Data Gathering for a Culture Specific Approach in MIR

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

ABSTRACT
In this paper we describe the data gathering work done within a large research project, CompMusic, which emphasizes a culture specific approach in the automatic description of several world music repertoires. Currently we are focusing on the Hindustani (North India), Carnatic (South India) and Turkish-makam (Turkey) music traditions. The selection and organization of the data to be processed for the characterization of each of these traditions is of the utmost importance.

Categories and Subject Descriptors
H.3.1 [**Information Systems**]: Information storage and retrieval – *content analysis and indexing.*

General Terms
Documentation

Keywords
Music Information Retrieval; Computational Musicology.

1. INTRODUCTION
CompMusic, Computational Models for the Discovery of the World's Music [1], is a research project whose goal is to advance in the field of MIR by approaching a number of the current research challenges from a culture specific perspective. It aims to advance in the automatic description of music through the development of information modeling techniques applicable to five non-Western music cultures: Hindustani (North India), Carnatic (South India), Turkish-makam (Turkey), Andalusian (North Africa), and Han (China).

The project involves three main tasks: (1) gathering of data related to the selected musical cultures, (2) investigating information processing methodologies for finding the specificities of each musical repertoire, and (3) developing culture specific systems for discovering the richness of each musical culture.

Our work is in line with the current research approaches in MIR that are based on combining audio content analysis with the extraction of cultural information from text sources. By cultural information we normally refer to any metadata obtained from the web that can complement the audio features. However, in CompMusic we give a more restricted and semantic meaning to it. For us it means the information that describes the musical significance of a piece within its own cultural tradition. Our aim is to use this information to build systems that can help understand and appreciate music and its cultural significance.

Properly gathering the data with which to study and characterize a musical piece in its cultural context is a research issue in itself. We have to make sure that the data is musically and culturally meaningful and that is organized in the most suitable way to be processed with our computational methods designed for a content+context music description approach.

Next I list the different types of data sources to be used in our project and discuss some issues behind the data gathering process. More information on the data sources is available on the project's website: http://compmusic.upf.edu.

2. EXPERT INFORMATION
By expert information we refer to the texts created by experts that can be processed to discover new semantic relationships of relevance for the understanding of the musical repertories selected. Of course this automatic process cannot substitute the qualitative study of musicological and cultural texts and our personal interaction with actual experts.

Academic publications, if available in digital text form, can be used for quantitative studies. We use a collaborative on-line reference manager, Mendeley[1], to gather, organize and share all the relevant academic publications. The editorial metadata of the publications is publicly available and, due to copyright issues, most of the actual articles are accessible only to the project researchers. A system like Mendeley is a powerful tool to access and process academic references automatically, thus it can be considered a structured text repository from which we can automatically extract musical knowledge.

Another source of expert information is Wikipedia[2]; from which we can extract musically relevant concepts to characterize our repertoires, even more so if we use a tool like DBpedia[3], which already extracts structured information from Wikipedia. However, quite a number of the entries related to the music cultures studied are incomplete and we are committed to help improve them. It is especially important that experts immersed in the different musical cultures contribute with their own cultural perspective.

3. AUDIO RECORDINGS
Given that one of the aims of the project is to extract musically meaningful descriptors from audio recordings, we are gathering a representative audio collection for each music culture. Experts advised us and we bought around 200 commercial CDs for each repertoire plus we got access to some personal CD collections, gathering more than 300 hours of audio recordings in each collection. The size will grow in time and we aim at reaching 500 hours of audio per collection in the next few years.

For the selection of the CDs it was important to choose recordings by recognized and representative artists, with reliable editorial data. The issue of editorial data has been a problem especially in the case of Hindustani music.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1] http://www.mendeley.com. http://www.mendeley.com/groups/944081/compmusic-indian. http://www.mendeley.com/groups/940871/compmusic-makam-maqam.

[2] http://www.wikipedia.org

[3] http://dbpedia.org

For specific experiments we are recording music fragments in controlled situations. To gather and organize these recordings we use Freesound[4], an open collaborative site of audio samples that supports different ways to describe, organize and access the samples and their metadata. We are encouraging contributions of audio recordings that might be relevant for the project.

4. AUDIO FEATURES

A lot of our research work requires starting from standard low-level features extracted from the audio recordings. Thus, it makes sense to pre-compute all the common low-level features for all the collections and store them together with the audio files. To generate these features we use Essentia[5], an audio analysis library developed by our research group that includes most of the common low and mid level feature analysis algorithms. We store the audio feature data in YAML[6] format.

5. MUSIC NOTATION

The musical cultures studied have developed as oral traditions but they use music notation and that is of relevance to study and describe some musical characteristics. However few music notations are available in digital form and we will have to input them in a format like MusicXML[7] if we want to process them.

The Turkish-makam tradition adapted western notation in the early 20th century. Many scores are available in printed from and some digital repositories exist. Türk Müzik Kültürünün Hafızası[8] includes more than 53000 scores in image format and the multimedia encyclopedia Mus2okur[9] has around 1500 scores.

In Hindustani music there is very little tradition of written music and the few compilations of music notations that have been published are not digitally available. On the other hand in Carnatic music the written compositions are more common and there are quite a number of printed compilations of music notations. Again, the difficulty is in the availability of digital formats.

6. LYRICS

A big percentage of the pieces that we are working on are songs with lyrics and a number of MIR problems can be tackled by using the lyrics. For the case of Carnatic music Sahityam.net[10] includes the lyrics of more that 1800 songs. For Hindustani music Swarganga[11] has a large database of bandishes and for Turkish-makam songs Mus2okur is a good reference.

7. EDITORIAL METADATA

To gather the editorial metadata of the CD collections we use MusicBrainz[12], an open repository of music metadata. It supports all the metadata associated to CDs plus other detailed information about the music. However MusicBrainz was designed to support western popular music and it lacks the support for some of our culture specific concepts. Given the community-based approach of MusicBrainz we are involved in the addition of the culture specific concepts that we need and in making sure that the system can properly support our music repertoires.

8. COMMUNITY INFORMATION

Apart from using expert information, it is also possible to extract domain specific knowledge from non-expert information, thus reinforcing or discovering new semantic relations. Computationally this can be done by mining the text and analyzing the discussions in on-line communities of music enthusiasts.

For the case of Carnatic music we have started analyzing Rasikas.org[13], which is a very active community of Carnatic music lovers. We have not yet been able to find comparable on-line communities from the other music cultures, but we are using Freesound as a test bed to develop general methodologies to characterize on-line communities and to understand the effect of the platform in the development of community interactions. This is very useful for us because we have developed and maintain the platform for this community of sound lovers.

9. GROUND TRUTHS

A number of research problems and methodologies that we use require the availability of manually labeled-data, ground truths. These are descriptions, or metadata, that need to accompany the audio recordings and that are generally used for validating the automatically generated data of an algorithm. We are creating ground truth data for a numbers of research tasks, like tonic analysis in Indian music or motive analysis in Turkish-makam music. We store this data as XML files accompanying the recordings and it will be available through the project website.

10. CONCLUSIONS

Proper data gathering is fundamental in any MIR project but even more so if we want to take a cultural specific perspective. We have to make sure that the data is comprehensive and meaningful in its cultural context.

Once we have the data, we need to facilitate its accessibility, especially in a data infrastructure as distributed and heterogeneous as in our case. For that we have developed a linking system using the APIs of the different data repositories to access them from a unified software interface, in an effort to contribute to the Linking open data initiative[14]. It is from this interface that we can process our entire distributed data sources and efficiently run our analysis algorithms.

11. ACKNOWLEDGMENTS

The CompMusic project has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement 267583.

12. REFERENCES

[1] Serra, X. 2011. A multicultural approach in music information research. *Proceedings of the International Society for Music Information Retrieval Conference*.

[4] http://www.freesound.org

[5] http://mtg.upf.edu/technologies/essentia

[6] http://en.wikipedia.org/wiki/YAML

[7] http://www.makemusic.com/musicxml

[8] http://www.sanatmuziginotalari.com

[9] http://www.musiki.org

[10] http://sahityam.net

[11] http://www.swarganga.org/bandishbase.php

[12] http://musicbrainz.org/user/compmusic/collections

[13] http://www.rasikas.org

[14] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData