A WEB-BASED APPROACH TO DETERMINE THE ORIGIN OF AN ARTIST

Sten Govaerts **Erik Duval**
K.U. Leuven
Department of Computer Science
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{sten.govaerts, erik.duval}@cs.kuleuven.be

ABSTRACT

One can define the origin of an artist as the geographical location where he started his career. The origin is an important metadata element, because it can help to specify subgenres, be an indicator of regional popularity and improve recommendations. In this paper, we present six methods to determine the origin, based on Web data sources: one extracts data from Last.fm, two query Freebase and three analyze biographies. We evaluate the different methods with 11275 artists. Circa 55% of the artists can be classified using biographies. The best Freebase method can classify 26% and the Last.fm based method 7%. When comparing on accuracy, the Last.fm and Freebase methods perform similarly with around 90% accuracy. For the biography-based methods we achieve 71%. To improve coverage, a final, hybrid method achieves 77% accuracy and 60% coverage. The accuracy of the continent classification is 87%. As a showcase for our classifier, we developed a mashup application that displays, among others, information about the origin of artists from radio station playlists on a map.

1. INTRODUCTION

In a previous project [1], we developed a music player for hotels, restaurants and pubs. Peculiar to our approach is that a user can describe the music he wants by referring to a situation, rather than by defining the usual search criteria on artist, title, etc. This system uses almost 40 metadata fields, manually annotated by music experts of Aristo Music (http://www.aristomusic.com), which is a very time-consuming and expensive labor. Currently, the Aristo Music database contains around 58000 songs. The time-consuming metadata annotation process is difficult to scale: in the case of Aristo Music for instance, it limits the ability to penetrate new markets. In order to assist the experts, we already achieved some success in automating the annotation process for some metadata fields [2].

This paper reports on ongoing work in the MuziK project that focuses on automatically generating the metadata fields that are most costly to do manually and most relevant for end users. For this purpose, we rely on a variety of approaches: digital signal processing [3], web-based classification [4] and data analysis [5].

This paper focuses on one of the parameters: the origin of an artist, defined here as the geographical location where an artist started his musical career. This can be hard to determine sometimes. It can be seen as the country of the artist's first success, where he lived most of his life or where some of the group members live. For example Georg Friedrich Händel was born in Germany, but went to England where his career really took off.

1.1 Relevance

Although the origin of an artist can thus be quite fuzzy, it is a useful piece of metadata in many cases.

- Some subgenres are based on geographical location of the artist, for example Britpop and Viking Metal.
- It can also be a good indicator of the popularity of an artist in a region, as most artists are often most popular in their country of origin. An artist popularity visualization based on Last.fm data shows this by comparing two countries, http://hublog.hubmed.org/archives/001085.html. There are of course exceptions with an international carreer, like the Spice Girls.
- Recommendations can also improve by using the artist's origin. They can be tuned to the location of the listener. Some musically very similar songs can be good or bad recommendations, depending on the region (and dialect): for example, a small stage art genre in Belgium and the Netherlands (called "Kleinkunst") is mostly expressed in regional dialects and if the recommendation for a song from Antwerp is a song originating from Amsterdam it could break the atmosphere, although they are musically very similar.

The data in Figure 1, collected during an internal time management evaluation, shows the average time in seconds needed by a music expert to annotate different metadata fields. Origin is the 6th most expensive element, and rather close to the top. Providing the origin often requires manual lookup work, hence making it very expensive to annotate. Annotating continent takes much less because it is derived from origin. Target region is used for localized music distribution and is done in batch, making it faster. One can determine the origin in a more or less precise way. Sometimes, the country is not specific enough due to differences in regional music styles or linguistic,

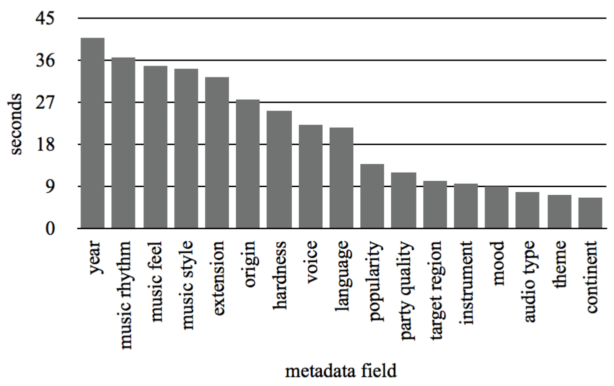

Figure 1. The average number of seconds to manually annotate a metadata field.

linguistic, cultural or religious differences within a country, as in for instance East coast vs. West coast rap. We rely on two levels: country and continent.

1.2 Related work

Today, specialized search engines enable searching for persons, which often requires linking different sources to a person. Researchers are using knowledge management systems to find experts [6]. Numerous examples can be found online: NNDB (http://www.nndb.com/), Pipl (http://pipl.com/), and Spock (http://www.spock.com/). They often offer information on birthplace and residence and some even recognize music artists. Except for Celma's demo [7], we did not find any related work specialized towards classifying the origin of music artists.

Origin is annotated by the experts based on either their personal knowledge of the artist or by looking up the origin information. Therefore, we try to automate this metadata field by extracting information from web data sources. In this paper, we present our approach to classify an artist's geographical origin. First, the web data sources and algorithms will be explained and evaluated in section 3. Afterwards, we discuss the concept and design of our mashup tool, which locates artists of radio station playlists as a demonstration of the technique, and present the conclusion and possible further work.

2. FINDING THE ORIGIN

Determining the origin of the artist will be very hard by using content-based techniques (e.g. by analyzing the signal). The best option is to analyze other data sources. In our quest to classify the origin we looked at a wide plethora of music related web resources. Most of these resources are in plain text, for example sites with reviews like Amazon.com, which makes it hard to extract geographical locations, but results can be achieved with for example named entity recognition [8]. One of the main problems is the use of the names of the inhabitants or adjectives of geographical locations, e.g. German vs. Germany. These kinds of words are called demonyms and are not recognized as geographical descriptors by most named entity recognizers, e.g. Open Calais (http://www.opencalais.com/). Luckily, more structured data sources are available: for instance, most Wikipedia artist pages contain a box with background information. Moreover, there are some online databases available for querying, e.g. Freebase (http://www.freebase.com). Tags are also a rich source of information and often contain geographical data. If an artist is really more popular in his country of origin (not proven), listening counts of an artist per country might also be an interesting source.

Our approach relies on several methods with different data sources, because no single data source covers all artists. In the remainder of this section, we will present our methods to determine the origin of an artist. Section 2.1 covers a screen scraping technique based on Last.fm (http://last.fm). Section 2.2 presents two approaches that rely on Freebase, and section 2.3 details a method to analyze an artist biography with demonyms.

2.1 Origin determination with Last.fm

Last.fm is a well-known music recommender system and a music community website with over 30 million users, making it a great resource of metadata for MIR, such as biographies and tags [9]. For some artists, Last.fm contains the origin and sometimes their different whereabouts over time, for example Radiohead is located in Abingdon, Oxfordshire, UK since 1986, according to http://www.last.fm/music/Radiohead.

We scrape the Last.fm artist page with Dapper (http://www.dapper.net/open/), which basically creates a web service out of unstructured data. Sometimes the data on Last.fm is incomplete: e.g. New York, without a country. To fill in these gaps, we use the Google Maps API (http://maps.google.com) for geo-coding to retrieve the ISO 3166-1 country code, used to identify the origin.

2.2 Origin determination with Freebase

Freebase [10] is a large collaborative semantic database, containing structured data, harvested from different resources, e.g. Wikipedia.org and MusicBrainz.org. Freebase allows querying through a REST-ful web service, using ontologies to describe the semantics and data interlinking. Different classes (/music/group_membership, /music/artist, /music/musical_group/member) describe music artists and within these, others (place_of_birth, nationality, origin, places_lived) describe geo-locations. We run a long complex query covering all the artist and geo-location classes of Freebase, which is used in 2 methods:

- **Based on the freebase origin class (freebase-origin):** the origin class is geo-coded with the Google Maps API to get the country code, which is the result.
- **Most frequent location (freebase-most_freq):** takes the nationality, birthplace and places where the artist lived and geocodes them all. The most occurring country code in all the locations is selected. Then out of all the locations with this selected country code, the most occurring city is selected. The location of that

city and country code is the result of the method. To conclude, this method returns the most frequent location of nationalities, birthplaces and residences of the group members.

2.3 Origin determination with biographies

Biographies often provide a lot of geographical information and thus can be a great source of information on the origin. As mentioned before, authors often describe geolocations with demonyms, for example "Anders Trentemøller is a Danish electronic musician…" (from http://www.last.fm/music/Trentemøller). We looked for natural language stemmers to transform demonyms, but none were found. One might be able to extend a stemmer with rules to cover all exceptions. We use a list of countries and their demonyms from Wikipedia, which also contains Anglo-Saxon cities and the states of the USA (http://en.wikipedia.org/wiki/Nationality). We noticed that the origin or residence of an artist is often mentioned in the first sentences of the biography. We implemented 3 variations that exploit this characteristic. The biography is split into natural language sentences and for every sentence the occurring demonyms and locations are noted.

- **Highest occurrence (bio-most_freq):** The result is the demonym or geographical location that occurs most often.
- **Favor first occurrences (bio-favour_1st):** For every encountered country code (cc) a list of sentence numbers s in which cc occurred is kept. Say, s_{tot} is the total number of sentences in the biography, then following formula is calculated for every country code cc:
$$R_{cc} = \sum_{i=0}^{length(s)} \frac{(s_{tot}+1)-s_i}{s_{tot}+1}$$
The result is the country code with the highest R_{cc}.
- **Weaker favoring first occurrences (bio-weak_favour_1st):** this method is equal to the previous, but another weighting function is applied for every country code cc:
$$R_{cc} = \sum_{i=0}^{length(s)} \frac{(s_{tot}+2)-s_i}{s_{tot}+1}$$
Again, the result is R_{cc}. This method tries to spread importance a bit more over the sentences.

3. EVALUATION

The approaches from section 2 are evaluated against a ground truth data set provided by Aristo Music. First, we describe the data set normalization, then the results are discussed and a combination of all methods is presented.

3.1 The data set

As ground truth, we use the origin and continent field from the Aristo Music database. They group metadata per song, although the origin is an artist property. For some artists, different songs indicate different origins, due to errors in the database. In those cases, we consider the origin that occurs most often. From the data set, we removed 10 artists that are not annotated with origin, as well as artists with an origin value like "Mixed" and "Others": these are used when for instance a group of artists collaborate, e.g. "George Michael & Aretha Franklin". These artists are removed from the ground truth – we identify them through connectors like "and", "feat." and "vs.". The Caribbean contains the Caribbean Sea and its islands and cannot be mapped to a single country code. Artists annotated with "Caribbean" are thus also removed from the ground truth based. The overall result is that 25% are removed: we keep 11275 artists (Table 1).

To classify continents, the origin is mapped to a continent lookup table. There are different ways to define continents, based on geography or political treaties. We use a Wikipedia table (http://en.wikipedia.org/wiki/List_of_-countries_by_continent_(data_file)). Table 1 also shows the distribution of artists over continents. Our data set reflects mainstream music taste in Europe, with a strong representation of North America and Europe.

As input for the methods, the artist page and biography is retrieved for all artists with the Last.fm API. All origins are geocoded to obtain the country code.

3.2 The results

Section 3.2.1 examines how well all approaches cover the data set. Section 3.2.2 discusses the accuracy of the results per method. Finally, we present a new method that increases the coverage and improves overall performance.

3.2.1 Coverage

The coverage is defined as the number of artists of the ground truth that have been determined. Figure 2 shows the percentage of artists for which an origin was found per method. For 59%, an origin can be found by at least one of the methods. The large difference in coverage is the main reason why we use multiple data sources.

Only a small percentage of Last.fm artist pages (7%) contain the origin. Freebase covers a bit more than 26%. Surprisingly, only 5% less artists have an explicit origin class in the Freebase ontology. For the three biography-based methods, the coverage is obviously the same (56%) and about double that of Freebase. As only 63% of artists

	#	%
total nb. artists original data	14880	100
total nb. artists cleaned data	11275	75,77
nb. artists removed	3605	24,23
nb. artists in Europe	7167	63,57
nb. artists in N. America	3612	32,04
nb. artists in S. America	195	1,73
nb. artists in Asia	141	1,25
nb. artists in Africa	80	0,71
nb. artists in Oceania	80	0,71

Table 1. The number of artists in the data set and for each continent in the cleaned data set.

have a biography, this means that almost all of the retrieved biographies contain geographical information. For 62% of the artists, an origin could be retrieved with at least one of the methods. The coverage of the combination method will be discussed later.

As discussed, the best method can only annotate 56% of the artists. One of the reasons is that Aristo Music is a Belgian company with a database that includes many local artists for whom neither biographies nor freebase entries are available. There is also a rich tradition of mardi gras music in Belgium, made by artists well known in a very small region (sometimes only a village or town) for a short period of the year. Their online presence is zero. There are also possible differences in artist name writing between Aristo Music and Last.fm. The latter often offers alternative pages for the different writings; most of these pages contain no biography or a very condensed one, disabling the biography-based methods.

3.2.2 Accuracy

The combination method makes use of the results of the other methods. Therefore the same data set cannot be used to evaluate all methods. A data set, called combo data set, contains 3000 randomly selected artists from the full data set to evaluate the combination method. The 8275 artists left are used to evaluate the Last.fm, Freebase and biography methods and is called the LFB data set. The coverage of the LFB and combo data set is almost equal to Figure 2 (up to about 1% difference).

The accuracy is defined as the number of correct classifications divided by the total number of artists covered by a method. The accuracy of the Last.fm, Freebase and biography-based methods for the continent and origin classification on the LFB data set is shown in Table 2. As expected, the continent classification is performing better, because small errors in the origin will not impact on the continent classification: for instance, a misclassification of a French artist as a German one will still result in a correct continent (Europe). The Last.fm screen scraping approach has the highest accuracy overall and the Freebase methods perform equally well and are close second. Their 95% confidence intervals overlap, so the best performer cannot be concluded, nor from performed t-tests.

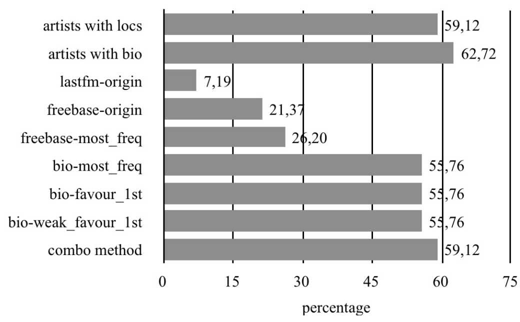

Figure 2. The percentage of annotated artists by the different methods.

methods	origin (%)	CI origin	cont. (%)	CI cont.
lastfm-origin	91,28	[89,06; 93,50]	95,32	[93,66; 96,98]
freebase-origin	90,60	[89,24; 91,96]	93,58	[92,44; 94,72]
freebase-most_freq	90,60	[89,37; 91,83]	94,36	[93,39; 95,33]
bio-most_freq	63,50	[62,11; 64,89]	77,21	[76,00; 78,42]
bio-favour_1st	70,95	[69,64; 72,26]	83,12	[82,04; 84,20]
bio-weak_favour_1st	66,16	[64,80; 67,52]	78,74	[77,56; 79,92]

Table 2. The accuracy and 95% confidence intervals for all methods for the classification of continent and origin on the LFB data set.

There is quite a drop in accuracy when using the biography as a source. Continent classification with biographies performs much better than origin. When comparing the results of the 3 different biography-based methods, it is clear that our assumption that the origin appears in the first sentences is valid, because the method favoring the first sentences most strongly, bio-favour_1st, is the best.

All methods have their issues. We noticed that some of them occur due to errors in geocoding of locations with the same name, e.g. Birmingham in USA and UK, or a mix-up between small geographical entities and their big neighboring countries, e.g. Jersey and England, Luxemburg and Belgium. Another problem occurs when different artists have the same name; in that case, they have to be identified on song level. Of course, there are also method dependent errors: for example, sometimes the Last.fm origin is given as a demonym instead of a country, e.g. “American” and “French”, and the geocoding will resolve this to another country, respectively to Americana, Brazil and Wattsburg, USA. The freebase-origin method often takes the country of birth as the origin of the artist, e.g. Akon was born in Senegal, but moved as a child to the USA. In the case of freebase-most_freq, it occurs that one band member dominates the locations, because Freebase contains many more geographical data on that one member as the rest. This is the case with Ry Cooder of Buena Vista Social Club: he originates in the USA, while this is a Cuban band. The Last.fm biography can contain multiple biographies from different artists with the same name. This of course confuses the biography-based methods.

In Figure 3, we can see the accuracy for every method for every continent. It is clear that the continent classification works better for Europe, North America and Oceania. This can be due to the strong representation of these continents in the ground truth or a stronger representation on the web (see Table 1). To be able to see a consistent trend, we need more data for the smaller continents. We noticed that the misclassifications of the biography-based methods for Asia can often be traced back to Japan, because of Japanese album releases. Some of the Belgian and Dutch artists are classified with former colonies. This happens when the biography contains the place of birth or performance locations. Bio-favour_1st outperforms the two other biography methods in Africa, Asia and South-

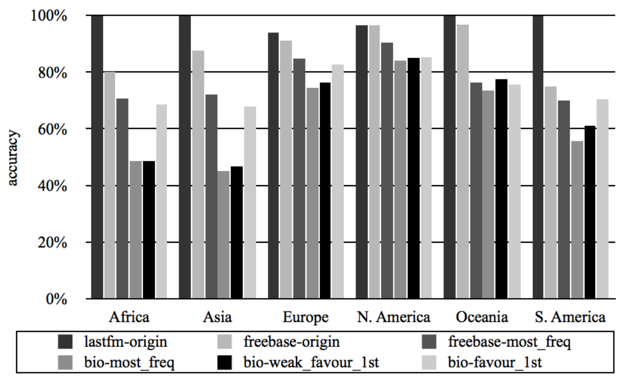

Figure 3. The accuracy of every method for every continent on the LFB data set.

America. Due to the high number of countries it is impossible to this on country level.

3.2.3 Improve by combination (combo_method)

We will now introduce a new method that maximizes coverage and improves performance beyond bio-favour_1st. The idea of the new method is to use all origins found by all methods to improve the coverage and select the origins smartly to improve accuracy.

The result is selected in an order based on the accuracy in Table 2. Since the 95% confidence intervals of the Last.fm and Freebase methods overlap, we don't know which one is significantly better. We used the highest accuracy and the spread of the confidence intervals to order. If a location of lastfm-origin is available, this is the result, because it has the highest accuracy. Otherwise freebase-most_freq is selected, because the confidence interval of freebase_most_freq lies encapsulated in that of freebase-origin, then freebase-origin and as last option bio-favour_1st. We can still improve this slightly by using the continent accuracy in Figure 3. Freebase-origin performs better in Africa, Asia and South-America. If the origin determined by freebase-origin is from these continents, this is selected before freebase-most_freq.

Obviously, the coverage of the new method (59%) for the complete data set equals the percentage found by all methods (Figure 2). The accuracy of the combo-method for classifying origin on the combo data set is shown in

	recall	precision	F-measure
US	0,8824	0,8468	0,8642
GB	0,8092	0,8513	0,8297
FR	0,7576	0,8117	0,7837
DE	0,6281	0,8444	0,7204
NL	0,6186	0,7849	0,6919
BE	0,5385	0,9825	0,6957
IT	0,6538	0,6939	0,6733
SE	0,8718	0,8293	0,8500
JM	0,7586	0,9565	0,8462
CA	0,8947	0,6071	0,7234

Table 4. Precision, recall and F-measure of the top countries categorized by combo_method.

methods	origin (%)	CI origin	cont. (%)	CI cont.
lastfm-origin	89,58	[85,26; 93,90]	94,79	[91,65; 97,93]
freebase-origin	90,85	[88,61; 93,09]	94,16	[92,33; 95,99]
freebase-most_freq	91,60	[89,65; 93,55]	94,70	[93,12; 96,28]
bio-most_freq	64,63	[62,33; 66,93]	77,81	[75,81; 79,81]
bio-favour_1st	71,04	[68,85; 73,22]	82,04	[80,19; 83,89]
bio-weak_favour_1st	66,26	[63,98; 68,54]	78,66	[76,69; 80,63]
combo_method	77,09	[75,12; 79,06]	86,29	[84,67; 87,90]

Table 3. The accuracy and 95% confidence intervals for all methods on the combo data set.

Table 3. The method with the highest coverage previously, bio-favour_1st, improves from 71% to 77% and for continent to 86%. The lower accuracy, compared to the Freebase and Last.fm methods, is caused by half of the results, which are from bio-favour_1st. The combo-method performs 14% less than the best method, freebase-most_freq, but covers more than double.

Table 4 shows the recall, precision and F-measure of the 10 countries with the most artists classified by combo_method. Overall the precision is higher than the recall. A higher precision is preferred for this task, because the main task of the classifier is to classify new artists. Belgium, the Netherlands and Germany have a rather low recall value. Currently, these countries are the main markets of Aristo Music, so there are comparatively more globally lesser-known artists in the database for these countries than others and thus less data available on the web. The lower precision for Canada is probably caused by misclassifications with US and mistakes by the biography analysis: for example, if the bio mentions "French speaking", then the artist may be classified with France.

4. MASHUP

As a showcase for the classifier, a mashup application was developed, that visualizes playlists from radio stations and locates the artists on a map, based on the origin classifier. We also link the data to YouTube videos, biographies and the Last.fm account.

The playlists are retrieved from Last.fm accounts, Twitter or scraped from radio station websites with Dapper. Yahoo! Pipes retrieves the data from these sources and adds the biography, pictures, track duration and Last.fm URL from the Last.fm API. Then we apply the origin classifier, implemented as a REST web service on Google App Engine. It requires the artist, the biography and the Last.fm artist page URL; e.g. http://artistlocator.appspot.com/?artist=Morrissey&bio=Test+bio&last_fm_url=http://www.last.fm/music/Morrissey. Figure 4, shows a screenshot of the mashup application, http://www.cs.kuleuven.be/~sten/lastonamfm/.

During its 24 days online, the mash-up attracted 220 individual visitors; almost 30% of them return. Most visitors come from Belgium, USA, France and UK. On 8 May, it was "Mashup of the Day" on Programmable Web (http://www.programmableweb.com/). A Belgian na-

Figure 4. A screenshot of the mashup application.

tional radio station discussed it on air. Another radio station intends to use it to analyze playlists of competitors.

5. CONCLUSION AND FUTURE WORK

Our aim was not to build a very complex classifier, but to see how far we could get using simple techniques. Real life systems can often benefit from this, e.g. short computation time. This rather simple approach leads to decent performance. From 6 different methods to classify the origin, the 4 best performing are combined in the final classifier to maximize coverage to 59%. The resulting origins are selected from the classifier with the highest accuracy. This results in a final accuracy of 77% for origin classification and 86% for continent classification.

The classifier can probably benefit from applying data mining techniques. For example the biographies can be analyzed for words that co-occur for artists of the same country. Something similar could be trained on the Last.fm tags of artists with the same origin to extract geographical information from tags. Another idea is to use named entity recognition, like Open Calais, to extract birthplaces and other geographical facts. This could complement the demonym analysis to get more detailed information, for example city names.

One important way to improve the classifier is by increasing the coverage. The identification of an artist online has to be improved to get the correct biography. This can be done with online identification services, e.g. MusicBrainz. More data sources can also enable further improvement. People search engines sometimes show the birthplace and could thus be leveraged. Additional data sources, such as Belgian rock/pop catalogs, could be exploited to cover local artists. An obstacle might the intellectual properties of such collections. The web can be used as a whole to extract information from by for example crawling music related sites, using search engines for classification [4] or using links from semantic search engines, like http://sig.ma.

In any case, we believe that our classifier is accurate enough to automatically annotate the origin of an artist for applications like that of Aristo Music. This can remove the need for costly manual effort and enable scalable automation for an important metadata element.

6. ACKNOWLEDGEMENT

We gratefully acknowledge the support of IWT Vlaanderen on the project: "MuziK: Muziek op maat van de klant" under grant IWT 080226 and Aristo Music NV for the evaluation data, metadata annotation time management study and their musical expertise.

7. REFERENCES

[1] N. Corthaut, S. Govaerts, E. Duval. “Moody Tunes: The Rockanango Project”, Proc. of the 7th Int. Conf. on Music Inf. Retrieval, 2006, pp. 308-313.

[2] S. Govaerts, N. Corthaut, E. Duval. “Mood-ex-Machina: Towards Automation of Moody Tunes”, Proc. of the 8th Int. Conf. on Music Inf. Retrieval, 2007, pp. 347-350.

[3] G. Tzanetakis, P. Cook. “Musical genre classification of audio signals”, IEEE Trans. on Speech & Audio Proc., Jul. 2002, 10(5), 293-302.

[4] M. Schedl, T. Pohle, P. Knees, G. Widmer. “Assigning and visualizing music genres by web-based co-occurrence analyis”, Proc. of the 7th Int. Conf. on Music Inf. Retrieval, 2006.

[5] J. Bergstra, A. Lacoste, D. Eck. “Predicting genre labels for artists using FreeDB”, Proc. of the 7th Int. Conf. on Music Inf. Retrieval, 2006.

[6] I. Becerra-Fernandez. “The role of artificial intelligence technologies in the implementation of People-Finder knowledge management systems”, Knowl.-Based Syst., 2000, 13(5), pp. 315-320.

[7] Ò. Celma, M. Nunes, “GeoMuzik: A geographic interface for large music collections”, Proc. of the 9th Int. Conf. on Music Inf. Retrieval, 2008.

[8] B. Pouliquen, R. Steinberger, C. Ignat, T.De Groeve. "Geographical information recognition and visualization in texts written in various languages", SAC '04, 2004, pp. 1051-1058.

[9] Paul Lamere. “Social tagging and music information retrieval”, Journ. New Music Res., 37(2), 101–114.

[10] K. Bollacker, C. Evans, et al. “Freebase: a collaboratively created graph database for structuring human knowledge”, 2008, ACM SIGMOD, pp. 1247-1250.

[11] M. Meire, X. Ochoa, E.Duval, “SAmgI: Automatic metadata generation v2.0”, 2007, Proc. of EdMedia07, AACE, pp. 1195-1204.