MUHSIC: An Open Dataset with Temporal Musical Success Information

Gabriel P. Oliveira, Gabriel R. G. Barbosa, Bruna C. Melo, Mariana O. Silva, Danilo B. Seufitelli, Mirella M. Moro

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

gabrielpoliveira@dcc.ufmg.br, grgb@ufmg.br,
{brunacamposmelo,mariana.santos,daniloboechat,mirella}@dcc.ufmg.br

Abstract. *Music is an alive industry with an increasing volume of complex data that creates new challenges and opportunities for extracting knowledge, benefiting not only the different music segments but also the Music Information Retrieval (MIR) community. In this paper, we present MUHSIC, a novel dataset with enhanced information on musical success. We focus on artists and genres by combining chart-related data with acoustic metadata to describe the temporal evolution of musical careers. The enriched and curated data allow building success-based time series to investigate high-impact periods (hot streaks) in such careers, transforming complex data into knowledge. Overall, MUHSIC is a relevant tool in music-related tasks due to its easy use and replicability.*

1. Introduction

Music and Computer Science have a long, fruitful relationship, as computing techniques have helped the music industry over a myriad of different problems. Especially on Music Information Retrieval (MIR), much has already been done since the seminal position paper by Byrd and Crawford [2002], from automatically classifying music genres to tackling user gender bias in recommendation algorithms [Melchiorre et al. 2021]. Specialized computing approaches are still necessary due to inherent music challenges, including dynamic evolution, volatile information, and heterogeneous data sources.

With more information available over the Web, solving such challenges inevitably faces huge data volumes, which brings additional intrinsic processing issues. Nonetheless, MIR tasks have much to benefit from online data sources such as Billboard and Spotify. Such platforms provide information related to the music ecosystem, including chart-based features and acoustic fingerprints. This meaningful set of characteristics is paramount to Hit Song Science (HSS), an emerging field within MIR that aims to predict a song's popularity from its features [Çimen and Kayis 2021, Pachet 2011].

Besides predicting song success, music-related data are also relevant to AI-based talent identification, mainly through popularity peaks within artist careers. Such peaks are commonly grouped into hot streak periods, defined as the continuous periods of success or productivity above normal [Liu et al. 2018, Garimella and West 2019, Janosov et al. 2020]. Indeed, identifying hot streaks is very useful for the music industry, as it is one way of investing in the right artist at the most relevant moment. Although the existing music-related datasets provide partial information to perform such a task [Cosimato et al. 2019, Silva et al. 2019], to the best of our knowledge, none of them contains temporal success information already processed for analysis of this type.

Table 1. Comparison of datasets with popularity data.

Year	Dataset	Reference	Size	Songs	Charts	Artists	Genres	Time Series
2011	MSD	[Bertin-Mahieux et al. 2011]	1,000,000	✓	×	✓	×	×
2016	TPD	[Karydis et al. 2016]	23,385	✓	✓	✓	×	×
2019	HSPD	[Zangerle et al. 2019]	1,000,000	✓	✓	×	✓	×
2019	SPD	[Cosimato et al. 2019]	101,939	✓	×	✓	✓	×
2019	MusicOSet	[Silva et al. 2019]	20,405	✓	✓	✓	✓	×
2020	MGD	[Oliveira et al. 2020]	13,880	✓	✓	✓	✓	×
2021	**MUHSIC**	**This work**	**22,635**	✓	✓	✓	✓	✓

In this paper, we introduce MUHSIC (***Mu**sic-oriented **H**ot **S**treak **I**nformation **C**ollection*), an open dataset with temporal information on musical success, focusing on artist and genre careers. Specifically, we provide chart-based success time series from 1958 to 2020, as well as the hot streak periods detected in each time series. Besides, the dataset also contains metadata about the most relevant music elements, i.e., songs, artists, and genres. This novel set of features and the easy use and replicability makes MUHSIC a valuable resource for different MIR applications, such as HSS and Music Genre Classification. Moreover, the success time series may be used for analyzing the temporal evolution of musical careers and identifying success trends in the music industry.

2. Existing Music Datasets

There are numerous datasets that include different information about music, including metadata, acoustic features, lyrics, and popularity data. The subset of information varies according to the work purpose. Nevertheless, the more data available, the more mining tasks may be performed over such data. Table 1 shows the top datasets that include popularity data, an information that is crucial to evaluate success.

The Million Song Dataset (MSD) [Bertin-Mahieux et al. 2011] is the only one in this list that does not have information on popularity. However, it is also one of the most used datasets in MIR, as it contains audio features and metadata for one million music tracks, with over 280 GB of data. It also gets a lot of critics due to lack of details on exactly how its data was extracted and integrated.

Then, the Track Popularity Dataset (TPD) [Karydis et al. 2016] is probably the first to provide data on musical track popularity by considering different sources from 2004 to 2014. It contains features tailored for music information retrieval (e.g., identification spaces and contextual similarity) and considers both popular and non-popular audio tracks. The Hit Song Prediction Dataset (HSPD) [Zangerle et al. 2019] was built upon the MSD by including data about its tracks that were in the Billboard Hot 100 charts.

In a different perspective, MusicOSet [Silva et al. 2019] and SpotGenTrack Popularity Dataset (SPD) [Cosimato et al. 2019] focus on quality and provide metadata, lyrics, acoustic features and song popularity. MusicOSet content enables large-scale evaluations of song and music collaboration-based recommendations, whereas SPD is tailored for other MIR tasks, such as genre classification and auto-tagging. Then, the Music Genre Database (MGD) [Oliveira et al. 2020] expands on such possibilities by providing information on music genres: genre collaboration networks and genre mapping.

Now, MUHSIC naturally shares content with its predecessors; nonetheless, it also

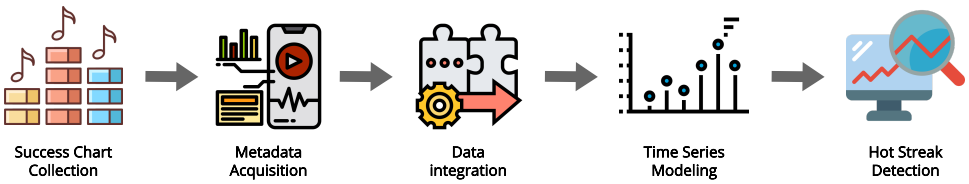

Figure 1. MUHSIC building methodology.

tackles temporal success through *time series modeling* and *hot streaks detection*, which are its most complex and important novelties. Moreover, aiming to embrace several music data mining tasks, MUHSIC provides comprehensive information on the temporal success of both artists and their songs aggregated in genres.

3. MUHSIC Dataset

In this section, we present MUHSIC itself, an open dataset focused on temporal success information. We first describe the building methodology, from data collection to the processing phase (Section 3.1). Then, we overview the main data within the dataset (Section 3.2) and perform a short exploratory analysis over it (Section 3.3). Finally, we go over its format and usage instructions (Section 3.4).

3.1. Building Methodology

MUHSIC building process has five steps: (*i*) collecting music charts to represent temporal success data; (*ii*) collecting song and artist metadata from Spotify; (*iii*) integrating all such data; (*iv*) modeling musical careers in success time series; and (*v*) generating hot streak information from such careers. Figure 1 shows these steps, which are detailed next.

Success Chart Collection. Billboard is an American-based music magazine (with operations in Canada, Brazil, Greece, Japan, South Korea, and Russia) widely known for its exclusive charts on trends across all musical genres. It publishes the *Hot 100* Chart,[1] which is also the main all-genre song ranking in the country and possibly the world, as according to IFPI,[2] the United States is the biggest music market worldwide. Billboard Hot 100 has been weekly published since 1958, and is currently built by summing up songs' sales, airplay and streaming consumption. To model artist success over time, we collect all Hot 100 charts from August 11, 1958 to August 22, 2020 (data collection time) by using the Python package *billboard.py*.[3] Each chart is composed of its 100 entries, ranked from the most popular song to the least popular on that week.

Metadata Acquisition. Billboard chart entries are composed only of the song name and its artists, which is usually not enough to answer complex research questions. Therefore, we improve the initial dataset from the previous step by collecting data from Spotify, the world's most popular audio streaming service with more than 365 million users in 178 markets (as of August 2021). Its API[4] enables to get extra information on artists and songs, including artist genres and debut date; and acoustic features for each song, such as key, mode, and energy.

[1]Billboard Hot 100 Chart: `https://www.billboard.com/charts/hot-100`
[2]IFPI Global Music Report: `https://gmr.ifpi.org/`
[3]billboard.py: `https://github.com/guoguo12/billboard-charts`
[4]Spotify Developer API: `https://developer.spotify.com/`

Data Integration. Building a dataset from multiple sources requires a data integration step to obtain a unique data structure with significant and valuable information. Therefore, as MUHSIC is built using data from Billboard and Spotify, it is necessary to link all songs from the Hot 100 entries to their correspondent Spotify song records with enhanced artist and acoustic features. We do so by using both probabilistic and fuzzy matching approaches for record linkage. Specifically, we use the *SequenceMatcher* class from the Python *difflib*[5] package, the *Jaro-Winkler* algorithm from the *python-string-similarity*[6] package and four similarity functions (i.e., **ratio, partial_ratio, token_sort_ratio, WRatio**) from the FuzzyWuzzy[7] package. We consider a match when the similarity between Billboard and Spotify records is at least 0.9. Consequently, there are some Hot 100 entries for which we could not find their match on Spotify. This may be due to divergences in the song/artist name spelling or the unavailability of songs in Spotify. Overall, we were able to map approximately 85% of all collected songs.

Time Series Modeling. As in most industries, music success evolves by following the audience tastes, worldwide trends, and other factors such as media platforms dynamics, new music styles, and new song releases. Here, we model success over time based on the Hot 100 charts and Spotify data for both artist and genres, by including aggregated acoustic features of the songs that appear in a given week (e.g., the number of explicit songs and the median acousticness) and success information as follows.

For each **artist**, we build a time series from the debut date (i.e., date of the first release obtained from Spotify) to the last chart collected. Thus, each point in the time series represents the success of such an artist in a given week, according to the Hot 100 chart. The success of an artist is given by the *rank scores* for all of their songs that appear on the week chart. The *rank_score* of a song i is $rank\_score(i) = max\_rank - rank(i) + 1$, where *max_rank* is the lowest possible rank (i.e., 100), and *rank(i)* is the song position on the chart. Then, the ranks scores of an artist need to be aggregated somehow. The easiest idea is to sum up all rank scores of the songs. However, that is not enough because an artist with the #1 song in one week is more successful than an artist with two songs in the middle of the chart (e.g., on positions #49 and #50). Hence, rank score aggregation uses the Discounted Cumulative Gain (DCG) [Aggarwal 2016], which emphasizes the most relevant records (i.e., the highest ranked songs on the chart) and penalizes songs that appear in lower positions by a logarithmic factor, as defined by Equation 1.

$$DCG = \sum_{i=1}^{n} \frac{rank\_score(i)}{log_2(i+1)} \tag{1}$$

We also build time series for analyzing the evolution of **genre** success over time. First, we assign artists' genres to their songs, as the songs themselves do not have such information. Then, for each week, we use DCG to aggregate all songs from artists belonging to a given genre that appear on that week chart.

Hot Streak Detection. Professional careers tend to have phases of high productivity, reaching the career peak. Hot streaks (HS) is the term commonly used for continuous

[5] *difflib*: `http://docs.python.org/3.6/library/difflib.html`

[6] *python-string-similarity*: `http://github.com/luozhouyang/python-string-similarity`

[7] *fuzzywuzzy*: `http://github.com/seatgeek/fuzzywuzzy`

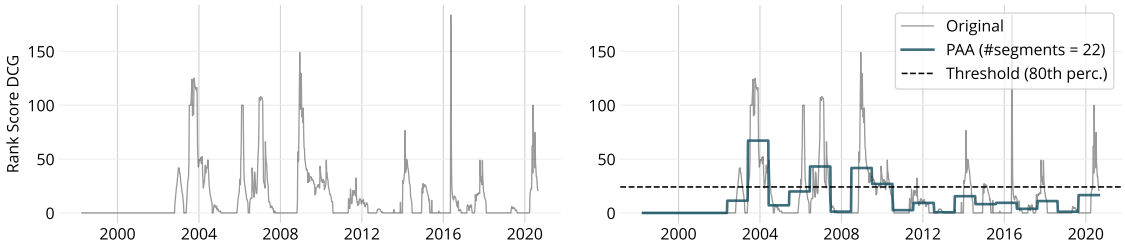

Figure 2. Beyoncé success time series (1998–2020). MUHSIC contains both the raw success time series (left) and the PAA fit with hot streaks (right).

periods of success above normal. Previous work reveals that hot streaks can arise at any time in a professional career [Garimella and West 2019, Liu et al. 2018]. In music, when an artist is at a hot streak, such an artist is also at the most profitable moment of a career. Therefore, we identify hot streak periods in both artist and genre time series (i.e., their careers) to allow further success analyses.

From the raw time series, we first apply Piecewise Aggregate Approximation (PAA) to reduce their dimensionality [Keogh and Pazzani 2000]. In short, PAA reduces a given time series (with n points) into a new series with N segments, $1 \leq N \leq n$. Their values are calculated by the average of the points within such frames. Therefore, the approximation of each point on the original time series is made by simply assigning the PAA value of its corresponding segment. Then, we define hot streaks as the periods when the success metric (approximated by PAA) is higher than a predefined threshold. The threshold value is specific for each time series and is calculated from the *activity rate* (AR), which is the ratio between the number of weeks in which the artist/genre appears on Hot 100 and the total number of weeks of the time series:

- $\mathbf{AR} \geq \mathbf{20}\%$: threshold is the *80th percentile* of the success metric;
- $\mathbf{15}\% \leq \mathbf{AR} < \mathbf{20}\%$: threshold is the *85th percentile* of the success metric;
- $\mathbf{10}\% \leq \mathbf{AR} < \mathbf{15}\%$: threshold is the *90th percentile* of the success metric;
- $\mathbf{AR} < \mathbf{10}\%$: threshold is the *95th percentile* of the success metric.

Figure 2 illustrates Beyoncé's career, as an example of the time series available in MUHSIC. From her success timeline measured by DCG (left), we detect three hot streak periods (right). Therefore, our dataset provides all such data for further success analyses.

3.2. Data Content

We organize MUHSIC in a relational schema to ease the understanding and further analyses. The dataset consists of 11 tables which contain all the information collected, curated, and enriched, as designed in Figure 3. Such tables may be classified into five categories: charts, songs, artists, genres, and associative tables. The first four categories represent the main elements of the musical ecosystem, and the associative ones connect them.

Charts. Music charts are the core of our dataset, as the success metric and the considered artists are all derived from them. Therefore, we put all Hot 100 charts in the *charts* table, which contains the ranking position, the track and artist names, which are used in the Spotify matching process. Other chart-related information extracted from Billboard includes peak position, number of weeks on the chart, and the previous week's position.

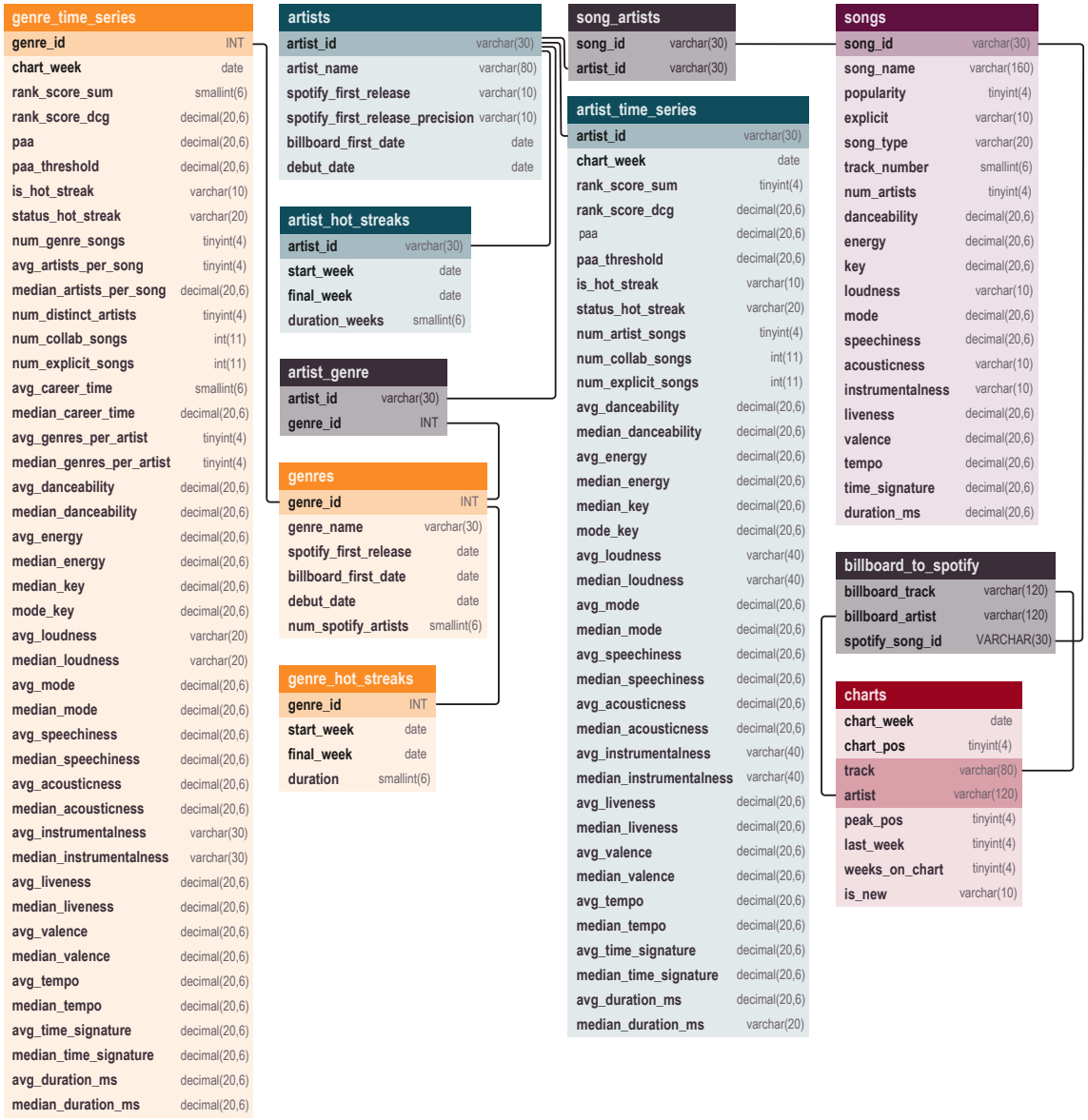

Figure 3. MUHSIC relational schema.

Songs. The table *songs* contains the detailed information of hit songs from Billboard Hot 100. All songs in the table were obtained from the Spotify matching explained in the previous section. Each record has relevant data, such as the *explicit* flag, which indicates whether the song lyrics contain language unsuitable for children. In addition, we make available acoustic features that describe aspects of the audio content, such as key, tempo, and *acousticness* (i.e., the probability of a song being acoustic).

Artists. This category contains three core tables with data about the artists themselves, their success time series, and their hot streak data. Specifically, table *artists* comprises the metadata obtained directly from Spotify and Billboard, such as the date of their first release (*spotify_first_release*) and the date of their first entry at Billboard (*billboard_first_date*). We consider the artist *debut_date* as the minimum between these two dates. Next, table *artist_time_series* describes the careers of all considered artists. For each artist and weekly chart, we calculate the aggregated rank scores in two ways (sum

Table 2. MUHSIC basic statistics: number of records and file size.

Table	Records	Size	Table	Records	Size
artists	6,066	1.5 MB	genres	998	80 KB
artist_genre	21,123	4.0 MB	genre_hot_streaks	2,248	128 KB
artist_hot_streaks	7,557	1.5 MB	genre_time_series	2,067,770	1.0 GB
artist_time_series	6,251,441	3.9 GB	songs	22,635	8.5 MB
billboard_to_spotify	26,919	5.5 MB	song_artists	26,790	5.0 MB
charts	322,380	29.6 MB			

and DCG), the PAA approximation, and the threshold used to define hot streak periods. The field *is_hot_streak* informs whether the week belongs to a hot streak or not. Besides, we provide the aggregated acoustic features for all songs that appear in that weekly chart. Finally, the table *artist_hot_streaks* summarizes the hot streak periods, informing the beginning and the end of each one, as well as its duration in weeks.

Genres. The genre category is similar to the artist one with three tables, as we also build genre time series and detect their hot streaks. Table *genres* contains the metadata from Spotify, including the number of artists belonging to a specific genre (*num_spotify_artists*). Table *genre_time_series* comprises the genre success time series, and most of the columns are similar to the corresponding artist table. However, we added new features that are possible due to this genre perspective. For instance, we calculate the number of distinct artists of the genre for each week (*num_distinct_artists*), as well as the average and median career time of such artists (*avg_career_time and median_career_time*). Then, table *genre_hot_streaks* summarizes the hot streak periods for the genre careers.

Associative Tables. The last category has the associative tables to represent many-to-many relationships and link tables from the other categories. For example, table *billboard_to_spotify* relates the Hot 100 entries to their Spotify correspondent instances. This relationship is made using columns *track* and *artist* from table *charts* and the *song_id* from table Spotify *songs*. Also, tables *song_artists* and *artist_genres* represent the list of artists who sing a hit song and the music genres from a given artist, respectively.

3.3. Exploratory Data Analysis

Here, we perform an exploratory data analysis over the new dataset, then presenting basic statistics that help understand the data. The final version of MUHSIC is composed of: 3,238 weekly charts; 22,635 distinct songs; and 6,066 artists belonging to 998 music genres. Such enriched and curated data enable to build success-based time series to explore artists' careers and genres' evolution. Table 2 summarizes basics statistics for each table.

One key feature of MUHSIC is the artists' music genre. Figure 4 presents the 25 most frequent genres in the dataset, according to the number of artists – in Spotify, the genre is linked directly to the artist, not to each song. We highlight the presence of widely popular super-genres such as *rock*, *pop*, *rap*, *r&b*, *soul*, and *country*. There is no standardization in the Spotify genres; hence, there are several derived genres that are also frequent, including *dance pop, brill building pop*, and *southern hip hop*. We use all such information to generate the genre success time series and to detect their hot streaks.

Next, we analyze the hot streak periods for artists through a simple characteriza-

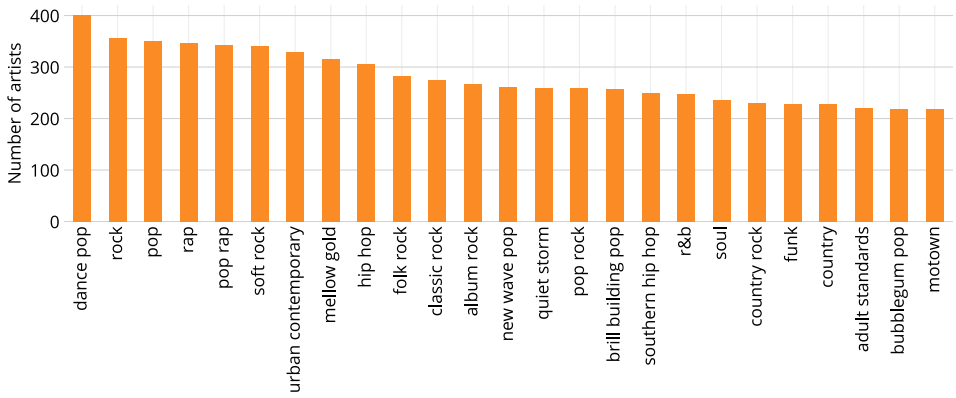

Figure 4. Top 25 music genres in MUHSIC sorted by the number of artists.

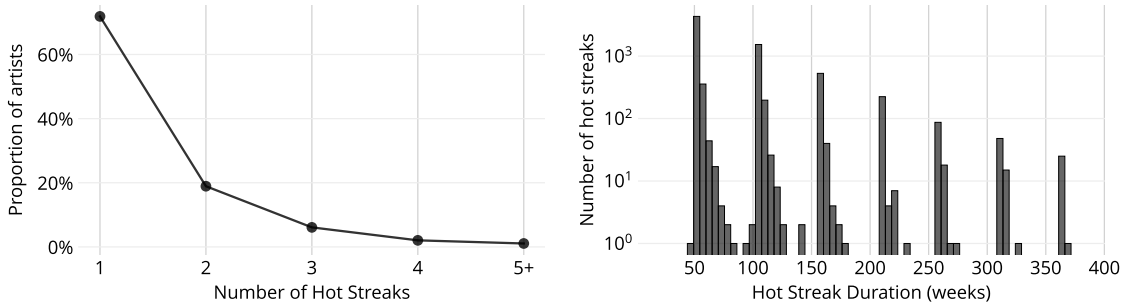

Figure 5. Hot streak statistics: hot streaks per artist (left) and duration (right).

tion analysis of table *artist_hot_streaks*. Figure 5 (left) shows the number of hot streaks (HS) per artists. In general, most artists (about 90%) have between one and two hot streak periods in their careers. Only a few manage to achieve more than five periods. Examples of such artists include Bruce Springsteen (8 HS), Michael Jackson (8 HS), and Mariah Carey (7 HS). Besides, most hot streaks last for around one year (52 weeks), as shown by Figure 5 (right). There are also many HS periods during two, three, and four years, suggesting a yearly pattern in this phenomenon. Note the log scale in the y-axis.

3.4. Format and Usage

MUHSIC is publicly available at Zenodo, an open dissemination research data repository [Oliveira et al. 2021]. Such a platform is aligned with the open science principle, as it shares, curates, and publicizes data and software for everybody. At Zenodo, the dataset is available in two distinct formats:

MySQL dump (.sql file), which creates the 11 tables of the relational schema from Section 3.2. This format is recommended for simple and complex queries, as MySQL is better at dealing with large amount of data; and

CSV files, which contains the same data of the relational schema. This format may be useful for processing data in Python or R to perform complex analyses and visualizations.

4. Applicability

In this section, we point out possible scenarios and applications that illustrate the potential impact and usability of MUHSIC for Music Information Retrieval (MIR) in Section 4.1

and Time Series Analysis (TSA) in Section 4.2. Finally, we also go over challenges and limitations in Section 4.3.

4.1. Music Information Retrieval (MIR)

Music Information Retrieval (MIR) is a multidisciplinary research area dedicated to developing computational tools for studying music-related data. Across diverse disciplines, MIR covers a wide range of research topics, including music classification, recommendation, genre identification, and others. In particular, the metadata and success-related content available in MUHSIC act as valuable resources to be used in different MIR tasks. Next, we highlight two examples where the dataset can be directly applied.

Hit Song Science (HSS). It is a debated subject within the MIR community, which aims to predict song success before being released [Pachet 2011]. The premise of HSS is hit songs comprise a specific feature set that makes them appealing to most people. Such attributes can then be exploited through machine learning methods to assess whether a song will become a chart-topping hit. In this context, MUHSIC is ready to be used as a labeled dataset for training and testing such methods. Also, MUHSIC allows a comprehensive exploration of similarities, patterns, and differences through learning models as well.

Music Genre Classification (MGC). It is one branch of MIR area, which aims to classify a song into one or more musical genres. Such a research topic is a nontrivial task, as the boundaries between genres remain blurred [Scaringella et al. 2006, Oliveira et al. 2020]. Overall, MUHSIC is a handful tool, as it allows an easy linkage of the song acoustic features to the genres of the artist who sings it. Moreover, the variety of music genres available in the dataset may assist in specializing the classification task, as it helps distinguishing between distinct sub-genres (e.g., *dance pop* and *indie pop* as *pop* sub-genres).

4.2. Time Series Analysis (TSA)

Success in the music industry has a temporal structure, as the audience tastes change over time. The dynamics of media platforms, the emergence of new music styles, and the artists' song releases are some factors that shape what listeners consume. In order to analyze such behavior, Time Series Analysis (TSA) is a direct application of the dataset. Such an application is useful for understanding and predicting artists' success according to one or multiple attributes as well [Janosov et al. 2020]. Here, we describe hot streak prediction as a potential TSA-related task which may benefit from MUHSIC.

Hot Streak Prediction. In the music industry, it is crucial for artists and record labels to direct their future releases to achieve or maintain their success levels. In such a context, hot streak prediction is a highly relevant task, as it allows predicting success based on historical data. Therefore, the success time series and hot streak data present in MUHSIC can be directly used in such a task, serving as input of prediction models (i.e., classification and/or regression). As an example, Oliveira [2021] uses MUHSIC in a model to assess the hot streak prediction task, achieving an F1-score of 0.761.

4.3. Challenges and Limitations

MUHSIC is not free from limitations, which may be improved in future versions. The key challenges relate to the heterogeneity of the data sources used in the data collection

and data integration phases. Another limitation is that the data sources consider only mainstream and popular music, generalizing the information, as discussed next.

Data Integration. It is one of the main issues in many Computer Science research fields (e.g., Hit Song Science) as there is no unique data source for all necessary features and information available. Also, the lack of an individual and universal identifier for each music makes integrating several data sources challenging. Data such as a given song's musical genre(s) is not standardized over data sources, mainly due to the blurred line between similar music styles. Specifically, to integrate data between Billboard and Spotify, the similarity functions to match such sources may not work well. Moreover, several songs on Billboard charts do not correspond with Spotify (older songs, contractual issues, etc.).

Data Quality. A problem faced in our dataset is the lack of accurate release dates for songs. For example, there are songs of which the release date precision is at the day level, whereas there are others in the month or year level. This is an issue that can induce bias, since it impacts the generation of the success time series and consequently the hot streak detection. One possible solution to tackle such a problem is relying on additional data sources, such as Wikipedia. However, this is not a trivial task, since it is directly related to the data integration problem.

Regional Markets. Most studies on HSS use data from the American market (e.g., Billboard Hot 100 Chart) probably because the United States is the biggest music market in the world. Studies that consider other music markets focus mainly on European countries, such as the United Kingdom. However, there are many other relevant markets with distinct characteristics and behavior, which require an individual analysis of success. For example, Japan, Australia, and Canada are among the top 10 music markets in the world,[8] with a vibrant music scene and popular regional genres. Therefore, as local engagement shapes the global environment, future work must consider the regional aspect, thus ensuring that music culture within such countries are accounted for [Oliveira et al. 2020].

5. Conclusion and Future Improvements

This paper introduced MUHSIC, an enhanced open dataset including metadata and musical success information regarding the main elements of the music ecosystem. We modeled our dataset in a relational schema with 11 tables containing all the collected, curated, and enriched information from Billboard and Spotify. Although MUHSIC shares much content with other music-related datasets, its novelty relies on temporal success by providing artist and genre time series as representatives of their careers. Such data is accessible and ready-to-use for complex tasks regarding Music Information Retrieval (MIR) and Time Series Analyses (TSA), including music genre classification and hot streak prediction.

As future improvements to overcome the limitations of our dataset, we first plan to consider additional data sources to extend further our feature sets, such as lyrics (Genius), awards (Grammy), and other relevant metadata (Wikipedia). Moreover, the next step would be including non-hit songs and artists to increase the data diversity. Finally, we intend to expand the market coverage of our dataset, moving from US-only to other regional markets since local engagement shapes the global music environment.

[8]IFPI Global Music Report: `http://gmr.ifpi.org/`

Acknowledgments. The authors would like to thank Prof. Anisio Lacerda for all relevant insights to model and build this dataset. This work was supported by CAPES, CNPq, and FAPEMIG, Brazil.

References

Aggarwal, C. C. (2016). *Recommender Systems - The Textbook*. Springer. doi:10.1007/978-3-319-29659-3.

Bertin-Mahieux, T. et al. (2011). The Million Song Dataset. In *Proc. of Int'l Society for Music Information Retrieval Conf. (ISMIR)*, pages 591–596.

Byrd, D. and Crawford, T. (2002). Problems of music information retrieval in the real world. *Information Processing & Management*, 38(2):249–272. doi:10.1016/S0306-4573(01)00033-4.

Çimen, A. and Kayis, E. (2021). A longitudinal model for song popularity prediction. In *DATA*, pages 96–104. SciTePress. doi:10.5220/0010607700960104.

Cosimato, A. et al. (2019). The conundrum of success in music: Playing it or talking about it? *IEEE Access*, 7:123289–123298. doi:10.1109/ACCESS.2019.2937743.

Garimella, K. and West, R. (2019). Hot streaks on social media. In *International Conference on Web and Social Media*, pages 170–180. AAAI Press.

Janosov, M., Battiston, F., and Sinatra, R. (2020). Success and luck in creative careers. *EPJ Data Sci.*, 9(1):9. doi:10.1140/epjds/s13688-020-00227-w.

Karydis, I., Gkiokas, A., and Katsouros, V. (2016). Musical track popularity mining dataset. In *IFIP AIAI*, pages 562–572. doi:10.1007/978-3-319-44944-9_50.

Keogh, E. J. and Pazzani, M. J. (2000). Scaling up dynamic time warping for datamining applications. In *ACM SIGKDD*, pages 285–289. ACM. doi:10.1145/347090.347153.

Liu, L., Wang, Y., Sinatra, R., Giles, C. L., Song, C., and Wang, D. (2018). Hot streaks in artistic, cultural, and scientific careers. *Nature*, 559(7714):396–399. doi:10.1038/s41586-018-0315-8.

Melchiorre, A. B. et al. (2021). Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5):102666. doi:10.1016/j.ipm.2021.102666.

Oliveira, G. P. (2021). Analyses of musical success based on time, genre and collaboration. Master's thesis, Universidade Federal de Minas Gerais, Brazil.

Oliveira, G. P., Barbosa, G. R. G., Melo, B. C., Silva, M. O., Seufitelli, D. B., Lacerda, A., and Moro, M. M. (2021). MUHSIC: An Open Dataset with Temporal Musical Success Information. *Zenodo*. doi:10.5281/zenodo.5168695.

Oliveira, G. P., Silva, M. O., Seufitelli, D. B., Lacerda, A., and Moro, M. M. (2020). Detecting collaboration profiles in success-based music genre networks. In *Procs. Int'l Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada.

Pachet, F. (2011). Hit song science. In Tao Li, Mitsunori Ogihara, G. T., editor, *Music Data Mining*, chapter 10, pages 305–326. CRC Press, New York, NY, USA.

Scaringella, N., Zoia, G., and Mlynek, D. (2006). Automatic genre classification of music content: a survey. *IEEE Signal Process. Mag.*, 23(2):133–141. doi:10.1109/MSP.2006.1598089.

Silva, M. O., Rocha, L. M., and Moro, M. M. (2019). MusicOSet: An Enhanced Open Dataset for Music Data Mining. In *SBBD DSW*, pages 408–417. SBC.

Zangerle, E., Huber, R., and Yang, M. V. Y.-H. (2019). Hit Song Prediction: Leveraging Low- and High-Level Audio Features. In *Proc. of Int'l Society for Music Information Retrieval Conf. (ISMIR)*, pages 319–326.