MUSICLEF: A BENCHMARK ACTIVITY IN MULTIMODAL MUSIC INFORMATION RETRIEVAL

Nicola Orio
University of Padova
orio@dei.unipd.it

David Rizo
University of Alicante
drizo@dlsi.ua.es

Riccardo Miotto, Nicola Montecchio
University of Padova
{miottori,montecc2}@dei.unipd.it

Markus Schedl
Johannes Kepler University
markus.schedl@jku.at

Olivier Lartillot
Academy of Finland
olartillot@gmail.com

ABSTRACT

This work presents the rationale, tasks and procedures of MusiCLEF, a novel benchmarking activity that has been developed along with the Cross-Language Evaluation Forum (CLEF). The main goal of MusiCLEF is to promote the development of new methodologies for music access and retrieval on real public music collections, which can combine content-based information, automatically extracted from music files, with contextual information, provided by users via tags, comments, or reviews. Moreover, MusiCLEF aims at maintaining a tight connection with real application scenarios, focusing on issues on music access and retrieval that are faced by professional users. To this end, this year's evaluation campaign focused on two main tasks: automatic categorization of music to be used as soundtrack of TV shows and automatic identification of the digitized material of a music digital library.

1. INTRODUCTION

The increasing availability of digital music accessible by end users is boosting the development of Music Information Retrieval (MIR), a research area devoted to the study of methodologies for content- and context-based music access. As it appears from the scientific production of the last decades, research on MIR encompasses a wide variety of different subjects that go beyond pure retrieval: the definition of novel content descriptors and multidimensional similarity measures to generate playlists; the extraction of high level descriptors – e.g. melody, harmony, rhythm, structure – from audio; the automatic identification of artist and genre. As it is well known, the possibility to evaluate the different research results using a shared dataset has always played a central role in the development of information retrieval methodologies, as it is witnessed by the success of initiatives such as TREC and CLEF, which focus on textual documents.

The same need has been perceived in MIR, motivating the development of an important evaluation campaign, the Music Information Retrieval Evaluation eXchange (MIREX). MIREX campaigns [1] are organized since 2005 [4] by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Due to the many limitations posed by the music industry, the organizers of the MIREX chose to distribute only publicly available test collections. Participants are in charge to create their own collection and after local experimentation submit their software that is run by the organizers. This approach has two drawbacks, which have already been debated by the MIR research community: the results of previous campaigns cannot be easily replicated and the results depend on the individual training sets and not only on the submitted algorithms.

A recent relevant initiative, that aims at overcoming the limitations imposed by not sharing the datasets between researchers, is the Million Songs Dataset (MSD). Thanks to MSD [2], researchers can access a number of features from a very large collection of songs [2]. Unfortunately, the algorithms used to extract these features are not public, limiting the possibility to carry out research on content description techniques. Another ongoing initiative related to the evaluation of MIR approaches is the Networked Environment for Music Analysis (NEMA), that aims at providing a web-based architecture for the integration of music data and analytic/evaluative tools [3]. NEMA builds upon the achievements of MIREX campaigns regarding the evaluation of MIR approaches, with the additional goal of providing tools for resource discovery and sharing.

Within this scenario, MusiCLEF is an additional benchmarking initiative, that has been proposed in 2011 as part of the activities of the Cross-Language Evaluation Forum (CLEF). CLEF focuses on multilingual and multimodal retrieval [4] and gathers researchers in different aspect of information retrieval, ranging from plagiarism and intellectual property rights to image retrieval.

The goal of MusiCLEF is to promote the development of

[1] http://www.music-ir.org/mirex
[2] http://labrosa.ee.columbia.edu/millionsong/
[3] http://www.music-ir.org/?q=nema/overview
[4] http://clef-campaign.org/

novel methodologies for music access and retrieval, which can combine content-based information, automatically extracted from music files, with contextual information, provided by users through tags, comments, or reviews. The combination of these two sources of information is still under-investigated in MIR, although it is well known that content-based information alone is not able to capture all the relevant features of a given music piece (for instance, its usage as a soundtrack or the year of release), while contextual information suffers from the typical limitations for new items and new users (also known as cold start).

Aiming at investigating and promoting research on the combination of textual and music information, MusiCLEF has a strong focus on multimodality that, together with multilingualism, is the main objective of the CLEF evaluation forum. Moreover, the tasks proposed for MusiCLEF 2011 are motivated by real scenarios, discussed with private and public bodies involved in music access and dissemination. In particular, MIR techniques can be exploited for helping music professionals to describe music collections and for managing a music digital library of digitized analogue recordings. To this end, the organizers of MusiCLEF exploited the ongoing collaborations with both a company for music broadcasting services (LaCosa s.r.l.) and a public music library (University of Alicante's Fonoteca).

Two tasks are proposed within MusiCLEF 2011, and both are based on a test collection of thousands of songs in MP3 format. To completely overcome copyright issues, only low-level descriptors will be distributed to participants. Figure 1 depicts the tasks workflow of MusiCLEF, which is described in more detail in the following sections.

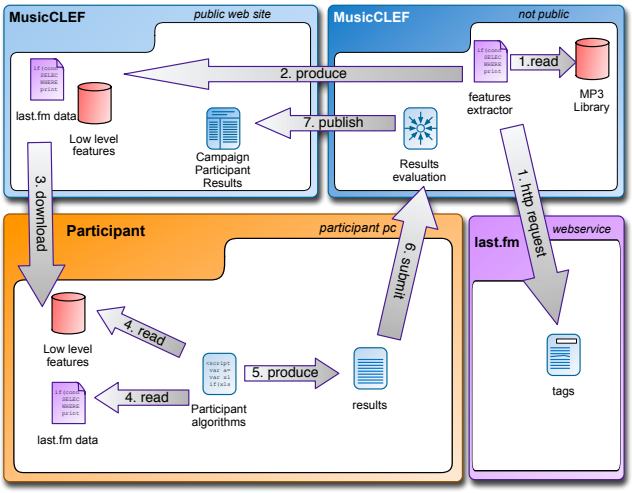

Figure 1: Task workflow in MusiCLEF.

It is important to note that, although the audio files cannot be distributed, the goal of MusiCLEF is to grant the participants with complete access to music features of the test collection. This means that the algorithms used to extract the music descriptors are public – and in particular are based on the set of tools provided by the MIRToolbox – but also that participants can submit their own original algorithm for feature extraction, that will be run locally. Therefore, MusiCLEF goals are to fill the gap between the other important initiatives in MIR evaluation: researchers can test and compare their approaches using a shared number of tasks, as in MIREX, while accessing a shared collection of content descriptors, as in MSD.

2. APPLICATION SCENARIOS

As mentioned in the previous section, a major goal of MusiCLEF is to maintain a tight connection with real application scenarios, in order to promote the development of techniques that can be applied to solve issues in music accessing and retrieval that are faced by professional users. The choice of focusing on professional users is motivated by the fact that they need to address a number of real-life issues that are usually not taken into account by music accessing systems aimed at the general public. At the same time, the evaluation of the effectiveness of the proposed automatic solution is easier to assess, because professional users have a clear idea of what are their information needs.

In the following we present the two professional partners of MusiCLEF, and we also describe the motivations that induced us to organize the two tasks mentioned in the previous section.

2.1 LaCosa s.r.l.

LaCosa was founded as a service provider of the major TV broadcasting – public and private – companies in Italy with the goal of managing and describing a large music collection of songs to be used for TV programs, including jingles, background and incidental music, and music themes for TV shows. LaCosa has a strong cooperation with RTI, a company that, apart from buying and storing songs issued by the major record companies, produces its own music catalogue. At present, RTI library contains about 320,000 songs of pop-rock, jazz, and classical music. Besides playing the role of music consultant, being one of the biggest private music repositories in Italy, RTI offers a number of services to external companies of music consultants, who can browse remotely the repository. Audio features distributed to the participants are thus extracted remotely, without downloading the audio files.

The typical job of a music consultant is to select a list of songs that are suitable for a particular application, for instance a TV commercial, the "promo" of a new program, the background music for a documentary, and so on. The availability of large online collections, such as Last.fm and YouTube, is representing an alternative to the services of a music consultant. For instance, journalists are increasingly selecting by themselves the music for their news stories, instead of asking to music consultants. The goal of LaCosa is

then to provide high quality descriptions, that are tailored to the particular application domain, in order to represent still a more interesting alternative to free recommendations.

Given these considerations, the requirements of LaCosa can be summarized as follows: How to improve the acquisition process, extracting the maximum amount of information about music recordings from external resources? How to provide good suggestion about possible usages of music material, minimizing the amount of manual work?

Because of the interest on the development of automatic systems for addressing these two requirements, LaCosa decided to provide at its own expenses a number of assessors to create the ground truth for evaluation. The involvement of professional users included also the definition of a vocabulary of 167 terms describing music genre (terms are organized in two levels, genre and subgenre), and of 188 terms describing the music mood. It is important to note that, in this case, the concept of mood is related to the usage of a particular song within a video production. As explained in more detail in Section 3, only a subset of the mood tags have been used in the evaluation campaign.

2.2 University of Alicante's Fonoteca

Some years ago, the local radio broadcast station *Radio Alicante Cadena Ser* transferred its collection of vinyls to the Library of the University of Alicante. This collection contains approximately 40,000 vinyls of an important cultural value, containing a wide range of genres. The library decided to digitize the vinyls, sound and covers, to overcome the preservation problems when allowing library users to access the discs and to enable its reproduction embedded in the library's Online Public Access Catalog (OPAC) with the name *Fonoteca* [5] .

The process was carried out following library cataloguing techniques to make the inventory of the collection. Vinyls were catalogued using Universal Decimal Classification, and classified into subjects based on the Library of Congress subject headings. Digitized covers and audio were linked to the corresponding records. The cataloguing data consists of the album's title, the name of the discographic company, the release year, its physic description, several entries for genres classified manually by the cataloguers, and finally notes about the content. Regarding the sound content, each vinyl was digitized in two files, one for each side. For 45 rpm discs each side usually contains only one song, while for 33 rpm LPs, which are more common in the collection, each side contains several tracks.

Having catalogued and digitized the material, some drawbacks emerge that strongly limit the browsing capabilities in the OPAC. The separation of tracks from a continuous stream could be easily solved in most cases just by finding silences between tracks. However, this may not be the case for live recordings or classical music tracks, where the music itself contains long rests. A related problem is the correct entitling of the tracks. Although some catalogued albums contain details of the contained tracks, there are many others, mainly operas, where the track names are not present. Another common situation is that of finding two different recordings of the same work whose tracks have been labeled using two different languages or naming schemes, e.g., "Symphony No. 9" knowns as "Novena Sinfonía" as well as "Choral Symphony". Audio fingerprinting techniques can hardly be applied to solve this task because of disc age, besides the fact that some of the discs may not have been reissued on CD and thus may not have been included in any audio fingerprint dataset.

Besides these drawbacks, the staff of the library demands some features that cannot be implemented given the current structure of the data. For example, given an album, find it in music sites like *Last.fm* or *Grooveshark*. Similarly, find a given song/track and its different recordings in those music sites and inside the library regardless of language or naming schemes. In order to locate music, they want the users to be able to query the library given metadata not contained in the catalog, like the lyrics of the songs.

3. CATEGORIZATION OF POP/ROCK MUSIC

The goal of the first task is to exploit both automatically extracted information about the content and user generated information about the context to carry out categorization. The task is based on a real application scenario: songs of a "commercial music library" need to be categorized according to their possible usage in TV and radio broadcasts or Web streaming (commercials, soundtracks, jingles). According to experts in the field, it is common practice to use different sources of information to assess the relevance of a given song to a particular usage. At first candidate songs are selected depending on the result of Web searches and on the analysis of user-generated tags. Since these sources of information are usually very noisy, experts make the final choice depending on the actual music content.

In order to simulate this scenario, participants of MusiCLEF are provided with three different sources of information: content descriptors, user tags, and related Web pages. Since CLEF campaigns aim at promoting multilingualism, tags and Web pages are in different languages. It was not mandatory, at least for MusiCLEF 2011, neither to use all the different languages nor to exploit all the source of information. In general, participants are free to select the descriptors that better fit the approach they want to test. To this end, the possibility of creating a baseline of individual sources of information is considered of interest for future MusiCLEF campaigns.

The dataset made available to participants includes mostly songs of pop and rock genres, which are the more often used in TV broadcasts. As mentioned in Section 2.1 a number of music professionals from LaCosa s.r.l. provided the categorization for the complete dataset of 1355 songs, which has been divided in a training set of 975 song and test set of the

[5] http://www.ua.es/en/bibliotecas/SIBID/fonoteca

remaining 380 songs. Being the first year, the ground truth is available for a limited number of songs but it is envisaged that the continuation of MusiCLEF over the years will create a shared background for evaluation.

The participants were asked to assign to each song in the test set the correct tags. Results were evaluated against the ground truth.

3.1 Definition of the Dataset

The task of music categorization can be considered an auto-tagging task, that is the automatic assignment of relevant descriptive semantic words to a set of songs. In the literature, several scalable approaches have been proposed for labeling music with semantics including social tagging, Web mining, tag propagation from similar songs, and content-based automatic strategies [3]. Regardless of the approach used, the output of a tagging system is generally a vector of tag scores, which measures the strength of the relationships tag-song for each tag of a semantic vocabulary (i.e. *semantic weights*).

The dataset built to carry out the auto-tagging evaluation campaign is composed of 1355 different songs, played by 218 different artists; each song has a duration between 2 and 6 minutes. One of the goals of the task is to have participants that may exploit, beyond content-based audio features, also other music descriptors (e.g. social and Web mined tags). For this reason we built the dataset using only well-known artists; this allowed us to gather a big amount of Web-based descriptors (i.e. the "wisdom of the crowd") for most of the songs in the dataset. We collected the songs starting from the "Rolling Stone 500 Greatest Songs of All Time" list [6], which was the cover story of a special issue of Rolling Stone (no. 963 of December 9 2004 – updated in May 2010). The song list was chosen based on votes by 172 musicians, critics, and music-industry professionals, and is almost entirely composed of English-speaking artists. Table 1 reports the top 10 positions of this rank list.

Starting from this list, we considered all the different artists as seeds to query a larger music database for gathering all the songs associated to every artist, excluding live versions that are usually of little interest for TV broadcasts. From this pool we randomly retained at most 8 songs per-artist, in order to fairly uniformly distribute songs between the different artist. As result, we had 161 artists associated with about 8 songs in the final collection.

Each song in the dataset has been manually annotated by music professionals from LaCosa. The vocabulary of tags defined by the experts was initially composed of 355 tags divided in two categories – genre (167) and usage (288) – loosely inspired by the Music Genome Project [7].

After that, all the songs have been tagged by the human experts with at least one tag for genre and five tags for mood. At the end, we discarded all the tags that were assigned to less than twenty songs; this led to the final released vocabulary of 94 tags.

Rank	Title	Artist
1	Like a rolling stone	Bob Dylan
2	(I can't get no) Satisfaction	Rolling Stones
3	Imagine	John Lennon
4	What's going on	Marvin Gaye
5	Respect	Aretha Franklin
6	Good Vibrations	Beach Boys
7	Johnny B. Goode	Chuck Berry
8	Hey Jude	Beatles
9	Smells like teen spirit	Nirvana
10	What'd I say	Ray Charles

Table 1*:* Top 10 songs of the Rolling Stone 500 Greatest Songs List (updated 2010).

3.2 Content- and Context-based Descriptors

Songs are also described by audio features. In particular, we precomputed timbre descriptors (Mel-Frequency Cepstral Coefficients) that are directly available to participants. Feature sets have been computed using the MIRToolbox [7] algorithms, which are publicly available. Moreover, participants can request the extraction of additional descriptors. In order to let participants perform their own feature extraction, we plan to make available also more general features in future years. In particular, we plan to provide the output of the triangular filterbanks before computing the log and the cosine transform of MFCCs. The rhythm based descriptors provided by the MIRToolbox will be precomputed as well.

We also provide social tags gathered from Last.fm as available on May 2011. For each song of the corpus, we used the Last.fm audio fingerprint service [8] and public data sharing AudioScrobbler website [9] to associate our music files to their songs and collect social tags for each song. Therefore, we release the list of social tags together with their associated score.

Category	Tags
Genre	bossanova, country rock, hymn, orchestral pop, slide blues
Mood	alarm, awards, danger, glamour, military, scary, trance

Table 2*:* A sample of the tags proposed to the music professionals for annotating the songs of the auto-tagging dataset.

[6] http://www.metrolyrics.com/rs/ (as in May 2011)
[7] http://www.pandora.com
[8] http://blog.last.fm/2010/07/09/fingerprint-api-and-app-updated/
[9] http://ws.audioscrobbler.com/2.0/

3.3 Web-mining

Web pages covering music-related topics have been used successfully as data source for various MIR tasks, in particular, for information extraction (e.g., band membership [5], artist recommendation [1], and similarity measurement [6, 8]. The text-based features extracted from such Web pages are often referred to as cultural or community metadata since they typically capture the knowledge or opinions of a large number of people or institutions. They therefore represent a kind of contextual data.

We first queried `Google` to retrieve up to 100 URLs for each artist in the collection. Subsequently, we fetch the Web content available at these URLs. Since usually the resulting pages typically contain a lot of unrelated documents, we alleviate this issue by adding further keywords to the search query, with an approach similar to [8]. We crawled various sets of Web pages in six different languages – English, German, Swedish, French, Italian, and Spanish – employing the following query scheme:

```
"artist name" (+music|+musik|+musique|+musica)
```

For MusiCLEF a total of 127,133 pages have been fetched.

The resulting information enables participants who would like to make use of structural information to derive corresponding features from the raw Web pages. In addition to these sets of Web pages, we provide precomputed term weight vectors. Taking into account the findings of a large scale study on modeling term weight vectors from artist-related Web pages [6], we first describe each artist as a virtual document, which is the concatenation of the HTML documents retrieved for the artist. We then compute per virtual artist document the *term frequencies* (tf) in absolute numbers. Further providing the *inverse document frequency* (idf) scores for the Web page set of each language will allow participants to easily build a simple $tf \cdot idf$ representation or apply more elaborate information fusion techniques. In summary, for the term vector representation of the dataset, we offer the following pieces of information:

- tf weights per virtual document of each artist
- global idf scores for each language
- corresponding lists of terms for each language

The twofold representation of the datasets (Web pages and generic term weights) leaves much room for various directions of experimentation. For example, Web structure mining and structural analysis techniques can be applied to the Web pages, while the provided term weight representation will certainly benefit from term selection, length normalization, and experimentation with different formulations for tf and idf.

4. IDENTIFICATION OF CLASSICAL MUSIC

The task of automatically identifying an audio recording is a typical MIR task, consisting of the clustering in the same group recordings of different performances of a composition. Also in this case, a real-life application scenario has been considered: loosely labeled digital acquisition of old analogue recordings of classical music should be automatically annotated with metadata (composer, title, movement, excerpt). Although systems for automatic music identification already give good results, the combination of segmentation and identification of continuous recordings is not well investigated yet. The participants are provided by a set of digital acquisitions of vinyls made by the Fonoteca, that has to be segmented and labeled.

An important aspect addressed by this task is the scalability of the approaches. To this end, we encourage participants to test the performance on the same task with a reference collection of increasing size, up to about 6,700 MP3s. This is achieved by providing additional information on the recording that can help filtering out part of the dataset. In particular, the additional information is consistent with the one founded in the real LP covers – author, performer, short title – and is the sole information that is reported by the Fonoteca catalogue. For this task, relevance judgments are provided automatically using available metadata and listening directly to the recordings.

Participants are provided with content descriptors of the complete dataset of 6680 single music files and with 22 additional digital acquisitions of 11 LPs (thus a total of 22 LP sides is available on individual MP3s). There are two different goal: to identify the songs belonging to the same group (for single files) and to match the content of the LP recordings with the corresponding songs.

4.1 Definition of the Dataset

Music identification usually focuses on pop music (hence its common designation as *cover song* identification). The reason for that might be attributed to the disproportion in commercial interests for the pop music market with respect to other genres. Nonetheless the need for the application of such technology to other styles is often felt by many music libraries and archives that, especially in Europe, aim at the preservation and dissemination of classical music.

The collection that we propose was created starting from the database of a broadcasting company consisting of about 320,000 music recordings in MP3 format (see Section 2.1). Our primary aim was to extract from it the largest possible sub-collection of classical music in order to build a shared dataset for the classical music identification task. We selected 2,671 such recordings, associated to works that are represented at least twice in the database. These recordings form 945 *cover sets* [10] ; the distribution of the set cardinalities follows a power law, and is represented in Figure 2. The distribution of the recordings with respect to the works' authors is depicted in Figure 3. The collection was finally

[10] The phrase "cover set" denotes a set of different recordings of the same underlying piece of music.

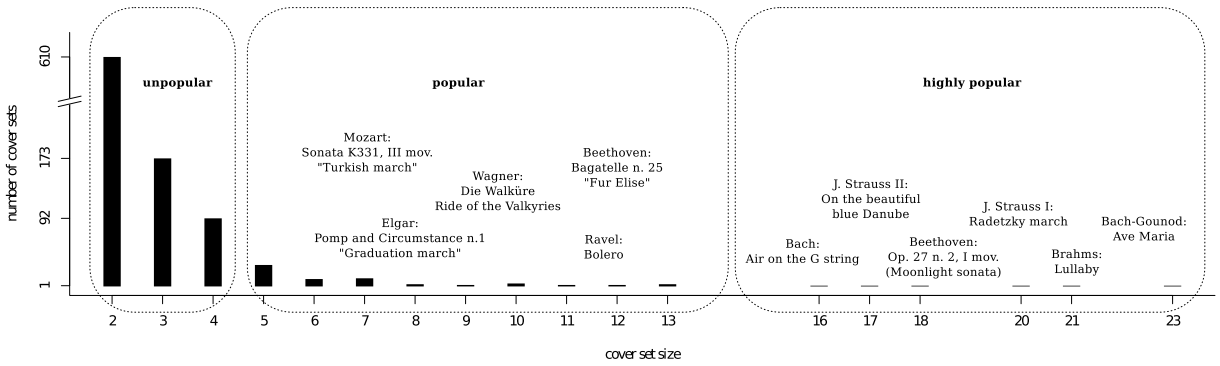

Figure 2*:* Distribution of cover set cardinalities for the classical music cover identification task.

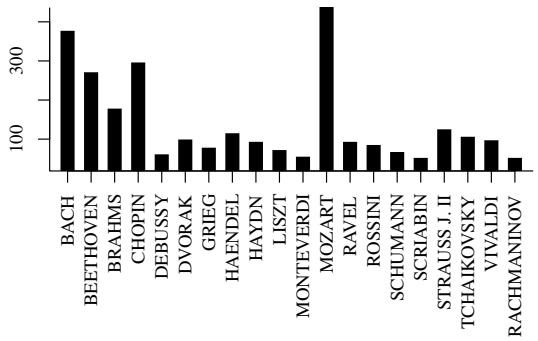

Figure 3*:* Number of files for the most represented authors.

augmented to 6680 pieces by adding recordings of classical music works by other authors.

4.2 Content-based Descriptors

Songs are described by audio features. In particular, we precomputed audio descriptors (chroma vectors) that are directly available to participants. Chroma vectors have been computed at different temporal and frequency resolutions. Also in this case, feature sets have been computed using the MIRToolbox [7] algorithms, which are publicly available. Moreover, participants can request the extraction of additional descriptors (which may include also additional chroma vectors computed with different algorithms). It is important to note that datasets of any size can be processed thanks to implicit memory management mechanisms developed in MIRtoolbox.

5. CONCLUSIONS

This paper introduces MusiCLEF, a new benchmarking activity that aims at fostering content- and context-based analysis techniques to improve music information retrieval tasks, with a special focus on multimodal approaches. A one-day MusiCLEF workshop is to be held in 2011 in Amsterdam as part of the Cross-Language Evaluation Forum (CLEF) conference, where participants can share their approaches and contribute to the future organization of MusiCLEF.

6. ACKNOWLEDGMENTS

The authors are grateful for the support of the staff of La-Cosa s.r.l. and the University of Alicante's Fonoteca. MusiCLEF has been partially supported by Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191. CLEF is an activity of PROMISE. This research is also supported by the Spanish Ministry projects DRIMS (TIN2009-14247-C02-02) and Consolider Ingenio MIPRCV (CSD2007-00018), both partially supported by EU ERDF, and by the Austrian Science Funds (FWF): P22856-N23.

7. REFERENCES

[1] S. Baumann and O. Hummel. Using Cultural Metadata for Artist Recommendation. In *Proc. of WEDELMUSIC*, Leeds, UK, Sep 2003.

[2] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. of ISMIR*, 2011.

[3] D. Turnbull et al. Five Approaches to Collecting Tags for Music. In *Proc. of ISMIR*, 2008.

[4] J. S. Downie et al. The 2005 Music Information retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. In *Proc. of ISMIR*, 2005.

[5] M. Schedl et al. Web-based Detection of Music Band Members and Line-Up. In *Proc. of ISMIR*, Vienna, Austria, Sep 2007.

[6] M. Schedl et al. Exploring the Music Similarity Space on the Web. *ACM Transactions on Information Systems*, 2011.

[7] O. Lartillot and P. Toiviainen. A Matlab Toolbox for Musical Feature Extraction from Audio. In *Proc. of DAFx*, 2007.

[8] B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proc. of ICMC*, Göteborg, Sweden, Sep 2002.