MusicOSet: An Enhanced Open Dataset for Music Data Mining

Mariana O. Silva, Laís M. Rocha, Mirella M. Moro

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

{mariana.santos,laismota,mirella}@dcc.ufmg.br

Abstract. *We present the* MusicOSet*, an open and enhanced dataset of musical elements (music, albums, and artists) suitable for music data mining. We describe the creation process and the data contents, along with usage examples and possible applications. The attractive features of* MusicOSet *include the enrichment of existing metadata and the popularity classification of the musical elements present in the dataset.*

1. Introduction

Individuals, organizations and governments are gradually realizing how the publication and availability of datasets can be useful. The fundamental role of datasets in several fields of research is irrefutable, especially for the initial progress of emerging topics and possibilities of experimental replications and thorough comparisons. For instance, public datasets are already an integral part of fields such as machine learning (Wine Quality [Cortez et al. 2009]), computer vision (ImageNet [Deng et al. 2009]), complex networks (SNAP [Leskovec and Krevl 2014]), social networks (GitSED [Batista et al. 2017]), digital libraries (DeduDLB [Silva and Brandão 2017]), biotechnology (MAMMOSET [Oliveira et al. 2017]) and Music Information Retrieval - MIR (MSD [Bertin-Mahieux et al. 2011]).

As in most scientific research fields, collecting and distributing datasets are important in MIR [Karydis et al. 2016]. Music Information Retrieval is a very important task in music data mining [Li et al. 2011]. Specifically, in such growing research domain, relevant musical content generally refers to audio files associated with lyrics, metadata and semantic information. While being a key piece in the progress of MIR research, the free distribution of such datasets and standardization are challenging tasks due to very restrictive copyright laws. However, to overcome these problems, many researchers follow an approach using free licenses (e.g., Creative Commons) [Goto et al. 2003; Defferrard et al. 2017] or just making acoustic feature vectors available [Bertin-Mahieux et al. 2011; Porter et al. 2015; Gemmeke et al. 2017], not audio data.

One issue in MIR is to apply multifaceted information from large musical databases for predicting hits. That and other issues evolved to a new research area called *Hit Song Science* (HSS), which aims to better understand the relationship between the intrinsic characteristics of the songs and their success. In other words, the goal is to predict whether a song offers the potential to become popular and commercially successful, thus reaching the top of the charts. In the MIR vision of HSS, the challenge is to gather a set of musical resources that can be mapped to music popularity. Once this mapping is ready, the process of predicting a new arbitrary song can be automated [Li et al. 2011].

There are different datasets in both MIR and HSS that cover a wide spectrum of the domain (see Table 1). However, none is directly applicable to extracting knowledge

of the popularity and intrinsic characteristics of musical elements (artists, songs and albums). Even worse, the main sources of music data extraction apply their respective track identification systems, making it challenging to collectively use multiple sources of musical data. Moreover, in most cases, there is no data available on less popular components. That is, the data collection contains only the popularity degree, with no information on the non-hits elements of the music industry.

To address the aforementioned challenges, we introduce the *MusicOSet*, an open and enhanced dataset of musical elements (artists, songs and albums) suitable for music data mining through the following features:

- Integration and centralization of different musical data sources;
- Calculation of popularity scores and classification of hits and non-hits musical elements, from 1962 to 2018;
- Enriched metadata for music, artists, and albums from the US popular music industry;
- Availability of acoustic and lyrical resources; and
- Unrestricted access in two formats (SQL database and compressed `.csv` files).

The remainder of this paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe the dataset, its creation processes as well as a detailed analysis of its content. We discuss how the data have been used and its applicability in Section 4. Next, we detail the potential challenges and limitations on using *MusicOSet* in Section 5 and conclude with future research directions in Section 6.

2. Related Work

There are numerous datasets publicly available that cover a broad spectrum in the music data mining area. These datasets provide plenty information related to music from different perspectives. However, most of them seek to provide content information (e.g., metadata, tags or acoustic features) by focusing on a particular purpose (e.g., recommendation systems, music information retrieval or music classification). Nevertheless, to embrace several tasks of music data mining, a dataset must provide a wide range of information in a centralized and easily accessible way, promoting the exploration of diverse musical aspects. Table 1 presents the most common datasets by comparing size (i.e., the total number of songs), metadata/acoustic features/lyrics/popularity data availability and release year, which is also the sorting field.

The RWC Music Database [Goto et al. 2003] is a copyrighted music database available specifically for search purposes. It was one of the first large-scale music databases containing six original components in different genres. However, the RWC size is currently considered small and it does not contain any further metadata. Another widely used dataset is the Computer Audition Lab 500-song (CAL500) [Turnbull et al. 2008]. CAL500 is a corpus of 500 tracks of songs chosen from a collection of western popular music authors. Each of the 500 songs was manually annotated by at least three people using a survey, with a total of $1,708$ musical annotations. Moreover, for each song, the dataset provides several features that have been extracted from audio files.

The Million Song Dataset (MSD) [Bertin-Mahieux et al. 2011] is perhaps one of the most used datasets in MIR. It provides audio features and metadata for one million contemporary popular music tracks. It stands out as one of the largest datasets currently available for research ends, totaling over 280 GB of data. Although MSD provides a great

Table 1: Comparison of the existing datasets

Dataset	Size	Metadata	Acoustic Features	Lyrics	Popularity Data	Year
RWC	365	yes	no	yes	no	2001
Cal500	500	no	yes	no	no	2007
MSD	1,000,000	yes	yes	no	no	2011
MusiClef	1,355	yes	yes	no	no	2012
TPD	23,385	yes	yes	no	yes	2014
Audio Set	2,084,320	yes	yes	no	no	2017
FMA	106,574	yes	yes	no	no	2017
HSPD	1,000,000	no	yes	no	yes	2019
MusicOSet	**20,405**	**yes**	**yes**	**yes**	**yes**	**2019**

deal of information, it is also criticized, mainly for the obscurity of the approaches used to extract content descriptors and the improbable integration of the different parts of the dataset. With a considerably reduced size, Schedl et al. introduced the MusiClef dataset [Schedl et al. 2013], a multimodal collection of professionally commented music. MusiClef includes editorial metadata, several audio features, annotation sets, collaboratively generated user tags, and MusicBrainz[1] identifiers to facilitate linking to other datasets.

Following a distinct approach to existing datasets, the Track Popularity Dataset (TPD) [Karydis et al. 2016] provides several sources of music popularity definition, within a period between 2004 and 2014. TPD also provides a mapping between different identification spaces, allowing the use of different data sources combined with metadata and contextual similarity information between tracks. More recently, the Hit Song Prediction Dataset (HSPD) based on the MSD was introduced [Zangerle et al. 2019]. With one million representative songs released between 1922 and 2011, the dataset also provides information about the MSD tracks that were included in the Billboard Hot 100 charts.

In a different perspective, Audio Set was introduced to bridge the gap in data availability between image and audio research [Gemmeke et al. 2017]. It is a large-scale dataset of hand-written audio events, which uses a carefully structured hierarchical ontology of 632 classes of literature-guided audio and manual curation. With a total of 2,084,320 songs, Audio Set exceeds MSD, becoming the largest set of music data. Concurrently, the Free Music Archive (FMA) was introduced as an open and easy-to-access dataset suitable to evaluate numerous music information retrieval (MIR) tasks [Defferrard et al. 2017]. The FMA consists of 343 days of audio and 917 GB, all under permissive Creative Commons licenses. It has complete metadata, including music title, album, artist and genres; user information, such as play counts, favorite items, and comments; along with high-quality audio files and some pre-calculated features.

As Table 1 shows, *MusicOSet* differs from all those datasets. It has more than 20 thousand songs, a regular size when compared to others. Nonetheless, what it lacks in number of songs, it makes up for in high quality information. In contrast to the datasets aforementioned, *MusicOSet* is the only one to provide all the attributes shown in Table 1.

[1] *MusicBrainz*: `https://musicbrainz.org/`

3. MusicSet

MusicOSet is an open and enhanced dataset of musical elements (artists, songs and albums) based on musical popularity classification. Provides a directly accessible collection of data suitable for numerous tasks in music data mining (e.g., data visualization, classification, clustering, similarity search, MIR, HSS and so forth). This section describes the entire creation and collection process, as well as its content, format and usage.

3.1. Creation Process

To create *MusicOSet*, the potential information sources were divided into three main categories: music popularity sources, metadata sources, and acoustic and lyrical features sources. Data from all three categories were initially collected between January and May 2019. Nevertheless, the update and enhancement of the data happened in June 2019.

Music Popularity Sources. Music popularity can be defined in different ways, including critics acclaim, social media and music platforms, sales profit, awards, etc. Another common approach is to rely on pop charts, such as the Billboard charts. To collect information of musical popularity, we used the *billboard.py*[2] Python API for access Billboard charts and perform the data collection. In total, we collected the last 56 years of the Hot 100 and Billboard 200 charts, ranging from 1962 (January 01, 1962) up to 2018 (December 31, 2018).

Metadata Sources. Subsequently, we used the Spotify, Genius, and Wikipedia platforms as sources of metadata and content. We choose these three web services since they all provide an API for research purposes. The *Spotipy*[3] library allows full access to all music data provided by the Spotify platform. For supplementary information, the *Wikipedia*[4] and *LyricsGenius*[5] libraries provide direct interfaces for accessing and analyzing data on Wikipedia and for songs, artists, and lyrics stored on Genius, respectively.

Acoustic and Lyrical Sources. To further enhance the *MusicOSet*, we included information on the lyrical and acoustic features of the collected songs. Acoustic fingerprints are condensed digital summaries of a song's phonic features [Ren et al. 2010]. These characteristics are the best measures available to capture the musical effect. That is, through acoustic features, the artistic style and creative experience are captured, characterizing the genome of a song. To collect information about the acoustic features of the songs from each album, we use the same *Spotipy* library. The fingerprints are produced by The Echo Nest[6], an online provider of musical intelligence that was acquired by Spotify in 2014. As for the lyrics information, we use the Python client for the Genius.com API, *LyricsGenius*.

Sources Integration. Determining when records are referring to the same real-world entity is essential in any Data Management effort that brings together data from multiple sources. This process is called *record linkage* and can be solved through probabilistic or fuzzy matching. Probabilistic text linkage is a very effective approach that uses string similarity functions, comparing two parts of the text and producing a single similarity metric. As

[2] *billboard.py*: `https://github.com/guoguo12/billboard-charts`
[3] *Spotipy*: `https://spotipy.readthedocs.io/`
[4] *Wikipedia*: `https://wikipedia.readthedocs.io/`
[5] *LyricsGenius*: `https://github.com/johnwmillr/LyricsGenius`
[6] *The Echo Nest*: `http://the.echonest.com/`

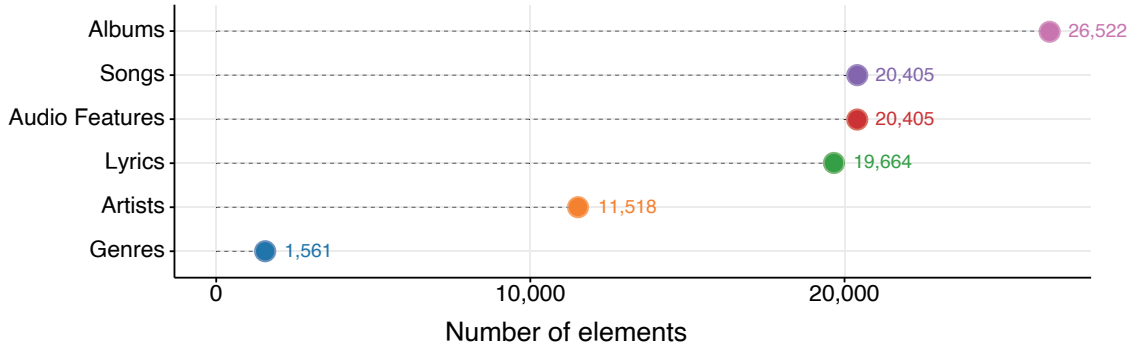

Figure 1: *MusicOSet* statistics

each data source has a different identification system, we used the *SequenceMatcher* class from the Python *difflib*[7] library, as well as the *Jaro-Winkler* algorithm from the *python-string-similarity*[8] library to map the music/artist/album records that refer to the same entity in all sources, with a similarity ratio of 0.9. However, because not all platforms incorporate information about all the gathered data, the mapping is not complete. In total, we were able to map approximately $82,5\%$ of the initial records collected.

3.2. Data Content

To facilitate the storage and visualization of data, we use a relational database management system (RDBMS) as a mechanism for storing the *MusicOSet* dataset. Figure 2 shows the database *schema*. Overall, it is composed of 13 tables that include the metadata/content, the acoustic and lyric features, and the popularity rating tables of the musical elements. Figure 2 also illustrates a division of the tables into three main segments: Popularity, Metadata and Song Features. Note that Popularity and Metadata are available in three different levels: artist (solo, duo or band), album (which is a collection of songs), and individual song. Such levels make the dataset more inclusive and easy to query, also enabling its use in different music data mining applications under varied aspects of the music industry. Figure 1 presents a quantitative description of the *MusicOSet* statistics.

Popularity. The popularity segment contains three tables (*ArtistChart*, *AlbumChart* and *SongChart*) consisting of ranking information collected from Billboard charts. Such tables include the inverted rating on a chart (*rank_score*); the highest rank achieved in any week of a year (*peak_position*); the number of weeks it has been on the charts in a year (*weeks_on_chart*); and the chart date. Specifically, the *SongChart* and *AlbumChart* tables were created from data collected from the Hot 100 and Billboard 200 charts, respectively. For creating the *ArtistChart* table, we weekly grouped the ranking information of the song/album artists featured in both previously mentioned Billboard charts. The other three remaining tables represent our popularity classification of the musical elements. Success may be measured by the presence of a song, album or artist on the ranking charts, such as Billboard. However, there is no equivalent for "unsuccessful songs" or "unpopular artists". With such restriction, there is no direct way to collect the unknown or less popular songs/albums/artists. To handle this limitation, we initially calculated a year-end score

[7] *difflib*: `docs.python.org/3.6/library/difflib.html`
[8] *python-string-similarity*: `github.com/luozhouyang/python-string-similarity`

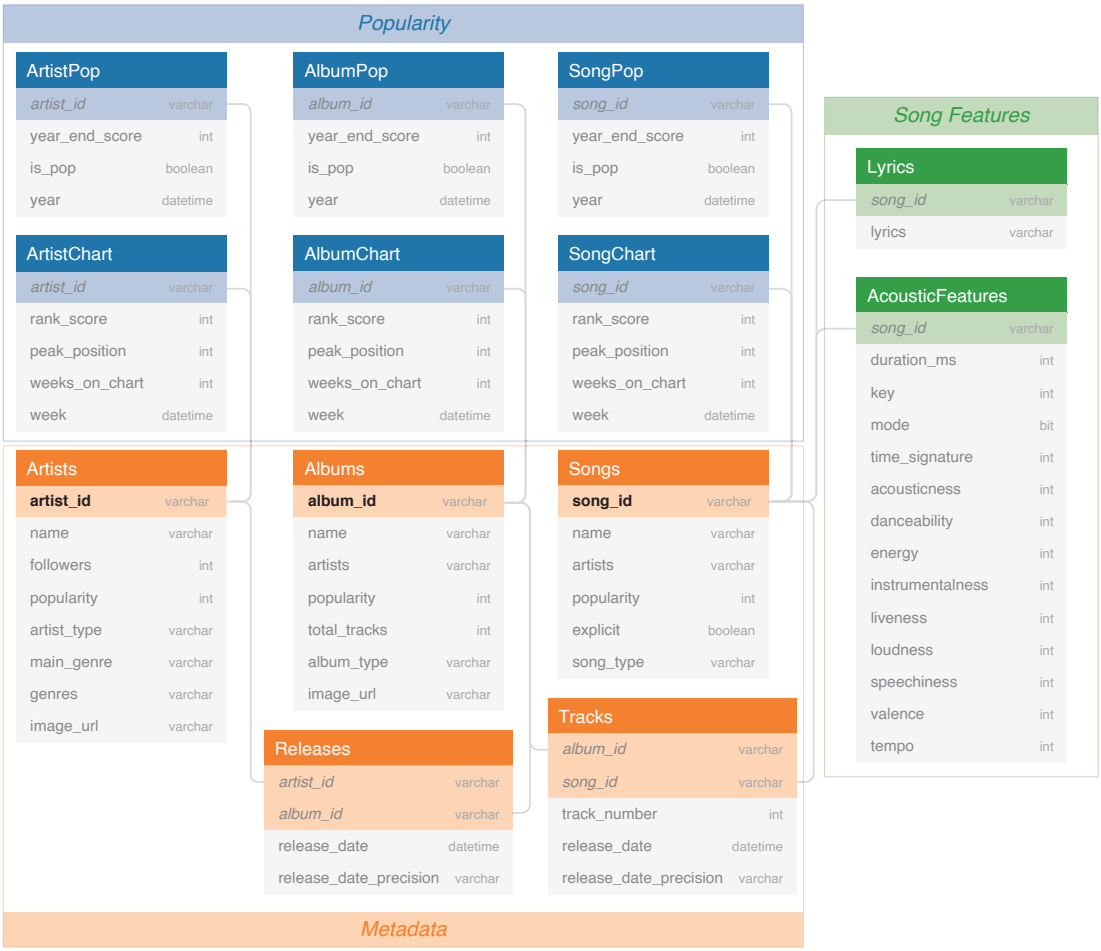

Figure 2: Schema for *MusicOSet*

that combines the scores of *peak_position* and *weeks_on_chart* and ranked the musical elements annually. Next, we assume that positive records (hits) correspond to the items that scored higher than the average for that year; whereas the negatives (non-hits) are the ones that obtained the lowest scores. Finally, we create a *boolean* field (*is_pop*), where *True* indicates a popular song/artist/album and *False*, the opposite.

Metadata. The metadata segment consists of textual and numeric information about songs, artists and albums. Basic information such as name, number of followers, popularity, and genre were collected directly from Spotify. The popularity field represents a value between 0 and 100, with 100 being the most popular. The song popularity is calculated by an algorithm and is based, in the most part, on the cumulative number of plays the track has and how recent those plays are. In other words, songs that are currently being played a lot receive higher popularity score than songs frequently played in the past only. Then, artist and album popularity are mathematically derived from song popularity. We also added information on types of song, artist and album. To capture the type of artists, we used the Wikipedia API to search for artist names and identify the presence of the terms “singer”, “band”, “duo”, “rapper” or “DJ”. As the type of songs, we distinguish only two types: solo songs (with only one artist present in its execution) or collaborative songs (where there is more than one artist). For the type of albums, we collected directly from Spotify,

which classifies the albums in three categories: album, single or compilation. To conclude, we also added URLs of artist images and album covers collected from both Spotify and Genius platforms. The remaining two tables (*Releases* and *Tracks*) were created to store album release information in the *Albums* table and song track information in the *Songs* table, respectively.

Song Features. Finally, the song features segment consists of only two tables: *Lyrics* and *AcousticFeatures*. The first table contains the lyrics of all songs present in the *Songs* table, which were collected using the *LyricsGenius* library. The second table contains acoustic fingerprints collected directly from Spotify[9]. Some acoustic fingerprints are objective (key, intensity, mode, tempo and time signature); others are more subjective (acousticness, danceability, energy, instrumentalness, liveness, speechiness and valence) and their values are calculated using The Echo Nest's music audio analysis tool [Jehan and DesRoches 2011]. We also consider the duration of the track as acoustic characteristics of a song.

3.3. Format and Usage

The *MusicOSet* is available including two separate parts: (A) a SQL file that creates the relational database and the 13 tables previously described (Section 3.1) and subsequently loads all data in the tables by a MySQL installation; (B) the same information as the tables in (A) but in `.csv` format to support fast use of the data and mitigate the need for a relational database. The full *MusicOSet* can be downloaded at `www.dcc.ufmg.br/~mirella/projs/bade`.

4. Applicability

A broad variety of music data mining tasks could be performed and analyzed using the *MusicOSet*. In this section, we share scenarios and possible applications, which help to illustrate the breadth and potential impact of the data available.

Metadata Analysis. One of the most direct applications for the dataset is the metadata analysis. Metadata analysis may involve, for example: music visualization, focusing on illustrating the metadata or acoustic contents; association mining, which refers to detecting correlations between different items in a set of data (e.g. acoustic characteristics, song lyrics, popularity, etc.); and clustering, which groups musical items into sets of similar objects based on their peer similarities. In addition to these tasks, other issues to explore over the dataset include: *How related are the artists? How have popular song lyrics changed over the years? Which musical genres have the highest average vocabulary? How different musical genres are correlated?* An example of a recent study that performs metadata analysis using *MusicOSet* is published in [Silva et al. 2019]. Through topological metrics and a clustering algorithm, the authors identified three well-defined communities with distinct collaboration patterns and notable discrepancies in levels of musical success. In addition, they found that successful artists are more likely to have profiles with a high degree of interaction and high diversification.

Hit Song Science. In the Hit Song Science (HSS) scenario, the main goal is to predict the success of songs before they are released. Therefore, researchers seek to identify features that make music more likely to be popular. There have been several notable studies on

[9]Spotify API Doc: `developer.spotify.com/documentation/web-api/reference/`

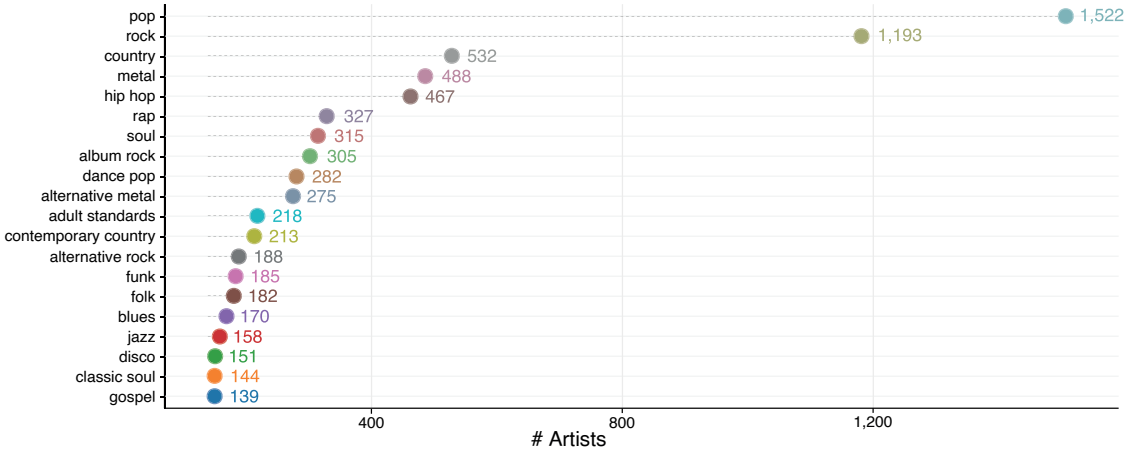

Figure 3: Distribution of artists in the 20 most common genres in *MusicOSet*

this topic, some of which focus on extracting acoustic and general lyric characteristics from songs, and then use standard classifiers to separate hits from non-hits. To study how the artists connect professionally can affect their musical success, Silva and Moro [2019] proposed a study using *MusicOSet* to assess whether there is a causal relationship between collaboration profiles and artist popularity.

Music Information Retrieval. Music Information Retrieval (MIR) is an emerging research area dedicated to meeting users' musical information needs [Li et al. 2011]. Musical recommendation and musical similarity are well explored issues in MIR due to the potential commercial value of a working system. The metadata and popularity information available on *MusicOSet* open up the possibility of a large-scale evaluation of song and music collaboration recommendation systems. Moreover, the available song features can assist the search for musical similarity. In principle, the user can provide an acoustic fingerprint set or a song lyric, and then the music search system will search for similar musical works based on the information provided by the user.

5. Limitations and Challenges

The *MusicOSet* is not free from limitations, which may be improved in future versions. The key challenges are related to the heterogeneity of the data sources used in the collection process. That is, due to the different identification systems, not all sources provide information about all the data gathered. Hence, some tables contain missing data. Another limitation is that the data sources consider only the mainstream and popular music, generalizing the information. Specifically, the data sources probably do not contain independent or less popular songs, artists or albums. Cultural and genre diversity is another issue: there is monopolization of US musical industry elements, as well as of pop and rock genres. This becomes explicit in Figure 3, which exposes the distribution of artists that have the 20 most common genres of the dataset.

In summary, although *MusicOSet* can be used to evaluate many tasks, some subsequent actions would further enhance the dataset. Additional features not considered in this first release, which are present in other datasets discussed in Section 2, could further enrich *MusicOSet*. For instance, the structure and content of the songs [Bertin-Mahieux et al. 2011], listener information [Schedl et al. 2013], extras artist metadata (e.g., related artists,

location, career time, etc.) or song metadata (e.g. track similarity, composer, publisher, genre, license, Spotify URL, etc.) [Karydis et al. 2016; Defferrard et al. 2017]. In addition, it would be critical to implement an automated Web-based collection and integration service that updates the dataset by capturing all sources.

6. Conclusion

This work introduces the *MusicOSet*: a cured, open and enhanced dataset of musical elements suitable for music data mining. Our contribution is related to integrating metadata, audio resources and musical popularity information. The *MusicOSet* is organized as a relational schema and made available in a public repository in two different formats. We also provide a statistical analysis of the dataset, as well as a discussion of the applicability and the main limitations in which the use of the *MusicOSet* involves. We believe that the dataset created along with the information described in this paper can be used for many music data mining tasks, such as MIR, classification, clustering, and prediction of successful songs (HSS).

Although *MusicOSet* remains two orders of magnitude behind the large-scale reference datasets analyzed here [Gemmeke et al. 2017; Bertin-Mahieux et al. 2011], for the best of our knowledge, it is the only one to provide a more complete set of attributes. In addition to providing enhanced metadata for songs, artists and albums from the US popular music industry, our dataset additionally makes available popularity scores, hit and non-hit ratings, as well as acoustic fingerprints and lyrics. All of this accessible in two formats (SQL database and compressed `.csv` files), with integration and centralization of different music data sources.

As future work, we plan to include new data sources to further expand and enrich *MusicOSet*, increasing the scope of potential applications. For instance, the added of different popularity sources, such as Last.fm and Spotify charts or even the number of Grammy Awards. Additionally, we plan to employ some traditional approaches to dealing with the missing data. A classic strategy would be to discard all elements for any sample that is missing one or more data components. The major problem with this method is the reduction of sample size. Therefore, we can alternatively fill in the missing data manually or input values using regression imputation. In the latter case, a regression model is estimated based on the observed values of variables to predict missing values. In other words, available information is utilized to predict the missing value of a specific variable.

Acknowledgments. The work is supported by CNPq, Brazil

References

Batista, N. A. et al. (2017). GitSED: Um Conjunto de Dados com Informações Sociais Baseado no GitHub. In *SBBD-Dataset Showcase Workshop*, pages 224–233.

Bertin-Mahieux, T. et al. (2011). The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval*, pages 591–596.

Cortez, P. et al. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

Defferrard, M. et al. (2017). FMA: A Dataset for Music Analysis. In *18th International Society for Music Information Retrieval Conference*.

Deng, J. et al. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Gemmeke, J. F. et al. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Goto, M. et al. (2003). RWC music database: Music genre database and musical instrument sound database. *Proceedings of the 4th International Conference on Music Information Retrieval*, pages 229–230.

Jehan, T. and DesRoches, D. (2011). Analyzer documentation. *The Echo Nest*.

Karydis, I. et al. (2016). Musical track popularity mining dataset. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 562–572.

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`.

Li, T., Ogihara, M., and Tzanetakis, G. (2011). *Music Data Mining*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition.

Oliveira, P. H. et al. (2017). MAMMOSET: An enhanced dataset of mammograms. In *SBBD-Dataset Showcase Workshop*, pages 256–266.

Porter, A. et al. (2015). Acousticbrainz: a community platform for gathering music information obtained from audio. In *16th International Society for Music Information Retrieval Conference*.

Ren, L. et al. (2010). Dynamic Nonparametric Bayesian Models for Analysis of Music. *Journal of the American Statistical Association*, 105:458–472.

Schedl, M. et al. (2013). A Professionally Annotated and Enriched Multimodal Data Set on Popular Music. In *Proceedings of the 4th ACM Multimedia Systems Conference (MMSys 2013)*, Oslo, Norway.

Silva, M. O. and Brandão, M. A. (2017). Deduplicação de Nomes e Redes de Co-autoria na DBLP. In *SBBD-Dataset Showcase Workshop*, pages 203–212.

Silva, M. O. and Moro, M. M. (2019). Causality Analysis Between Collaboration Profiles and Musical Success. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*. [to appear].

Silva, M. O., Rocha, L. M., and Moro, M. M. (2019). Collaboration Profiles and Their Impact on Musical Success. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 2070–2077, Limassol, Cyprus. ACM.

Turnbull, D. et al. (2008). Semantic Annotation and Retrieval of Music and Sound Effects. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):467–476.

Zangerle, E. et al. (2019). Hit Song Prediction: Leveraging Low- and High-Level Audio Features. In *Proceedings of the 20th International Society for Music Information Retrieval Conference 2019 (ISMIR 2019)*. [to appear].