management (SCM) functionality of Git as well as adding its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for each project. It provides a number of different access levels, from free through to paid services.

. 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.

- **Spotify Top 100 Weekly Charts in 2021** (published)
  The distributions will be accessible through:
  - *Domain-specific repository*:
    GitHub (GitHub) type repository

    GitHub Inc. is a Git-based software repository that provides version control. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for each project. It provides a number of different access levels, from free through to paid services.

    . 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** (CSV) type model and format

  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:

- **Billboard Top 100 Weekly Charts in 2021** (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.

- **Spotify Top 100 Weekly Charts in 2021** (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.

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.

Linchun Hou, Spring Wan, and Lotte Telnekes 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

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 produce

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

- **Billboard Top 100 Weekly Charts in 2021**
  - It contains personal data.
  - It does not contain sensitive data.
- **Spotify Top 100 Weekly Charts in 2021**
  - 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 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 are collecting the artist names from the charts to study their music popularity.

The data collection is not subject to ethical legislation.

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/31b69e22-291a-484d-b0b9-46aabd5cf110.

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/
  This is a group project for Digital Humanities: Tools and Methods (LHU011M05) which is a course offered by University of Groningen. We will be adhering to the university's research policy.