A Professionally Annotated and Enriched Multimodal Data Set on Popular Music

Markus Schedl
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

Cynthia C. S. Liem
Delft University of Technology
The Netherlands
c.c.s.liem@tudelft.nl

Nicola Orio
University of Padua
Padua, Italy
orio@dei.unipd.it

Geoffroy Peeters
UMR STMS IRCAM-CNRS
Paris, France
geoffroy.peeters@ircam.fr

ABSTRACT

This paper presents the MusiClef data set, a multimodal data set of professionally annotated music. It includes *editorial meta-data* about songs, albums, and artists, as well as MusicBrainz identifiers to facilitate linking to other data sets. In addition, several *audio features* (generic low-level descriptors and state-of-the-art music features) are provided. Different sets of *annotations* as well as *music context* data – collaboratively generated user tags, web pages about artists and albums, and the annotation labels provided by music experts – are included too. Versions of this data set were used in the MusiCLEF 2011 and in the MusiClef 2012 evaluation campaigns for auto-tagging tasks.

In this paper, we report on the motivation for the data set, on its composition, on related sets, and on the evaluation campaigns in which versions of the set were already used. These campaigns likewise represent one use case, i.e. music auto-tagging, of the data set. The complete data set is publicly available for download at http://www.cp.jku.at/musiclef.

Categories and Subject Descriptors

Information systems [**Information retrieval**]: Evaluation of retrieval results—*Test collections*; Information systems [**Information retrieval**]: Specialized information retrieval—*Music retrieval*; Applied computing [**Arts and humanities**]: [Sound and music computing]

General Terms

Measurement, Documentation, Standardization

Keywords

MusiClef, Multimodal Music Data Sets

1. INTRODUCTION

MMSys 2013 Oslo, Norway

With the digital information age, a continuously expanding collection of media items has become available to an ever growing online audience. This collection also encompasses many music items. As the amount of items has grown too large to be humanly overseeable, intelligent automated techniques are necessary to describe and organize them, and to allow human users to navigate among them and retrieve the items they are looking for. For the music domain, this boosted advances in an area known as Music Information Retrieval or Music Information Research (Music-IR).

The relatively abstract nature of music makes both the development of automated music analysis and description techniques, as well as their evaluation, challenging and non-trivial tasks. In addressing these, it has increasingly been emphasized that both development and evaluation should take place with users and real-life scenarios in mind, and that the experience of music is not just established by an audio signal, but follows from a combination of this signal with multimodal contextual information [3, 6, 14].

With these considerations in mind, the MusiClef initiative was set up, focusing on benchmarked Music-IR evaluation for real-life use scenarios. As part of this initiative, a professionally annotated multimodal music data set has been developed, which is introduced in this paper.

The remainder of the paper is organized as follows: A general overview of benchmarking initiatives in Music-IR, and corresponding data sets, is given in Section 2. The motivation for and history of the MusiClef data set, together with a description of the real-world auto-tagging use case it was originally intended for, is presented in Section 3. In Section 4, the multimodal music data set is described in detail. Not part of the data set, but strongly related to its use for auto-tagging is a reference implementation we provide and elaborate on in Section 5. Eventually, we summarize the work and present possible extensions of the data set in Section 6.

2. BENCHMARKING IN MUSIC-IR

Benchmarking in Music-IR faces several challenges. First of all, due to copyright restrictions, the music data on which evaluation takes place can typically not be shared in its original form. Following this, it is not trivial to establish transparency in the interpretation of obtained results.

At present, the most strongly profiled benchmarking forum for Music-IR tasks is the Music Information Retrieval Evaluation eXchange (MIREX), which is held annually as part of the International Society for Music Information Retrieval (ISMIR) Conference. MIREX covers various

Music-IR tasks proposed by the community, ranging from "Audio Music Similarity and Retrieval" to "Audio Tag Affinity Estimation" and from "Classical Composer Identification" to "Genre Classification". Different data sets are used, depending on the task. However, because of the copyright restrictions mentioned above, the evaluation data typically cannot be shared with MIREX participants. Hence, they must locally experiment on their own collections, after which they submit their algorithms to be run on the evaluation set by the organizers. Unfortunately, this leads to results that cannot easily be replicated. Furthermore, the performance of the submitted algorithms strongly depends on the quality of the individual training sets used by the participating groups.

A recent and promising initiative to overcome these limitations is the `Million Song Dataset` (MSD), which allows researchers to access a number of features from a very large song collection [2]. Features include audio descriptors, lyrics, listening histories, and semantic tags. In 2012, a music recommendation benchmarking initiative[1] was organized based on this data set. However, the feature set is static and the used feature extraction algorithms are not fully public, limiting possibilities to carry out further research on content description techniques.

In 2011, `Yahoo! Labs` organized the `KDD Cup`[2] on music recommendation. This initiative was controversially discussed in the Music-IR community. On the one hand, the offered real-world data sets cover an outstanding number of over 300 million ratings for more than 600,000 music items; on the other hand, the data sets are very abstract in nature – no information about data items and users, other than anonymous identifiers, are given. Since the task was a pure rating prediction task (common in the field of recommendation systems, but not in Music-IR), it was not perceived a very interesting one by the majority of Music-IR researchers. The same additional shortcomings as for the MSD Challenge hold as well.

3. THE MUSICLEF INITIATIVE AND ITS AUTO-TAGGING CAMPAIGN

3.1 Motivation

As a response to the Music-IR benchmarking issues described in the previous section, in 2011, a lab named `MusiCLEF` was run in the `Cross-Language Evaluation Forum`[3] (CLEF) [9]. In this lab, major effort was spent on acquiring and annotating data for Music-IR benchmarking corpora, meeting the following requirements:

- The data should consist of *multimodal resources*, in order to reflect both content-related and contextual information;
- The data should be *relevant to a real-life use case* and *annotated by professionals*;
- The corpora should be *as transparent and flexible as possible*. While copyright restrictions still hold for the raw music data, it should be encouraged that the computed feature descriptors to be shared are computed using implementations which are clearly documented in literature, and preferably are publicly available. Furthermore, during benchmarking runs, the corpora should allow for specialized feature computation on demand, in which benchmarking participants would be able to submit their own feature extractors.

[1] `http://www.kaggle.com/c/msdchallenge`
[2] `http://kddcup.yahoo.com`
[3] `http://www.clef2011.org`

`MusiCLEF 2011` yielded two corpora and benchmarking task setups: one dedicated to automated tagging of popular music (auto-tagging), and one dedicated to identification of classical music vinyl recordings. In 2012, the auto-tagging task was further refined, and run as a "`MusiClef` Multimodal Music Tagging" Brave New Task in the `MediaEval`[4] multimedia evaluation campaign [8, 7], which formed the basis for the data set described in this paper. The auto-tagging task will be described in the following subsection.

3.2 MusiClef 2012 Auto-Tagging Campaign

The use case for the auto-tagging corpus initiated at `MusiCLEF 2011` was provided by a professional broadcasting service provider. Frequently, suitable music is sought to accompany broadcasting productions. In order to make optimal and efficient use of the music tracks available in a library, it is then important that these tracks have meaningful annotations (possibly originating from external resources), which preferably are obtained in an automated way, and on which automatic searches can be performed. Concrete questions posed were (1) how to improve the information acquisition process, extracting the maximum amount of information about music recordings from external resources and (2) how to provide good suggestions of possible usages of music material, minimizing the amount of manual work. Considering ongoing work in the Music-IR research field, these problems were connected to the challenge of *music auto-tagging*, based on rich multimodal resources.

Music auto-tagging typically involves training a supervised learner on a training data set that associates feature representations with semantic tags. In order to increase computational efficiency, optionally some feature selection or dimensionality reduction technique might be employed to the input feature vectors before training the classifier. After training is finished, the classifier is used to predict labels to previously unseen music items. A schematic illustration of the process can be found in Figure 1.

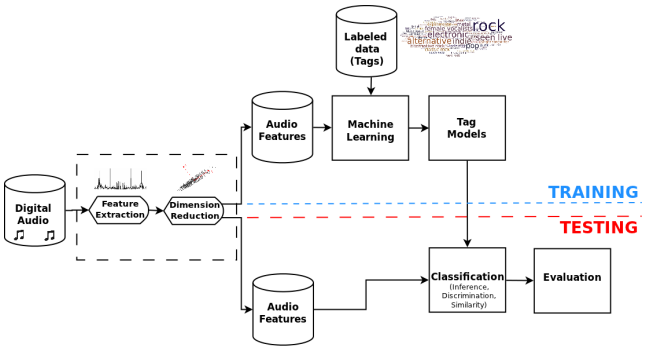

Figure 1: Illustration of a music auto-tagger [18].

For the `MusiClef` auto-tagging use case, a data set with popular music was established, for which the commissioning broadcasting service partner provided song-level annotations by its professionals regarding genre and mood[5]. The benchmark organizers computed audio features for the songs, and enriched the data by gathering supporting collaborative annotations and contextual information from the web.

In its practical setup, the `MusiClef 2012` auto-tagging task was intended to follow best practices from the evaluation cam-

[4] `http://www.multimediaeval.org`
[5] For this use case, the concept of 'mood' is related to the usage of a particular song within a video production.

paigns in the Quaero program[6], meant to promote transparency. In particular, the implementation of the evaluation framework was made public, and together with this, a reference implementation was provided. This implementation was not intended to take part in evaluation result rankings, but to exemplify a naive, nevertheless comprehensive approach to the benchmarking task, which could serve as a baseline. Finally, with the release of the MusiClef data connected to this paper, the ground truth corresponding to the test data of the auto-tagging campaign will be made public, such that complete annotation data is available. Further data specifications are given in the following section.

4. MUSICLEF DATA SET

In this section, we will give the specifications of the data set connected to this paper. This data set, which can be identified as MIR:MusiClef:2012:MMSys:version1.0 (see [10] for the naming rationale), is a revised and expanded version of the data set that was used for the MusiClef 2012 auto-tagging benchmarking campaign, which would be identified as MIR:MusiClef:2012:MediaEval:version1.0. The revisions and expansions have been performed such that ground truth annotations are available for both original train and test split items, and that the data has been additionally enriched with multimodal resources in such a way that it will also be useful to Music-IR use scenarios other than auto-tagging. For the sake of simplicity, we will refer to our current data set as the MusiClef data set.

Since one of the motivations to generate this data set was multimodality in representing music items, the MusiClef data collection consists of five parts: *editorial metadata*, *audio features*, *collaboratively generated user tags*, *web pages*, and the *annotation labels provided by music experts*. In Figure 4, we summarize the parts of the proposed corpus. In Table 1, we provide a formal description of the corpus using the methodology proposed in [10]. In the remainder of this section, the corpus components are explained and specified in further detail.

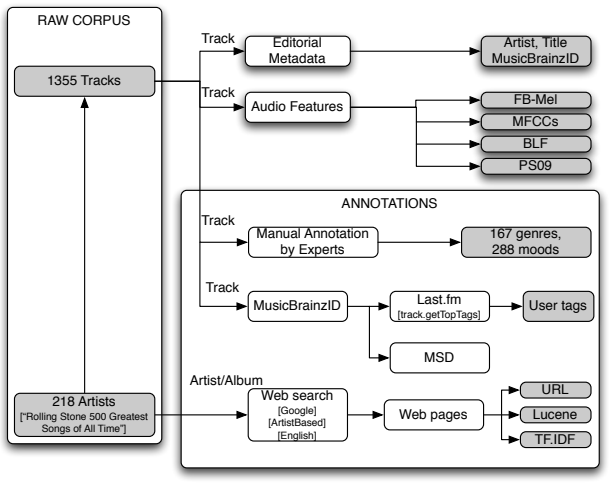

Figure 2: Various content types and annotations attached to the music tracks or artists.

[6]Quaero is a program promoting research and industrial innovation on technologies for automatic analysis and classification of multimedia and multilingual documents, gathering around 30 French and German public and private research organizations.

Table 1: Description of the annotated corpus according to [10].

(C1) Corpus ID: MIR:MusiClef:2012:MMSys:version1.0
(A) Raw Corpus
(A1) Definition: The Corpus is made of (a13) sampled real items. The sampling is a "Popularity-oriented sampling".
(A2) Type of media diffusion: full duration music items in stereo high-quality. It should be noted that the audio of the data set is not distributed but is represented by audio features. These audio features are not considered as annotation in this table.
(B) Annotations
The data set is multimodal, i.e. the raw corpus is distributed with different type of annotations.
(B1) Origin: The annotations distributed with this corpus have been obtained by (b15) Manual annotation (expert annotations into Genre and Mood) and (b12) Aggregation (Last.fm user tags and web pages). Last.fm tags are obtained by (b14) crowdsourcing.
(B21) Concepts definition: The definition of the concepts (meaning of the tags of Last.fm, of the words in the text documents, of the genres and moods) are not given. They are defined by their application to the specific music items.
(B22) Annotation rules: A vocabulary of annotations, divided into genre and mood, has been agreed on with the supervisors of a commercial library of production music. Each song was annotated with at least one tag for genre and five tags for mood.
(B31) Annotators: The annotation has been made by a group of music professionals who routinely provide textual descriptors for commercial music libraries as part of their job.
(B32) Validation / reliability: Since each song was annotated by only one expert, no cross-checking has been carried out. The time required to annotate a single song has been logged.
(B4) Annotation tools: Annotators accessed a web interface that randomly assigned them a number of songs to annotate. They could listed to the whole song through the interface and annotate it using a set of checkboxes.
(C) Documents and Storing
(C2) Audio identifier: Artist names and track titles are provided. MusicBrainz identifiers are provided for the artists and the tracks.
(C2) Storage: The data set is accessible at http://www.cp.jku.at/musiclef. Each type of annotation is stored in a specific file. The files are available in CSV and XML formats.

4.1 Raw Corpus

Selection of the corpus content

In order to avoid sparsity of particular data sources that are prone to low data coverage, one of the requirements for the collection was to select well-known songs by popular artists. This way, we can expect that enough social tags are available for each song and enough web pages are available for each artist. We collected the songs starting from the "Rolling Stone 500 Greatest Songs of All Time", which lists songs that have been recorded by a total of 218 different artists.

The initial list of 500 songs was extended by adding at most 8 songs for each artist, obtaining a final list of 1,355 songs. We purposely excluded live versions and cover songs, because recordings of the former frequently show low audio quality and the latter can give inconsistencies between tags related to the performer and web pages related to the composer.

Editorial metadata

Due to the integration of various data sources in the MusiClef data set, it is crucial for recognition and linking purposes to provide editorial metadata. We hence include lists of artist, track, and album names. Furthermore, in order to facilitate cross-linking with other data sets or information resources, we provide MusicBrainz[7]

[7]http://www.musicbrainz.org

identifiers. These are not only provided at the artist level, but — upon availability — at the track level as well.

Audio features

For copyright reasons, audio can not be distributed. Content is hence made available through the distribution of pre-computed audio features.

First, a set of low-level features has been computed using the publicly available `MIRtoolbox` [4]. For each track, we provide the following two sets of low-level audio features, each one computed on a frame-basis, i.e. values represent temporally stationary content descriptors.

FB-Mel are obtained by decomposing the signal through a bank of 40 triangular filters, distributed on a Mel scale. This descriptor represents a general feature to be used as input for further processing; in particular, it is used for computing the following descriptor.

MFCC Mel Frequency Cepstral Coefficients are the most widely used audio features in speech processing [12]. They are obtained by computing a Discrete Cosine Transform (DCT) over the logarithm of FB-Mel descriptors. This allows to decorrelate the various dimensions of the feature vectors. The first 20 coefficients are included in the data set.

In addition to these low-level features, and different from the basic features provided in the MSD, we further offer features computed by two state-of-the-art audio feature extraction algorithms [17] and [11].

BLF Block-Level Features are a combination of several features that model temporal aspects of the audio by dividing the signal into blocks [17].

PS09 The MIREX submission made by Pohle and Schnitzer in 2009 is an aggregation of features that describe rhythmic aspects (*Fluctuation Patterns*, *Onset Patterns*, and *Onset Coefficients*) and features that model timbre (*MFCCs*, *Spectral Contrast Coefficients*, and two highly specialized descriptors *Harmonicness* and *Attackness*) [11].

We provide feature vectors and similarity estimates between all tracks, where the algorithms produce those. Both algorithms performed very well in various MIREX tasks (in particular, "Audio Music Similarity" and several auto-tagging tasks) during the past few years.

4.2 Annotations / Music Context

Manual annotations by experts (track level)

All songs in the data set have been manually annotated by a group of professional music consultants. The expertise of these annotators include providing textual descriptors to commercial music libraries, in particular for production music where tags are used as the primary means to select music soundtracks for video broadcasts. The vocabulary of tags was defined in tight collaboration with music consultants, and it was initially composed of 355 tags: 167 for genre and 188 for mood. Manual tagging was carried out through a web interface, from which it was possible to listen to the complete songs and select the associated tags through a number of checkboxes, divided in genre and mood. Annotators were required to provide at least one tag for genre and five tags for mood. Songs were randomly assigned to annotators and were presented in random order. Each song has been tagged by one annotator.

Table 2: A subset of the tags used by the professional annotators.

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

From the initial set, we kept only the tags that have been assigned to at least 10 songs, obtaining a final list of 94 tags. A subset of the tags is reported in Table 2, from which it can be seen that both mood and genre tags cover a variety of categories.

User tags (track level)

We used the API provided by `Last.fm`[8] to gather the collaborative user tags associated to each song. More precisely, we used the function *track.getTopTags*. The distribution of social tags across songs is not uniform, ranging from only one tag – for about 3% of the songs – to more than 200 tags – for about 20 songs. For each tag, it is also reported the weight assigned by the `Last.fm` API, which is an integer number between 0 and 100.

Web pages (artist and release level)

Web pages covering music-related topics have been used successfully as data source for various Music-IR tasks, in particular for information extraction (e.g., band membership [16], artist recommendation [1], and similarity measurement [19, 15]. 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 aspects of the "music context".

We first query `Google` to retrieve up to 100 URLs for each artist in the data set. Subsequently, we fetch the web content available at these URLs. Since the resulting pages typically contain a lot of unrelated documents, we add additional keywords to the search query, employing an approach similar to [19]. We crawled various sets of web pages in six different languages, to further foster linguistic multimodality – English, German, Swedish, French, Italian, and Spanish. We used the following query scheme:

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

In addition, we performed another crawl, including album names, using the query scheme:

```
"artist name" "album name" +music
```

We restricted this crawl to English in order to avoid reducing coverage, presuming that much fewer web pages are available on the album level. Table 3 gives some statistics on the corresponding data set.

We provide the actual web pages, corresponding URLs, `Lucene`[9] indices, and $tf \cdot idf$ feature vectors over the entire terms in the corpus. Including the raw web pages enables extracting structural information and derive additional features. In addition to the sets of web pages, we provide pre-computed term weight vectors. Taking into account the findings of a recent large scale study on modeling term weight vectors from artist-related web pages [15], we first describe each artist as a virtual document, by simply concatenating

[8] `http://www.last.fm`
[9] `http://lucene.apache.org`

Table 3: A summary of the web-page-data set.

Lang.	Query Scheme	Pages	Terms
English	"artist" +music	20,907	1,828,291
German	"artist" +musik	21,343	1,759,041
French	"artist" +musique	20,907	1,787,870
Italian	"artist" +musica	21,342	1,447,465
Spanish	"artist" +musica	21,467	1,345,758
Swedish	"artist" +musik	21,487	1,668,628
English	"artist" "album" +music	52,626	3,257,695

the HTML documents retrieved for the artist. We then compute per virtual artist document the *term frequencies* (tf) in absolute numbers, and the *inverse document frequencies* (idf), again interpreting as document each virtual artist document.

5. MUSICLEF 2012 REFERENCE CODE

While not intended as a part of the formal corpus, in order to illustrate corpus usage for a concrete use case, we also provide the reference implementation that was released as part of the `MusiClef 2012` auto-tagging benchmarking campaign. To this end, Gaussian models were trained on the audio MFCC representations. Jointly with the `Last.fm` user tags and $tf \cdot idf$ web page features, classification was applied through a 1-nearest neighbor approach, using as proximity measures symmetrized Kullback-Leibler divergence (for audio) and cosine similarity (for text). The evaluation code, focused on several common Information Retrieval measures (accuracy, recall, precision, specificity, and F-measure), is released together with this reference implementation.

Finally, noting the large diversity in the professional tag vocabulary, we conjectured that different types of tags (e.g., contrast 'hopeful' to 'travel') may need different types of approaches in terms of modalities. Therefore, we made a functional categorization of the tags (also released as part of the reference implementation, see [8] for further details) to allow a deeper result analysis, including categories related to affect, genre, sound quality, but also specific occasions or places for which the song would be appropriate.

Employing our reference implementation, we ran evaluations for several fusion strategies, obtaining results for five cases: (1) consideration of audio features only, (2) consideration of user tag features only, (3) consideration of web page features only, (4) majority vote, considering all three data resources, and only keeping tags indicated by at least two of the resources, and (5) taking the union of the tags obtained for each of the three data resources. F-measure results for these different fusion strategies and obtained on our functional categorization are illustrated in Figure 3, indeed implying that the usage of resources in different modalities is beneficial.

6. SUMMARY AND OUTLOOK

We presented the `MusiClef` data set, a professionally annotated, multimodal collection of popular music. The data set consists of five parts: editorial metadata about the music items (artists, albums, songs, `MusicBrainz` identifiers), audio features, web pages, collaborative tags, and professional annotations.

Although the data set is in a stable state as of September 2012, the highly dynamic nature of some data sources included, in particular web pages and collaborative tags, demands for updating these parts of the collection. For instance, to include news about an artist that might influence human perception of his or her music it is necessary to re-crawl the web page set from time to time.

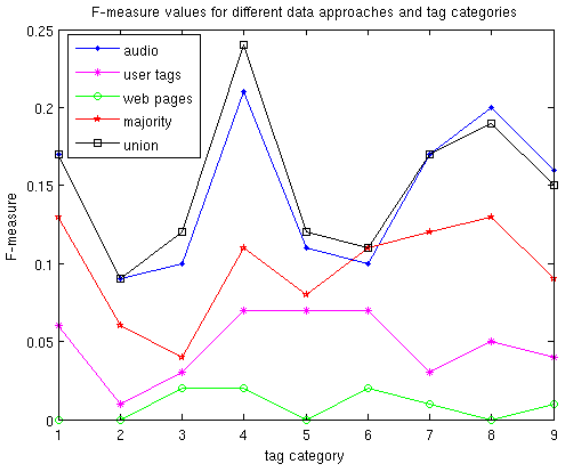

Figure 3: F-measure values for different data fusion strategies (indicated in the legend), considered per tag category. The correspondence between horizontal axis indices and categories is: (1) activity/energy, (2) affective state, (3) atmosphere, (4) other, (5) situation: occasion, (6) situation: physical, (7) sociocultural: genre, (8) sound: temporal, (9) sound: timbral.

We further plan to incorporate microblog data gathered from `Twitter`[10]. It has been shown that microblogs are a valuable data source for music similarity and retrieval tasks [13]. However, as `Twitter` is very restrictive in making publicly available their user data by means other than their API, we are not allowed to share the actual microblogs. We are hence evaluating way to share higher level representations, for instance, instead of the microblogs themselves, we may be able to share artist, album, and song names that are mentioned in tweets.

We are also considering to include representations of album covers. As it has been shown that information derived from images of album cover artwork can be used for music tagging purposes [5], this would further strengthen the multimodality of the data set. But again, we have to carefully take into account possible copyright restrictions, and will thus only be able to include features such as color histograms.

In summary, we believe that the `MusiClef` data set represents a truly multimodal set that should be established as a standard data set for multimodal music retrieval tasks. The set is publicly available for download from `http://www.cp.jku.at/musiclef`.

7. ACKNOWLEDGMENTS

The authors would like to thank David Rizo, Riccardo Miotto, Nicola Montecchio, and Olivier Lartillot for their support in starting the `MusiCLEF` initiative. `MusiClef` has been partially supported by the PROMISE Network of Excellence, co-funded by EU-FP7 (no. 258191), by the Quaero Program funded by Oseo French agency, by the MIReS project funded by EU-FP7-ICT-2011.1.5-287711, and by the Austrian Science Funds (FWF): P22856-N23. The work of Cynthia Liem is supported in part by the Google European Doctoral Fellowship in Multimedia.

[10] `http://www.twitter.com`

8. REFERENCES

[1] S. Baumann and O. Hummel. Using Cultural Metadata for Artist Recommendation. In *Proceedings of the 3rd International Conference on Web Delivering of Music (WEDELMUSIC)*, Leeds, UK, 2003.

[2] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[3] J. S. Downie, D. Byrd, and T. Crawford. Ten Years of ISMIR: Reflections on Challenges and Opportunities. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, October 2009.

[4] O. Lartillot and P. Toiviainen. A Matlab Toolbox for Musical Feature Extraction from Audio. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2007.

[5] J. Lībeks and D. Turnbull. You Can Judge an Artist by an Album Cover: Using Images for Music Annotation. *IEEE Multimedia*, 2011.

[6] C. C. S. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic. The Need for Music Information Retrieval with User-centered and Multimodal Strategies. In *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, pages 1–6, Scottsdale, AZ, USA, 2011.

[7] C. C. S. Liem, N. Orio, G. Peeters, and M. Schedl. Brave New Task: MusiClef Multimodal Music Tagging. In *Working Notes Proceedings of the MediaEval 2012 Workshop*, Pisa, Italy, 2012.

[8] N. Orio, C. C. S. Liem, G. Peeters, and M. Schedl. MusiClef: Multimodal Music Tagging Task. In *Proceedings of the 3rd Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*, Rome, Italy, 2012.

[9] N. Orio, D. Rizo, R. Miotto, N. Montecchio, M. Schedl, and O. Lartillot. MusiCLEF: A Benchmark Activity in Multimodal Music Information Retrieval. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, USA, 2011.

[10] G. Peeters and K. Fort. Towards A (Better) Definition Of The Description Of Annotated M.I.R. Corpora. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, October 2012.

[11] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On Rhythm and General Music Similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, October 2009.

[12] L. Rabiner and B. Juang. *Fundamentals of speech recognition*. Prentice-Hall, New-York, 1993.

[13] M. Schedl. #nowplaying Madonna: A Large-Scale Evaluation on Estimating Similarities Between Music Artists and Between Movies from Microblogs. *Information Retrieval*, 15:183–217, June 2012.

[14] M. Schedl and A. Flexer. Putting the User in the Center of Music Information Retrieval. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, October 2012.

[15] M. Schedl, T. Pohle, P. Knees, and G. Widmer. Exploring the Music Similarity Space on the Web. *ACM Transactions on Information Systems*, 29(3), July 2011.

[16] M. Schedl, G. Widmer, T. Pohle, and K. Seyerlehner. Web-based Detection of Music Band Members and Line-Up. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[17] K. Seyerlehner, G. Widmer, and T. Pohle. Fusing Block-Level Features for Music Similarity Estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, September 2010.

[18] M. Sordo. *Semantic Annotation of Music Collections: A Computational Approach*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2012.

[19] B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proceedings of the International Computer Music Conference (ICMC)*, Göteborg, Sweden, 2002.