

# Exploiting web content for enhancing semantic descriptions of music

Òscar Celma

July 21, 2004

## 1 Introduction

My research is mainly focused on the Music Information Retrieval (MIR) field. Nowadays, an immeasurable amount of digital multimedia material is available (on the World Wide Web, in broadcast data stream, in digital storage media) and this amount constantly grows. The intrinsic value of this information depends on how easily we can manage, search, retrieve and access to it.

As the number of available media increases, so does the need for a user to locate the desired audio files in an efficient way. Searching in digital libraries has been widely studied for several years, mostly focusing on retrieving textual information using text-based methods. These kind of queries can be complemented and improved with advanced retrieval methods focused on content-based descriptors (extracted from the audio by applying signal processing and machine learning techniques), even though some musical knowledge management and representation is necessary.

## 2 MIR and Database Management Systems

Database management systems (DBMS) have been widely used to implement efficient text based information retrieval systems, solving many of the problems encountered in that field. However, in the area of multimedia, and in particular in MIR, there is still a lot of ongoing research and there are open questions concerning architectural and design aspects related to DBMS.

### 2.1 Music Ontologies and Semantic Web

Managing audiovisual files implies to structure its associated metadata and content-based descriptors —using description schemes, taxonomies and ontologies— to organize a powerful musical knowledge representation.

Moreover, Semantic Web approach adds a powerful value to collected musical data (and metadata), such as reasoning and classifying given instances within a given ontology.

Thus, DBMS offers ways to manage ontologies (and its instances) in an efficient way. There are several database paradigms (Relational Databases, Object-Oriented, native XML DB, etc.) that can model this approach in a novelty manner.

## 2.2 Content gathering

As mentioned in the previous section, another important issue is the process of gathering content —data and metadata— that is already available on the web. Data mining techniques allows to extract relevant information, thus, allowing to integrate valuable information within a MIR system. For this reason, studying and developing a web-crawler agent focused on musical information is a need. The information extracted by this agent could be converted into usable knowledge and to improve and refine a MIR system.

## 3 Objectives

The main goal is to study the relationships between information available in the Web with information that can be automatically extracted from music files. Structures for representing both types of information are needed. Database architectures and functionalities that are capable to synergetically combine both areas are still to be studied and developed. In order to attain the main objective, the following threads —which follows a natural data flow— have to be addressed:

- **Gathering music content through web mining:**
  - To study the infrastructure of a Web Crawler
  - Adapt Web Crawler's infrastructure making it fitted to the requirements of music description
  - To study different DBMS paradigms for hypertext information storage and retrieval
  
- **Knowledge representation ontologies:**
  - To study different methodologies and languages for building ontologies
  - Investigate methodologies for (semi-automatic) creation of ontologies related to the Music field
  - To study different DBMS paradigms to manage ontologies

- **Learning and Reasoning capabilities:**

- To study different logic approaches related to knowledge representation ontologies (first order logic, descriptive logics, etc.)
- To study reasoning mechanisms and adapting some of them to the specificities of music description

The planned output of the research is a system capable of:

1. Improving and adding rich metadata annotations to users' music collections
2. Bridging the gap between content-based music and metadata information, allowing a researcher to infer new semantic descriptors based on both approaches
3. Generating new metadata information automatically from a music collection, combining both strategies