Contribution to the modelling of multimedia metadata in a distributed architecture

Ana-Maria Manzat, Florence Sedes, Romulus Grigoras

Institute de Recherche en Informatique de Toulouse
Ana-Maria.Manzat@irit.fr
Florence.Sedes@irit.fr
Romulus.Grigoras@enseeiht.fr

Nowadays almost any human activities create electronic documents. These documents are more or less complex. Not only that they are generated, but their content is needed sooner or later and the retrieval of relevant documents has become an important problem in the present. The retrieval of multimedia documents is a problem because of the important size of the content's collection and also because of the distributed aspect of the collection and the heterogeneity of the multimedia content.

A fair amount of research has been conducted in the field of information retrieval. There are approaches that are focused on the management of multimedia documents in a distributed architecture (e.g., CAIM project[1], CANDELA project[2]). In an Information Retrieval system, the most important phase is the indexation process. This process generates the metadata (data about data, data that describes the document, its content and also its context). These metadata are used in the retrieval process of documents that respond to a user query.

Now there are many people who show interest in the metadata field, researchers and industrials as well. They work together for establishing standards for meta-data and for creating new ones and improving the existing ones. In the present there are many meta-data standards defined for audio files, video files, text and image like Dublin Core, EXIF, STMPE, MPEG-7,etc. These standards are used for specifying certain information about the multimedia document. Some of the existing formats contain only a limited number of elements (e.g., Dublin Core) and others that are too exhaustive (e.g., Mpeg-7). Thus, the information systems must handle many standards and they become very complex.

The thesis is accomplished under the ITEA2 project LINDO[3] (Large scale distributed INDexation of multimedia Objects). This project is focused on managing the multimedia indexation process inside a distributed environment. An important issue of LINDO is the integration of different indexation engines in the system and their deployment in real time, while the system is running.

[1] http://caim.uib.no/
[2] http://www.hitech-projects.com/euprojects/candela
[3] http://www.lindo-itea.eu

In this context, the core problem of my thesis concerns the development of a generic modelling of the multimedia metadata in a distributed architecture.

Our main idea is that the format to be proposed should contain the existing metadata standards with limited changes in these standards. The integration of the standards must be done with the elimination of redundancy. In the different existent formats there are elements with the same meaning but different names. Our format will take into consideration the synonymy between the elements to be integrated. Some mapping rules must be proposed in order to resolve the synonymy and the equivalence between the standards.

The format we will propose should have an extensible hierarchical structure (e.g., new elements can be added anywhere in the structure). The levels of the structure are not a priori fixed. Such a structure can integrate easily the notion of granularity of the metadata (i.e., we can have metadata related to the hole document, to its content, to a part of the document, and so on).

A generic metadata format has several advantages. It solves the interoperability problem in the Information Retrieval Systems, where the indexing engines have as output different formats of metadata. If these outputs can be integrated in the generic format the system can use this format in the retrieval process and in the management of the multimedia documents. Another advantage is that once the mapping is done between the standards, it can be used to integrate these standards in the generic format and also the same mapping rules can be employed to obtain certain standard elements from the generic format.

Our generic format will be used in the LINDO project in order to integrate any indexation engine in the system without changing the system itself, the output of the new indexation engine will be transformed in the generic format that the system can handle.