IMPROVING GENRE ANNOTATIONS FOR THE MILLION SONG DATASET

Hendrik Schreiber
tagtraum industries incorporated
hs@tagtraum.com

ABSTRACT

Any automatic music genre recognition (MGR) system must show its value in tests against a ground truth dataset. Recently, the public dataset most often used for this purpose has been proven problematic, because of mislabeling, duplications, and its relatively small size. Another dataset, the Million Song Dataset (MSD), a collection of features and metadata for one million tracks, unfortunately does not contain readily accessible genre labels. Therefore, multiple attempts have been made to add song-level genre annotations, which are required for supervised machine learning tasks. Thus far, the quality of these annotations has not been evaluated.

In this paper we present a method for creating additional genre annotations for the MSD from databases, which contain multiple, crowd-sourced genre labels per song (Last.fm, beaTunes). Based on label co-occurrence rates, we derive taxonomies, which allow inference of top-level genres. These are most often used in MGR systems.

We then combine multiple datasets using majority voting. This both promises a more reliable ground truth and allows the evaluation of the newly generated and pre-existing datasets. To facilitate further research, all derived genre annotations are publicly available on our website.

1. INTRODUCTION

Automatic music genre recognition (MGR) is among the most popular Music Information Retrieval (MIR) tasks [5]. Until 2012, the majority of datasets used for MGR research was private and the most popular public dataset was GTZAN [13, 14]. Unfortunately, GTZAN has some documented deficiencies [12]. Additionally, with 1,000 excerpts from ten different genres, GTZAN is relatively small by today's standards. Desirable as dataset for MGR, in terms of size and available features, is the Million Song Dataset (MSD) [2]. But by 2012, when it was still very new, only three of the 345 publications (0.7%) surveyed in [13] had used it. This may be explained by the fact that the MSD does not contain explicit genre annotations. The authors of all three publications first had to derive song-level genre labels for a subset of the MSD as ground truth. For this purpose, Hu [7] and Schindler [10] both used album-level genre labels scraped from the All Music Guide website [1]. Dieleman et al. [3] selected 20 commonly used genres from the MusicBrainz artist tags contained in the MSD—an approach similar to what the MSD author suggested for the MSD Genre Dataset, a "simplified genre dataset from the Million Song Dataset for teaching purposes" [2]. With the exception of [10], the used ground truths aren't re-usable or well documented. And in the case of [10], they have not been evaluated and don't allow for multiple genre annotations per song.

In the spirit of [9], what is required to help facilitate MGR research using the MSD, is a song-level ground truth with a documented level of accuracy that also allows for ambiguity. In the following sections we will first derive (where necessary) and then compare four different genre datasets for the MSD. In Section 2 we describe how we created the beaTunes Genre Dataset (BDG). In Section 3, we apply a similar approach to the Last.fm Dataset [3], creating a Last.fm Genre Dataset (LFMGD). In Section 4 we explore to which degree the BDG, LFMGD and the datasets created by Hu (HO) [4] and Schindler (Top-MAGD) agree, and derive two new datasets, by combining multiple sources. Finally, in Section 5 and Section 6, we define benchmark partitions to promote repeatability of experiments using the new datasets and point to additional raw data.

2. BEATUNES GENRE DATASET

beaTunes [5] is a consumer application that encourages its users to correct song metadata using multiple heuristics. It also supports sending anonymized metadata to a central database, which matches it to metadata sent by other users. Much like tags on Last.fm, this allows keeping track of multiple user-submitted genres per song. For example, one song may have been associated with the label `Rock` by five users, while three users regarded the same song a `Pop` song. The database currently contains more than 870

[1] http://allmusic.com/
[2] http://labrosa.ee.columbia.edu/millionsong/blog/11-2-28-deriving-genre-dataset
[3] http://labrosa.ee.columbia.edu/millionsong/lastfm
[4] http://web.cs.miami.edu/home/yajiehu/resource/genre/
[5] http://www.beatunes.com/

million user song submissions of which 772 million are labeled with a genre and mapped to more than 85 million songs. Furthermore, the database stores each user's system language. In the remainder of this section we describe, how we used the existing genre labels to assign top-level genres (seeds) to each song and matched them to songs in the MSD.

2.1 Genre Label Normalization

In the beaTunes database, more than one million different, user-submitted genre labels are stored. Some of these are slight spelling variations of popular genre names like `Hip-Hop` or composites of multiple genres like `Hip-Hop/Rap`. Others describe custom categorization schemes, ratings, or are simply noise. In order to extract the most-used and thus most important genre labels from the database, we first normalized their names and then ranked them by usage count. The following normalization procedure was employed (building on [6]):

1. Convert to lowercase
2. Remove whitespace
3. Convert `'n'`, `and`, and `&` in different spellings of `R&B`, `D&B`, and `Rock'n'Roll` to `n`
4. Replace `alt.` with `alternative` and `trad.` with `traditional`
5. Tokenize with `+&/,;:!\[]()` as delimiters
6. From each token, remove all characters that aren't letters or digits
7. Sort tokens alphanumerically
8. Concatenate tokens with `/` as delimiter

This effectively treats composite labels like `Hip-Hop/Rap` as their own genre, but makes sure that `Hip-Hop/Rap` is equal to `Rap/Hip-Hop`. The special treatment in step 3 for `R&B`, `D&B`, and `Rock'n'Roll` is necessary, as the `&` character is also used as delimiter in composite labels (e.g. `Christian & Gospel`). After normalization, almost 700,000 different genre labels remain. However, 50% of all user-submitted songs are covered by the 16 most-used genres, 80% by the top 131 genres, and 90% by the top 750 genres.

2.2 Language-Specific Counts

Since genre labels reflect how listeners with a specific cultural background perceive music and what it means to them [1, 4, 8], we investigated how the collection's top genre rankings differ when taking the user's system language into account. Not surprisingly, by and large they are quite similar—with `Rock`, `Pop`, `Hip-Hop` and `Jazz` occurring in most top tens (Table 1). But there are a few notable exceptions. English speaking listeners are the only ones with `Country` (ranked 9th) in their top ten genres, French speakers rank `Reggae` (5th) higher than others, Spanish speakers rank `Latin` (5th), `House` (7th), and `Otros` (8th) high, and Japanese speakers rank `J-Pop` (3rd) near the top. Clearly, these differences are indicative of cultural preferences and should be taken into account when creating genre taxonomies. Therefore, in the remainder of this paper, we have only used the beaTunes label-submissions of English-speaking users.

2.3 Inferring Genre Taxonomies

As the beaTunes database contains on average about nine user submissions (i.e. genre labels) per song, we can record co-occurrences of labels on a per-song basis and thus infer relationships between them. Latent Semantic Analysis (LSA) with cosine similarity has been used for this purpose before [11]. But because we did not plan on using the cosine distance as metric, we did not deem it necessary to use Singular Value Decomposition (SVD) to keep the dimensionality low. Instead, we opted for a much simpler method. We filtered out rarely used labels and restricted ourselves to the top 1,000 genres covering over 93% of all user submissions with genre information.

Formally, we define $G := \{\texttt{Rock}, \texttt{Pop}, ...\}$ with $|G| = 1000 = n$ as the set of the n top genres, which are stored as distinct values in the vector $g \in G^n$ with $g := (\texttt{Rock}, \texttt{Pop}, ...)$. Each user submission is defined as a sparse vector $u \in \mathbb{N}^n$ with

$$u_i = \begin{cases} 1, & \text{if } g_i = \text{user-submitted genre} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

To establish the connection between a song s and its user labels u, we simply add up all u's belonging to the song and divide by the number of u's. Thus each song is represented by a vector $s \in \mathbb{R}^n$ with $0 \leq s_i \leq 1$ and $\sum_{i=0}^{n-1} s_i = 1$, denoting each genre's relative strength. To compute the co-occurrences for a given genre g_i with all other genres g, we element-wise average all s for which $s_i \neq 0$ is true. I.e.:

$$C_{g_i} := \bar{s}, \quad \text{for all } s \text{ with } s_i \neq 0; C \in \mathbb{R}^{n \times n} \quad (2)$$

The result is the matrix C that allows us to see how often a given genre co-occurs with another genre. Note that C is not symmetric as it would have been, had we used SVD with cosine similarity. So just because `Alternative` co-occurs with `Rock` fairly strongly ($C_{Alternative,Rock} = 0.156$), the opposite is not necessarily true ($C_{Rock,Alternative} = 0.026$, see Table 2). From the C values for the beaTunes database, it is also obvious, that `Rock` and `Pop` can be distinguished very well—both labels co-occur much more with themselves than with the other (`Rock`: 0.609/0.057, `Pop`: 0.593/0.077).

We exploit the asymmetry of C to construct a taxonomy by defining the following two rules:

(1) If a genre a co-occurs with another genre b more than a minimum threshold τ, and a co-occurs with b more than the other way around, then we assume that a is a sub-genre of b. More formally:

$$\begin{aligned} & a \text{ is a sub-genre of } b, \text{ iff} \\ & a \neq b \\ \wedge \quad & C_{a,b} > \tau \\ \wedge \quad & C_{a,b} > C_{b,a} \\ & \text{for all } a, b \in G \end{aligned} \quad (3)$$

	All ($N = 772.1$)	English ($N = 521.1$)	German ($N = 97.9$)	French ($N = 43.3$)	Spanish ($N = 27.1$)	Japanese ($N = 11.0$)
1.	Rock	Rock	Pop	Rock	Rock	Rock
2.	Pop	Pop	Rock	Pop	Pop	Pop
3.	Alternative	Alternative	Electronic	Jazz	Jazz	**J-Pop**
4.	Jazz	Hip-Hop/Rap	Hip-Hop	Hip-Hop	Soundtrack	R&B
5.	Hip-Hop	Hip-Hop	Jazz	**Reggae**	**Latin**	Soundtrack
6.	Hip-Hop/Rap	R&B	Alternative	R&B	Dance	Jazz
7.	Soundtrack	Soundtrack	Dance	Soundtrack	**House**	Electronica/Dance
8.	R&B	Jazz	R&B	Blues	**Otros**	ロック(Rock)
9.	Electronic	**Country**	Rock/Pop	Electronic	Blues	Altern. & Punk
10.	Country	Altern. & Punk	Soundtrack	Rap	Electronica	Hip-Hop/Rap

Table 1. Top ten genres used by beaTunes users with different languages. N denotes the number of submissions in millions.

Co-Occurrence Rank	1.	2.	3.	4.
Rock	Rock (0.609)	Pop (0.057)	Alternative (0.026)	Rock/Pop (0.016)
Pop	Pop (0.593)	Rock (0.077)	Rock/Pop (0.014)	R&B (0.013)
Alternative	Alternative (0.394)	Rock (0.156)	Pop (0.052)	Alternative/Punk (0.036)
R&B	R&B (0.566)	Pop (0.061)	Soul (0.036)	R&B/Soul (0.033)
Soundtrack	Soundtrack (0.754)	Rock (0.024)	Pop (0.022)	Game (0.011)
...	...	...	...	...

Table 2. Genre labels in the beaTunes database and their top four co-occurring labels ordered by relative strength given in parenthesis. The underlying values from the co-occurrence matrix C were computed taking only submissions by English speakers and the 1,000 most-used labels into account.

(2) Because this rule allows a genre to be a sub-genre of multiple genres, we add:

$$\begin{aligned} & a \text{ is a } \textit{direct} \text{ sub-genre of } b, \text{ iff} \\ & a \text{ is a sub-genre of } b \\ \wedge \quad & C_{a,b} > C_{a,c} \\ & \text{with } c \neq a \wedge c \neq b; a, b, c \in G \end{aligned} \tag{4}$$

By finding all direct sub-genres and their parents, we can now create a set of trees. The number of created trees depends on the threshold τ. We found, that to properly distinguish between genres like `Pop`, `Rock`, `Dance`, `R&B`, `Folk`, and `Other`, $\tau := 0.085$ proved to be useful, resulting in 141 trees. The roots of these trees are typically the names of seed-genres like `Jazz`, `Pop`, `Rock`, etc. (see Figure 1).

Not all generated trees have children. For example, the tree with the seed-genre `Groove` consists of just the root. Although `Groove` co-occurs with `R&B`, `Rock`, `Funk`, and `Soul`, the co-occurrence rates with genres other than itself are all below τ. Even the co-occurrence with itself is low (0.157). This suggests, that `Groove` is not really a genre, but more a property of a genre. Another example for a root-only tree is `Calypso`. Here the co-occurrence with itself is much higher (0.606) and indeed `Calypso` qualifies as stand-alone genre that simply does not have any sub-genres in this database.

Naturally, the generated taxonomies are only simplified mappings of the more complex relationship graph represented by C. In reality, genres aren't necessarily exclusive members of one tree or another (e.g. fusion genres). An ontology is the much better construct. But, as we will see, for the purpose of mapping most sub-genres to their seed-genre, trees are useful.

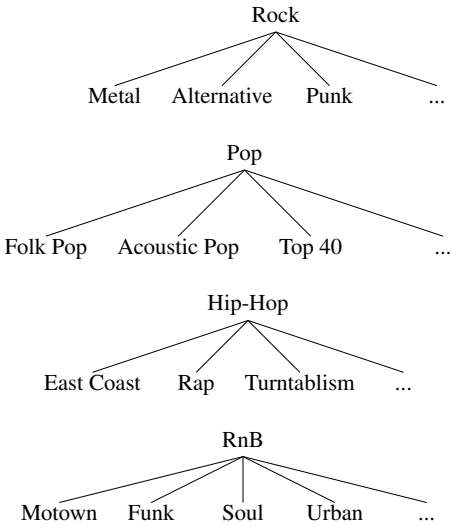

Figure 1. Partial, generated trees for the seed-genres `Rock`, `Pop`, `Hip-Hop`, and `R&B`.

2.4 Matching with Million Song Dataset

To create song-level genre annotations for the MSD, we queried the beaTunes database for songs with artist/title pairs contained in the MSD and were able to match 677,038 songs. In order to ease the comparison with the HO and Top-MAGD datasets, we associated each matched song with the seed-genre of its most often occurring genre label, taking advantage of the taxonomies created in Section 2.3. `Motown`, for example, is represented by its seed-genre `RnB`. In many cases, the found seed-genres are

Co-Occurrence Rank	1.	2.	3.	4.
rock	rock (0.128)	alternative (0.023)	pop (0.021)	indie (0.021)
pop	pop (0.107)	rock (0.037)	femalevocalists (0.024)	80s (0.018)
alternative	alternative (0.076)	rock (0.062)	indie (0.037)	alternativerock (0.023)
indie	indie (0.108)	rock (0.045)	alternative (0.034)	indierock (0.026)
electronic	electronic (0.119)	dance (0.026)	trance (0.021)	electronica (0.019)
...	...	...	...	...

Table 3. Tags in the Last.fm dataset and their top four co-occurring labels ordered by relative strength given in parenthesis, based on the co-occurrence matrix C, computed taking the 1,000 most-used labels into account.

equal to the All Music Guide labels used by Top-MAGD (`Blues`, `Country`, `Electronic`, `International`, `Jazz`, `Latin`, `Pop_Rock`, `Rap`, `Reggae`, `RnB`, `New Age`, `Folk`, `Vocal`). With a few exceptions: In our generated taxonomy for English users, `Blues` and `Vocal` are not seed-genres, but rather sub-genres of `Rock` and `Jazz`, respectively. Therefore, in these cases we used the label itself instead. We also translated `World` to `International`, and `Pop`, `Rock`, and `Pop/Rock` to `Pop_Rock`, and `Hip-Hop` to `Rap`. All songs we could not map to a Top-MAGD label were dropped, leaving us with 609,865 songs—90% of the originally matched songs. We call this dataset the beaTunes Genre Dataset (BGD).

3. LAST.FM GENRE DATASET

The Last.fm dataset is similar to the beaTunes database, in that it also contains multiple user-submitted labels per song which are each associated with a weight. Therefore we can use the same method to build a co-occurrence matrix and construct genre trees. The main difference lies in the kind of labels used. While the beaTunes labels are almost exclusively genre names, Last.fm tags vary a lot in content. Many are also genre labels, but others describe a mood, situation, location, time, or something completely different. As the dataset contains 522,366 different tags, it is not feasible to manually extract only the genre related ones. Therefore we again chose to incorporate the 1,000 most-used tags into computed genre trees. Because a single Last.fm song is often associated with many more tags than a beaTunes song with genre labels, we had to choose a different τ. Just like for BGD, we wanted to be able to see genres like `Electronic`, `Jazz`, `Pop` and `Rock` as seed-genres and therefore chose $\tau := 0.040$, which allows for this (see Table 3 for sample co-occurence values).

To create the Last.fm Genre Dataset (LFMGD), we associated each song with the seed-genre of the strongest tag that has a seed-genre corresponding to a Top-MAGD label or already corresponds to one of the Top-MAGD labels itself. In either case, we adjusted the spelling suitably. We also translated `hiphop` to `Rap`, and `pop`, `rock`, `poprock` to `Pop_Rock`, and `world` to `International`. Again, all songs not easily mappable to a Top-MAGD label were removed from the set. This left us with 340,323 (67.4%) of the 505,216 tracks originally labeled with at least one tag.

	Top-MAGD	LFMGD	BGD
HO	56.6%	*52.7%*	54.9%
Top-MAGD	-	75.8%	**84.1%**
LFMGD	-	-	81.0%

Table 5. Pairwise agreement rates for all four datasets for 136,639 MSD tracks occurring in all sets. The highest agreement is set in **bold**, the lowest in *italic*.

Dataset	Top-MAGD	LFMGD	BGD
Agreement Rate	90.4%	87.2%	**95.8%**

Table 6. Agreement rates for genre labels in Top-MAGD, LFMGD, and BGD when compared with the 133,676 tracks in CD1, found by majority voting.

4. CONSTRUCTING GROUND TRUTH

To construct a reliable ground truth, we evaluated agreement rates between the existing and constructed datasets using the genre labels from Top-MAGD. We then combined the more promising sets (Section 4.1). Because Top-MAGD labels as the lowest common denominator are somewhat unsatisfying, we then used just LFMGD and BGD to construct an additional dataset with finer genre granularity (Section 4.2).

4.1 Truth by Majority

After removal of duplicates [6], we found 136,639 tracks occurring in all four datasets Top-MAGD, LFMGD, BGD, and HO, all labeled with Top-MAGD genres. As a relative measure of trustworthiness, we calculated their pairwise agreement rate (Table 5). While the rates between Top-MAGD, LFMGD, and BGD are above 75%, those involving HO are below 57%. Unlike the other sets, HO was created with a combined classifier and is not the result of crowd-sourcing or any kind of expert annotation. Therefore a lower agreement rate was to be expected. The almost 20 percentage points difference illustrates that HO is not suitable as ground truth.

Since the other datasets were in relatively high agreement and we did not have a strong reason to believe, that

[6] `http://labrosa.ee.columbia.edu/millionsong/blog/11-3-15-921810-song-dataset-duplicates`

Top-MAGD	Blues	Country	Electronic	Folk	International	Jazz	Latin	New Age	Pop_Rock	Rap	Reggae	RnB	Vocal
Blues	**78.1%**	0.1%	0.0%	0.4%	0.1%	1.5%	0.0%	0.0%	**17.9%**	0.0%	0.0%	1.7%	0.0%
Country	0.0%	**86.1%**	0.0%	2.3%	0.1%	0.2%	0.0%	0.0%	**11.4%**	0.0%	0.0%	0.0%	0.0%
Electronic	0.0%	0.0%	**82.4%**	0.0%	0.4%	0.2%	0.1%	0.1%	**15.7%**	0.9%	0.0%	0.2%	0.0%
Folk	0.0%	0.8%	0.1%	**49.2%**	**14.9%**	0.0%	0.2%	0.1%	**34.3%**	0.0%	0.0%	0.0%	0.3%
International	0.1%	0.0%	7.8%	0.3%	**83.6%**	0.7%	0.8%	1.2%	4.5%	0.0%	0.0%	0.4%	0.7%
Jazz	0.1%	0.0%	2.2%	0.0%	0.5%	**76.2%**	1.2%	0.8%	6.5%	0.2%	0.0%	1.5%	**10.8%**
Latin	0.0%	0.0%	0.3%	0.0%	0.7%	0.3%	**95.6%**	0.0%	2.5%	0.2%	0.0%	0.0%	0.3%
New Age	0.0%	0.0%	2.7%	0.0%	1.4%	0.9%	0.0%	**93.5%**	1.5%	0.0%	0.0%	0.0%	0.0%
Pop_Rock	0.0%	0.1%	1.0%	0.2%	0.6%	0.1%	0.9%	0.0%	**96.0%**	0.1%	0.0%	0.8%	0.2%
Rap	0.0%	0.0%	3.0%	0.0%	0.7%	0.0%	0.2%	0.0%	4.5%	**91.0%**	0.0%	0.4%	0.0%
Reggae	0.0%	0.0%	2.2%	0.0%	1.7%	0.1%	1.0%	0.0%	**21.5%**	1.2%	**72.2%**	0.1%	0.0%
RnB	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	3.3%	0.4%	0.0%	**95.8%**	0.0%
Vocal	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.9%	0.0%	1.3%	0.0%	0.0%	0.0%	**95.8%**

BGD	Blues	Country	Electronic	Folk	International	Jazz	Latin	New Age	Pop_Rock	Rap	Reggae	RnB	Vocal
Blues	**97.6%**	0.0%	0.0%	0.2%	0.0%	0.1%	0.0%	0.0%	1.4%	0.0%	0.0%	0.6%	0.0%
Country	0.1%	**97.8%**	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	1.6%	0.0%	0.0%	0.1%	0.0%
Electronic	0.2%	0.0%	**91.2%**	0.0%	0.4%	0.3%	0.0%	0.6%	6.5%	0.6%	0.1%	0.2%	0.0%
Folk	0.4%	1.8%	0.0%	**93.9%**	0.2%	0.0%	0.0%	0.1%	3.6%	0.0%	0.0%	0.0%	0.0%
International	0.0%	0.2%	0.7%	0.9%	**93.8%**	0.5%	0.7%	0.5%	2.2%	0.0%	0.5%	0.0%	0.0%
Jazz	0.1%	0.0%	0.4%	0.0%	0.1%	**97.5%**	0.2%	0.1%	1.1%	0.1%	0.0%	0.4%	0.0%
Latin	0.1%	0.0%	0.5%	0.3%	1.6%	0.7%	**91.3%**	0.0%	4.9%	0.2%	0.3%	0.1%	0.0%
New Age	0.1%	0.0%	0.6%	0.1%	0.9%	0.4%	0.0%	**97.4%**	0.6%	0.0%	0.0%	0.0%	0.0%
Pop_Rock	0.3%	0.3%	1.3%	0.6%	0.1%	0.1%	0.2%	0.0%	**96.4%**	0.1%	0.1%	0.3%	0.0%
Rap	0.1%	0.0%	0.9%	0.0%	0.0%	0.1%	0.0%	0.0%	1.4%	**96.5%**	0.2%	0.8%	0.0%
Reggae	0.2%	0.0%	0.3%	0.0%	0.2%	0.0%	0.0%	0.0%	0.4%	0.5%	**98.3%**	0.1%	0.0%
RnB	0.1%	0.0%	0.2%	0.0%	0.1%	0.2%	0.0%	0.0%	3.8%	0.6%	0.0%	**94.9%**	0.0%
Vocal	1.3%	0.0%	0.0%	0.0%	0.4%	**16.3%**	0.4%	0.0%	**16.3%**	0.0%	0.0%	0.4%	**64.9%**

LFMGD	Blues	Country	Electronic	Folk	International	Jazz	Latin	New Age	Pop_Rock	Rap	Reggae	RnB	Vocal
Blues	**92.3%**	0.4%	0.2%	0.2%	0.0%	1.2%	0.0%	0.0%	5.3%	0.1%	0.2%	0.0%	0.0%
Country	0.2%	**91.8%**	0.0%	1.2%	0.0%	0.2%	0.0%	0.0%	6.4%	0.1%	0.1%	0.0%	0.0%
Electronic	0.0%	0.0%	**85.0%**	0.1%	0.2%	2.7%	0.1%	0.2%	9.8%	1.1%	0.7%	0.0%	0.1%
Folk	1.3%	3.7%	0.0%	**87.6%**	0.1%	0.6%	0.0%	0.0%	6.3%	0.0%	0.1%	0.0%	0.2%
International	0.1%	0.2%	2.4%	**17.2%**	**64.7%**	3.0%	1.4%	1.2%	7.9%	0.3%	1.4%	0.0%	0.3%
Jazz	0.4%	0.1%	0.5%	0.0%	0.3%	**95.1%**	0.1%	0.1%	2.8%	0.1%	0.1%	0.0%	0.3%
Latin	0.2%	0.2%	1.6%	1.2%	2.4%	3.8%	**59.0%**	0.1%	**29.6%**	0.6%	0.9%	0.0%	0.3%
New Age	0.1%	0.1%	**12.1%**	2.9%	1.4%	**17.6%**	0.9%	**54.8%**	9.8%	0.1%	0.0%	0.0%	0.1%
Pop_Rock	1.2%	1.1%	3.4%	2.8%	0.1%	1.0%	0.1%	0.1%	**88.9%**	0.4%	0.7%	0.1%	0.2%
Rap	0.0%	0.1%	1.4%	0.0%	0.0%	1.2%	0.3%	0.0%	4.0%	**92.2%**	0.3%	0.4%	0.0%
Reggae	0.0%	0.0%	0.2%	0.0%	0.1%	0.2%	0.1%	0.0%	2.2%	0.3%	**96.6%**	0.2%	0.0%
RnB	3.2%	0.2%	0.5%	0.1%	0.0%	9.1%	0.1%	0.0%	**20.5%**	3.9%	0.4%	**61.2%**	0.7%
Vocal	0.4%	1.7%	0.4%	0.8%	2.1%	**16.7%**	1.3%	1.3%	**25.9%**	2.1%	0.0%	0.0%	**47.3%**

Table 4. Confusion matrices between CD1 and Top-MAGD, BGD, and LFMGD. Values greater 10% are set in **bold**.

one of them is better than the other, we constructed a Combined Dataset 1 (CD1) from them using unweighted majority voting. CD1 contains only those tracks, that are labeled exclusively with the Top-MAGD genre set and for which the majority of labels from Top-MAGD, LFMGD, and BGD are identical. MSD duplicates were removed. Out of 136,991 tracks we found a majority genre for 133,676 (97.6% of all), of which 98,149 were found by unanimous consent (73.4% of majorities). To document ambiguity, we recorded both the majority decision and the minority vote, if there was one. This may be used in the evaluation of MGR systems, e.g. for fractional scores, or as indicator for uncertainty. The majority genre distribution of CD1 is shown in Figure 2. `Rock_Pop` is with 59.8% by far the most dominant genre, `Vocal` with 0.2% the most underrepresented one.

When comparing Top-MAGD, LFMGD, and BGD to the majority labels from CD1, we found that BGD matches best with 95.8%, followed by Top-MAGD with 90.4%, and LFMGD with 87.2% (Table 6). We believe that the relatively low agreement rate for LFMGD indicates room for improvement in the used mapping procedure from tags to genres, rather then problems with the original Last.fm dataset. Even though Top-MAGD was derived from album-level genre labels, it agrees with CD1 remarkably well, which attests to the quality of the set. BGD might be seen as the best of both worlds: its data source is song-level like LFMGD and at the same time somewhat limited to a genre vocabulary—more like Top-MAGD than LFMGD. This means the problematic mapping from free-form tags to genres is much easier. Overall, one might interpret these numbers as estimates for an upper boundary of MGR systems that test against a ground truth with only one genre label per song.

To provide more detail regarding the individual weaknesses of the datasets relative to CD1, we created confusion matrices (Table 4). In Top-MAGD the largest misclassifications occur for `Folk` (34.3%), `Reggae` (21.5%), `Blues` (17.9%), `Electronic` (15.7%), and `Country` (11.4%), which are all categorized as `Pop_Rock`. BGD classifies `Vocal` relatively poorly: 16.3% are misclassified as `Jazz` and 16.3% as `Pop_Rock`. LFMGD tends to misclassify `Latin`, `RnB`, and `Vocal` as `Pop_Rock` (29.6%, 20.5%, 25.9%), and `Vocal` as `Jazz` (16.7%). In summary, most errors occur with songs falsely identified as `Rock_Pop`. Additionally, `Vocal` tends to be misclassified as `Jazz`. We suspect this happens mainly, because `Vocal` is not seen as a genre, but rather as a style.

4.2 Truth by Consensus

Similar to Top-MAGD, almost 60% of all songs in CD1 are labeled `Pop_Rock`, Obviously, this rather coarse labeling is unsatisfying. Therefore we decided to create another

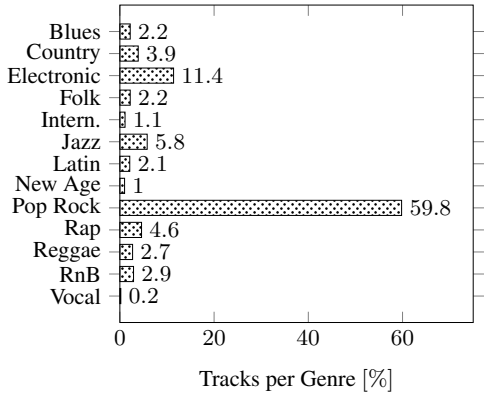

Figure 2. Majority genre distribution of tracks in CD1.

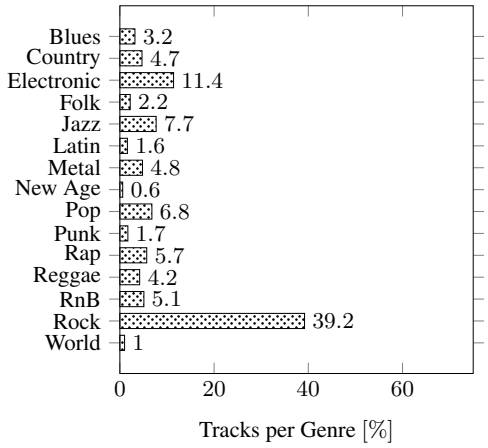

Figure 3. Genre distribution of tracks in CD2C.

dataset, Combined Dataset 2 (CD2), which differentiates between these two genres and adds two additional ones that are popular among users of beaTunes and Last.fm (`Metal` and `Punk`). Because `International` is hardly used in user-submitted tags and thus seems artificial, we used `World` instead. We also translated `Soul` to `R&B` in order to group them together, and removed `Vocal`, because it is the genre BGD and LFMGD confused most in CD1.

As sources for CD2 we used suitably modified versions of LFMGD and BGD and found 280,831 songs that both fit our genre-set, occur in both datasets, and aren't duplicates. 191,401 (68.2%) of the songs in CD2 have only one genre label, found by consensus. For convenience, we created another dataset called Combined Dataset 2 Consensus (CD2C) containing just those songs. As shown in Figure 3, the genre distribution for CD2C is a little more even than CD1—`Rock` being represented with a 39.2% share, `Pop` with 6.8%, and `New Age` with 0.6%.

5. BENCHMARK PARTITIONS

Inspired by [10] we provide three kinds of benchmark partitions for CD1, CD2, and CD2C in order to promote repeatability of experiments beyond x-fold cross validation. These partitions are:

- "Traditional" splits into training and test sets, with sizes 90%, 80%, 66%, and 50%; no stratification.
- Splits into training and test sets, with sizes 90%, 80%, 66%, and 50% and genre stratification.
- Splits with a fixed number of training samples per genre (1,000/2,000/3,000). Genres with fewer songs than the training size were dropped.

As CD2 songs are not always labeled with a majority genre, we used the first listed genre for stratification.

6. ADDITIONAL DATA

BGD and LFMGD represent simplified views on reality, suitable for comparisons with other, similar datasets like Top-MAGD. They both assign only one genre per song and the genre labels themselves are very limited. Both simplifications are problematic [9], which is why the combined datasets presented in this paper contain multiple genre labels where feasible. But for both BGD and LFMGD there is actually much more information available on a per-song basis. We are publishing it on our website in the hope that it proves useful for further research. Specifically, this includes:

- Multiple genre annotations/tags per song along with relative strength, and number of user-submissions to judge reliability.
- Co-occurrence matrices computed as described in Section 2.3.
- Derived genre taxonomies.

All data can be found at `http://www.tagtraum.com/msd_genre_datasets.html`.

7. CONCLUSION AND FUTURE WORK

Reliable and accessible annotations for large datasets are an important precondition for the development of successful music genre recognition (MGR) systems. Some often-used reference datasets are either relatively small or suffer from other deficiencies. To promote the adoption of the Million Song Dataset (MSD) for MGR research, we both evaluated existing and created two new genre annotation datasets for subsets of the MSD. Given that the large sizes of the datasets render manual validation almost impossible, we used either majority voting or consensus to validate existing data, and allowed for ambiguity in the created ground truths. In direct comparison with the generated ground truth CD1, 90.4% of the compared Top-MAGD labels were in agreement. To further promote experimentation and comparability, we also provided traditional and stratified benchmark partitions, as well as most of the data the combined datasets were derived from. In the process of creating the new datasets, we used simplifications like English-only labels and trees instead of graphs. Future work is needed to overcome these simplifications and better model the real world.

We hope the provided datasets prove useful for future publications in order to create better MGR systems.

8. REFERENCES

[1] Jean-Julien Aucouturier and Francois Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[2] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pages 591–596, 2011.

[3] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pages 669–674, 2011.

[4] Franco Fabbri. A theory of musical genres: Two applications. *Popular Music Perspectives*, pages 52–81, 1981.

[5] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *Multimedia, IEEE Transactions on*, 13(2):303–319, 2011.

[6] Gijs Geleijnse, Markus Schedl, and Peter Knees. The quest for ground truth in musical artist tagging in the social web era. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 525–530, 2007.

[7] Yajie Hu and Mitsunori Ogihara. Genre classification for million song dataset using confidence-based classifiers combination. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1083–1084. ACM, 2012.

[8] Jin Ha Lee, Kahyun Choi, Xiao Hu, and J Stephen Downie. K-pop genres: A cross-cultural exploration. In *Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR)*, pages 529–534, 2013.

[9] Cory McKay and Ichiro Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 101–106, 2006.

[10] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)*, pages 469–474, 2012.

[11] Mohamed Sordo, Oscar Celma, Martin Blech, and Enric Guaus. The quest for musical genres: Do the experts and the wisdom of crowds agree? In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 255–260, 2008.

[12] Bob L Sturm. An analysis of the gtzan music genre dataset. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12. ACM, 2012.

[13] Bob L Sturm. A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66. Springer, 2014.

[14] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.