Mining metadata from the web for AcousticBrainz

Alastair Porter
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
alastair.porter@upf.edu

Dmitry Bogdanov
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
dmitry.bogdanov@upf.edu

Xavier Serra
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
xavier.serra@upf.edu

ABSTRACT

Semantic annotations of music collections in digital libraries are important for organization and navigation of the collection. These annotations and their associated metadata are useful in many Music Information Retrieval tasks, and related fields in musicology. Music collections used in research are growing in size, and therefore it is useful to use semi-automatic means to obtain such annotations. We present software tools for mining metadata from the web for the purpose of annotating music collections. These tools expand on data present in the AcousticBrainz database, which contains software-generated analysis of music audio files. Using this tool we gather metadata and semantic information from a variety of sources including both community-based services such as MusicBrainz, Last.fm, and Discogs, and commercial databases including Itunes and AllMusic. The tool can be easily expanded to collect data from a new source, and is automatically updated when new items are added to AcousticBrainz. We extract genre annotations for recordings in AcousticBrainz using our tool and study the agreement between folksonomies and expert sources. We discuss the results and explore possibilities for future work.

CCS Concepts

•**Information systems** → **Digital libraries and archives;** *Multimedia databases; Data mining;* •**Applied computing** → **Sound and music computing;**

Keywords

Music Information Retrieval; Databases; Genre

1. INTRODUCTION

Many Music Information Retrieval (MIR) tasks require collections of audio annotated with semantic information (datasets). Traditionally these collections have been annotated manually, but with the increasing trend of working with ever larger datasets it becomes infeasible to annotate these datasets by hand. Often the focus of these smaller datasets is very narrow. For example, many datasets commonly used in genre classification contain no more than 10 genre labels [8]. It is also common for datasets to contain annotations covering only popular music. There are datasets containing more focused annotations which cover genres which are otherwise ignored in more general datasets, as well as datasets focused on specific musical properties (voice/instrumental music, gender, dark/bright timbre)[4].[1]

There has recently been discussion on the question of if in fact automatic recognition systems are actually identifying musical properties in the sense that we think of them. The most remarkable example is work by Sturm considering genre recognition systems [7]. Design of datasets is further complicated because of the subjective nature of some of the semantic categories, including genres or moods.

The web provides rich sources for both organized and free-form information about music, and a significant amount of MIR research is focused on exploring this data. In this paper we present a system for mining metadata and semantic information associated with recordings from the web. We then focus on a problem of genre annotation and present new datasets containing annotations created using our collected data. They include explicit genre annotations (AllMusic, Discogs, and Itunes) and genre annotations inferred from Last.fm. We evaluate agreement between Last.fm annotations and other sources.

2. MINING METADATA FROM THE WEB

In order to facilitate MIR experiments we decided to collect a dataset of music metadata and link it to existing audio analysis sources. To this end we developed MetaDB, a database and framework to collect and store annotations from many different online sources. We collect data related to the AcousticBrainz database [4].[2] AcousticBrainz is a community project containing MIR features extracted from audio files. Users who contribute to the project run software on their computers to analyze their personal audio collections and submit the analysis to the AcousticBrainz database. AcousticBrainz currently contains over 4 million submissions, covering about 2 million unique music recordings. Items in AcousticBrainz are referenced using MusicBrainz Identifiers (MBIDs). We initially collect metadata related to genre, style, and mood.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

[1] https://acousticbrainz.org/datasets/accuracy
[2] https://acousticbrainz.org

2.1 MetaDB infrastructure

The main interface to MetaDB is a webserver written in Flask.[3] External services can interact with it using two endpoints, *submit*, to add a new MBID to the database and *get*, to get metadata for a given MBID. We expect to create further endpoints to, for example, perform bulk lookups of metadata from a particular source.

Data is collected using *scrapers*, which are python modules following a specific API. Scrapers normally obtain data by connecting to an external website, although this is not a requirement. We can have more than one scraper per website (*source*) and they can retrieve data about recordings, releases, or artists. If the website does not understand MBIDs directly, we use metadata from MusicBrainz (e.g., artist name, track title) as the search criteria.

The scrapers may collect more information than is necessary for our task, so we have *filters*, which take the output of a scraper and transform it into data needed for a task. For example, we may collect folksonomy tags and then have a filter to extract only tags which represent genres or other semantic categories.

Data is stored in a PostgreSQL database.[4] Each item in the database references the MBID of a recording and the scraper used to collect the data. If a new scraper or filter is added, we can process all existing items in the database in a single batch process. When a submission is made to AcousticBrainz we automatically submit its MBID to MetaDB. Each scraper is run, collecting data for that recording. By continually running the scrapers we ensure that we always have up-to-date data.

The source code for MetaDB, including scrapers, is available under the GNU GPL v3 License.[5]

2.2 Data sources

Using MetaDB, we collected annotations and folksonomy tags from five different sources. We focused on sources which had structured data which was easy to collect. This data mainly contains genre and style annotations, but also mood and theme. While we only perform analysis on genre and style in this paper, this does not limit the future scope of our research and we anticipate collecting more types of data.

- **MusicBrainz**: a database of editorial metadata built by a community. It contains information about artists, releases, and recordings, as well as other detailed metadata and relationships. The scraper obtains the following data: A list of releases that this recording appears on and the date of each of the releases; A list of individual artists appearing on the recording and releases; Folksonomy tags on recordings, releases, and artists.
- **Discogs**: a similar community-built music database [1]. It contains information about relationships and performance roles on recordings (instruments, engineers, etc.). Discogs contains explicit genre and style annotations (genres are broad while styles are more specific). The Discogs community defines meanings for each genre and style, we may expect a consensus on genre use. Multiple genre and style labels per release are allowed. The scraper collects metadata at a release level. It queries the Discogs API with artist name, release title, and year. In the case that multiple releases are returned all are stored. The scraper obtains genre, style, label, and country and year of release.
- **AllMusic**: an online music guide [5] edited by an expert editorial staff (unlike MusicBrainz and Discogs). It contains a detailed list of genres, sub-genres, and styles, and many releases are annotated with one or more of each. AllMusic has no publicly available API, and therefore the scraper parses HTML data directly from the website. It performs an album search using the artist and release name and takes the first result where both have an exact match. The scraper collects genre, style, mood, themes and the album review if present.
- **Itunes**: an online music store with a public API for navigation of its content. The scraper performs a search using recording and artist metadata from MusicBrainz and selects the first result which has a title and artist name match. The scraper collects the "primaryGenreName" field, which is a single annotation provided by the API.
- **Last.fm**: a social music platform with collaborative tagging. It contains tag annotations and counts on a recording level. The tags are freeform and they tend to include commonly recognized genres. Tag counts are weighted, where the most applied tag for a recording has a weight of 100. We expect tags to represent the "wisdom of the crowds". The scraper uses the Last.fm API to get tag names and counts querying with artist and track names.

2.3 Statistics and data preprocessing

Among the recordings present in AcousticBrainz we were able to retrieve explicit genre annotations for 764,555 recordings from AllMusic, 720,597 from Discogs, and 957,529 from Itunes. Only 331,345 recordings were annotated by all three sources, which is less than half the size of all sources. Pairwise comparisons revealed that the largest intersection is between AllMusic and Discogs (537,786 recordings).

In all three sources genres are organized by trees. We collected the tree information from reference pages for AllMusic[6] and Itunes.[7] For Discogs we used the information found in the release submission interface[8] and the reference guide.[9] If an item was annotated by a more specific style from the tree but not the genre that this style is a part of (which sometimes happens for AllMusic and Itunes), we also added that genre as an annotation.

Annotations from all of our sources are of varying specificity. Annotations from AllMusic can fall into one of three levels (genre, sub-genre, sub-sub-genre), while Discogs and Itunes have genres and sub-genres. AllMusic contains the largest total number of genres (1186), followed by Discogs (491), and Itunes (253). The number of top level genres (roots) also varies (21 for AllMusic, 15 for Discogs, and 38 for Itunes). AllMusic contains many narrow genres including ones related to specific regional settings (e.g., "Midwest Rap"), Discogs and Itunes tend to be more generic in their categories. Some genre names appear multiple times in the same tree (e.g., "Electro" appears both under "Electronic" and "Hip-Hop" genres; "Latin Jazz" corresponds to both "Jazz" and "Latino" in Itunes).

In terms of coverage, in the data that we collected rock

[3] http://flask.pocoo.org/
[4] https://www.postgresql.org/
[5] http://github.com/MTG/metadb
[6] http://www.allmusic.com/genres
[7] https://affiliate.itunes.apple.com/resources/documentation/genre-mapping
[8] https://www.discogs.com/release/add
[9] https://reference.discogslabs.com/browse/genre

was the predominant genre, followed by pop, electronic, and jazz. The top 5 root genres are:

- **AllMusic**: Pop/Rock (62.34%), Electronic (10.20%), Jazz (9.15%), International (7.14%) and Classical (6.57%).
- **Discogs**: Rock (50.00%), Electronic (28.03%), Pop (7.78%), Jazz (7.65%), Classical (5.95%).
- **Itunes**: Rock (28.51%), Alternative (14.21%), Pop (11.16%), Electronic (7.27%), Jazz (6.20%).

We obtained Last.fm tags for 1,316,106 recordings, of which more than 67% contain genre-related tags according to string matching between tags and the genres from all of our sources. We perform basic preprocessing before matching strings. We then inferred three different sets of genre annotations from the tags by matching them to each genre tree.

For each recording, each tag is mapped to a genre and its parent genres. We preserve tag weights from the Last.fm API. The weight of a genre is the sum of its own weight plus the weights of all of its children. We inferred genre annotations for 841,571 recordings using the Discogs tree, 810,655 using AllMusic and 788,426 using Itunes. 717,151 recordings contained annotations in all three datasets.

We also collected tags for 392,881 recordings from MusicBrainz, however as this was significantly smaller than the other datasets we did not use it. We release datasets for Discogs, AllMusic, Itunes, Last.fm, and MusicBrainz, and the software used to perform our analysis.[10]

3. EXPERTS VS THE CROWD

We provide an initial exploration of the data that we collected. This analysis addresses the question of how folksonomy agrees with expert genre annotations. We can expect that annotations from different sources differ because of the experience of annotators and their familiarity with the content [3]. We took the intersection of recordings annotated by all four sources (Discogs, AllMusic, Itunes, and Last.fm) totaling 213,084 recordings. We refer to the datasets of each source as A, D, I for expert annotations and LA, LD, LI for inferred annotations from Last.fm.

3.1 Methodology

We test if Last.fm tags are a source of annotations which is as comprehensive as expert sources. We compare each of our pairs of datasets (A with LA, D with LD and I with LI). Each pair has an identical genre tree which make a direct comparison possible. We define three measures and three strategies for each pairwise comparison.

We apply lexical measures [2] to compare genre tree annotations between sources X and Y considering one of the trees (X) as a reference for each recording.

- **Lexical precision (P_X)**: the ratio between the number of genre matching both trees and the number of genre entities in Y
- **Lexical recall (R_X)**: the ratio between the number of genre entities matching both trees and the number of genre entities in X
- **Lexical F-measure (F)**

If we consider Y as a reference, lexical recall becomes a lexical precision and vice versa. Lexical F-measure is invariant to whether X or Y is a reference.

Lexical measures will treat genres of any level of specificity equally, and a match of a sub-genre will guarantee a match of its parents. We can compare trees bringing hierarchical information into the lexical measures using three strategies:

- **Contains**: consider all genre entities in lexical measures.
- **Leaves**: consider only leaf genre entities in lexical measures. This is a pessimistic strategy as it penalizes more specific annotations by not counting matches for their parents as correct.
- **Roots**: consider only top-level genre entities in lexical measures. We expect it to be an optimistic strategy because it works on a broader level avoiding conflicts on the sub-genre level.

During the initial process of creating the Last.fm genre annotations we kept all tags regardless of their weight. We repeat our evaluations with an increasing threshold value from 0 to 100 discarding genres whose weights fall below the threshold. We compute lexical measures using our strategies averaged across recordings matching annotations in A and LA, D and LD, and I and LI.

	P	R	F
AllMusic (A)			
contains	0.06	0.44	0.10
leaves	0.10	0.09	0.09
roots	0.49	0.59	0.53
Discogs (D)			
contains	0.08	0.61	0.15
leaves	0.28	0.13	0.18
roots	0.69	0.75	0.72
Itunes (I)			
contains	0.44	0.34	0.38
leaves	0.16	0.11	0.13
roots	0.44	0.38	0.41

Table 1: Lexical precision, recall and F-measure for the considered strategies at threshold=100 with A, D and I as a reference.

3.2 Results and discussion

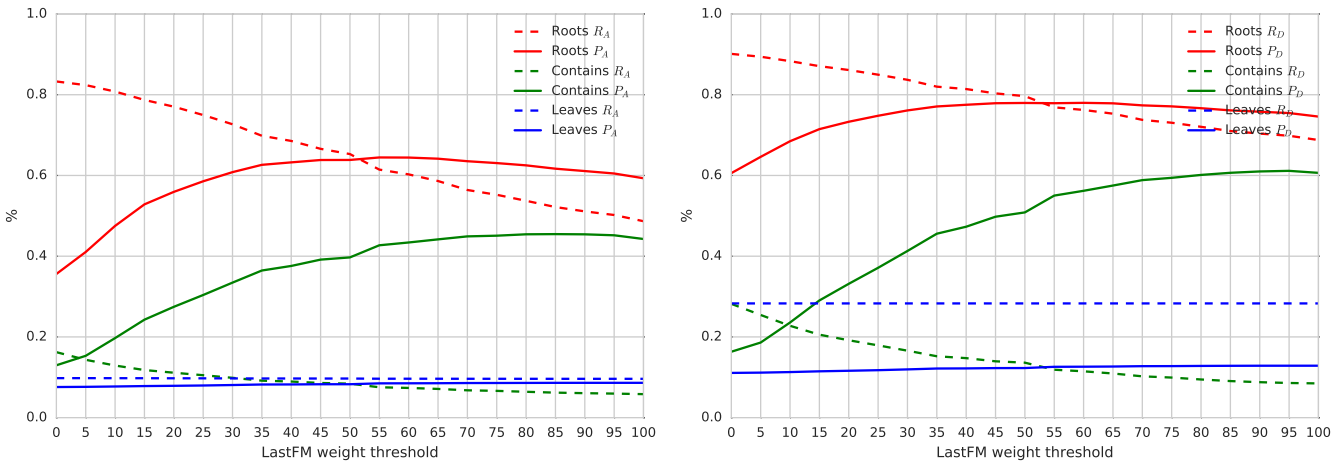
We present results for our comparison in Figure 1. As can be expected, we observe that recall falls as tags are gradually removed. Even with all of the tags in Last.fm (threshold=0) we only achieve a maximum recall of 80% to 90% for top-level genres (roots). This shows that user tags do not always cover the expert annotations, at least when using our method of matching tags to genres. When we also consider sub-genres, the maximum recall is below 20%.

Precision increases as tags are removed, which indicates that low-weighted tags do not contribute to the definition of genres according to our expert annotations. We see a decrease in precision as too many tags are removed, which suggests that a reasonable threshold for tag filtering is 50 to 55%.

When we consider our most specific annotations (Leaves), we see that there is little agreement. This value changes little as the threshold changes, indicating that the few matches which exist between the sources have high tag weights and so are not filtered.

Table 1 presents precision, recall and F-measure at threshold $t = 100$. According to the F-measures, the best agreement between sources was achieved when comparing top-level genres (roots). Folksonomy tags agree more with Discogs, perhaps because it contains less top level genres than AllMu-

[10]http://labs.acousticbrainz.org/dlfm2016



(a) AllMusic (A vs LA)

(b) Discogs (D vs LD)

Figure 1: Lexical precision and recall for the considered strategies with increasing threshold.

sic. Further investigation is needed to discover the reasons for poor agreement.

We are matching tags to genres by matching strings, with only a small amount of pre-processing. This simple process may be reducing the number of tags that we match to genres. We observed a number of false matches (e.g., "instrumental" occurs in Discogs as a sub-genre of "Hip-Hop", although it is a generic term which occurs in tags for many recordings). Our datasets are biased toward few genres (e.g., Pop and Rock) and we should consider separate analysis for each top-level genre.

4. CONCLUSIONS AND FUTURE WORK

We have presented a database system for scraping and storing metadata for music recordings. This database uses MusicBrainz as a source of stable identifiers, and to obtain textual metadata to match to other websites. We collected genre annotations for recordings from a number of different sources and stored them in the MetaDB database. The MetaDB software can be extended to collect metadata of any kind from any website. We want to add more scrapers to collect additional information including moods, themes, instrumentation and reviews.

Using the data that we collected we performed preliminary analysis of how folksonomy tags agree with expert genre annotations. We found agreement when comparing broad genre descriptors, but as the granularity of genre definitions increases, agreement between sources decreases.

We observed a disagreement between social tags and expert genre annotations, but we have not yet considered disagreement between all of the sources, nor have we looked at specific reasons for disagreement. In future work we will analyze confusions in genre annotations between expert sources. Some similar work has already been done but only considering broad genre categories and single-label annotation [6]. Grouping similar genres is another challenge to be considered. Our ultimate goal is to improve and extend the audio analysis data available in AcousticBrainz. In particular we need metadata to create new and more accurate models for semantic annotation of music.

5. ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688382.

6. REFERENCES

[1] D. Bogdanov and P. Herrera. Taking Advantage of Editorial Metadata to Recommend Music. In *International Symposium on Computer Music Modeling and Retrieval)*, 2012.

[2] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.

[3] A. J. Craft, G. A. Wiggins, and T. Crawford. How many beans make five? the consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In *International Society for Music Information Retrieval Conference*, 2007.

[4] A. Porter, D. Bogdanov, R. Kaye, R. Tsukanov, and X. Serra. AcousticBrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval Conference*, 2015.

[5] A. Schindler, R. Mayer, and A. Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *International Society for Music Information Retrieval Conference*, 2012.

[6] H. Schreiber. Improving genre annotations for the million song dataset. In *International Society for Music Information Retrieval Conference*, 2015.

[7] B. Sturm. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 10 2014.

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