Modeling and Flexible exploitation of Audio Documents

Mohamed Mbarki, Chantal Soulé-Dupuy and Nathalie Vallès-Parlangeau
SIG-D2S2 IRIT Paul Sabatier University
Toulouse I University
Mbarki@irit.fr, soule@univ-tlse1.fr and nathalie.valles@univ-tlse1.fr

Abstract

The overwhelming growth of multimedia information as well as its dematerialization and accessibility through the World Wide Web, present new research challenges in modeling, storage, retrieval and document analysis. The users need an easily handle of this huge multimedia information. Having in mind this need, the development of content based multimedia indexing and retrieval appears to be in the spotlight of multimedia manipulation research.

This paper presents a system designed for the management of multimedia documents warehouse. It aims to face the problems of efficient media processing and representation throw logical and semantic classification, modeling and exploitation.

1. Introduction

Current electronic documents fall into the category of complex objects (also called semi-structured data) owing to the definition of standards, such as XML (eXtensible Markup Language), SMIL (Synchronized, Multimedia Integration Language), and MPEG-4, MPEG-7. So, any document, that it is or not multimedia, can be generally considered from the viewpoint of its contents (which we often qualify as semantics), its logical structure and its presentation. It is then necessary to define a model of document to identify the common characteristics of a document category. These viewpoints on the multimedia document have several interests. First of all they must allow the representation of heterogeneous information. Then, they must facilitate the handling of this information according to their structure (organization and presentation) and their contents (semantics). Lastly, they must allow the realization of multiple analyses of existing multimedia documents owing to multimedia indexing.

The goal of document warehousing is to constitute a shareable repository in which information can be seen as a whole or like a piece of global information according to the need of any user. Starting from this integrated information, the warehouses must allow their processing according to several viewpoints, and via several techniques (document retrieval, multidimensional analysis, etc).

This paper presents our approach as regards modeling and exploitation of document warehouse. The metamodel is detailed, as well as its features. We describe how to instantiate this metamodel by audio documents. We also show techniques that we use to exploit the warehouse content.

The remainder of the paper is organized as follows. First, we outline some works devoted to the storage and the manipulation of documents. Then, we propose our document warehouse metamodel. The next section describes the information extraction we propose to instantiate the warehouse. Finally, we give an example of the experiments carried out within the framework of the warehouse exploitation.

2. Related Works

The document warehouse organizes and structures multimedia information for content retrieval. In this context, a multimedia document modeling is one of the key issues. The term is used to determine which information should be stored in a warehouse and to reflect the relationships between the document parts. In order to be able to handle the various types of data including text, images, videos and audio, several models were proposed. These models can be classified in two categories according to their levels of completeness in the holding of the multimedia documents description.

The first category gathers works which aims at modeling each type of media separately. These approaches do not manage the fitting of several media in only one document:

- Loisant E. and al in [8] propose a metamodel that can be used to describe any type of media. The goal of this metamodel is to provide an independent media base to generate specific models to only one type of media.
- Moënne-Loccoz N. and al in [11] provide a model to manage the specificities of video documents. This model ensures in particular the recognition of the temporal aspect and the diversity of the video document descriptors (high and low level).

The models of the second category cover the totality of the media that compose the documents. They transcribe links that connect the various mono-media components of the same document:

- Amous I. and al in [1] extend classic approaches by adding a set of metadata specific to each type of media in order to formalize information relating to the document content.
- Darmont J. and al in [3] propose an approach that presents the multimedia documents within a unified format by using XML language. This facilitates their structuring in document databases. Indeed, they propose a conceptual model that generalizes and presents any type of document in the form of a complex object. They use some characteristics of these documents to index them.

These works are intended to describe architectures that integrate the structures and the description data (metadata) within the same model. Such model allows the management of heterogeneous sets of semi-structured documents. All of them suppose that the semi-structured documents cannot have always a pre-defined structure and that each document has its own structure.

Nevertheless, we can notice that documents describing the same type of information, have usually similar structures (example cv, documentary emission, etc.) and/or are annotated by the same set of metadata. It would be then interesting to be able to find these similarities and to deduce generic documents classes and not to remain at a specific level. The use of these generic classes will facilitate the exploitation of the bulky documents warehouses contents by focusing research on only the needed collection.

Moreover, the majority of these models do not provide a clear separation between descriptions of the structure and of the document contents. Which induces a lack of clearness, consequently documents handling becomes harder.

To be able to extract more semantic information from multimedia documents, some researches has been reported on audio content classification and highlight detection. For example, Liu Z. and al. in [7] study the problem of classifying TV broadcast into five different categories: news, commercial, weather forecast, basketball game, and football game by using a set of low-level audio features for characterizing semantic content of short audio clips. In sports video analysis [17], highlight events are detected based on special audio effects like cheering, ball-hit, and whistling. While in film indexing [12] sounds like car-braking, siren, gunshot, and explosion are used to identify violent scenes in action movies.

We propose to melt the results given by this content classification approaches to the previous structure classification to provide an efficient storage and exploitation of multimedia documents.

The multimedia description standard MPEG-7 is an international standard since February 2002. It defines a huge set of description classes for multimedia content, for its creation and its communication.

The fundamental difference between this standard and the multimedia document warehouse modeling is revealed on how both approaches attribute importance to single descriptive elements. For example, MPEG-7 proposes a variety of tools (MediaLocator DS) to specify the "location" of a particular image, audio or video segment by referencing the media data. This is important for cross-referencing of metadata and media data. However, such effort is not of so much interest for document warehouse exploitation.

MPEG-7 standardizes the information exchange of descriptive information. However, it is not suitable to serve as a multimedia document model. In spite of the difference in requirements of MPEG-7 and a multimedia document warehouse model, they have the potential to work together to build a distributed multimedia system [6]. To conclude, this standard provides us a structured metadata description for semantically rich media content. But it should not be considered as a competitor to broadly used multimedia document models.

3. Our Approach

Models studied previously have a lack of flexibility in the multi-media documents handling. This is mainly related to the fact that either they are focused on only one media, or they present the totality of the document insofar as semantics and structure are blended. However, it is important to be able to model all the concepts related on both structural aspects (hierarchical, temporal, space) and semantic aspects. In this section we present our approach to provide a flexible exploitation of multimedia documents. We show our multimedia document warehouse metamodel and the techniques we use to handle the warehouse content.

3.1. Document Warehouse Metamodel

We propose a metamodel of document warehouses that cover any type of information (text, image, video, and audio).

The warehouse constitutes so a centralized and perennial repository for the various used and manipulated documents.

In order to integrate heterogeneous and disseminated documents, we propose an instantiation process ensuring the structure and the content extraction of the document.

Our metamodel contains two parts:

- the structural description: it provides a description of warehouse documents based on their structure elements,
- the metadata description: it identifies and organizes the metadata for the various logical structures of the warehouse in order to describe the document contents. Thus, the documents belonging to a same logical structure will be described by the same set of metadata.

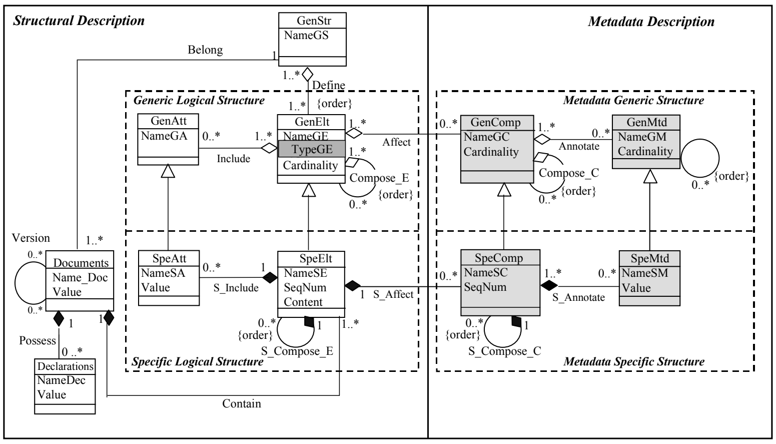


Fig 1. The Metamodel of Multimedia Document Warehouses

In what follows, we identify the structural and the metadata descriptions. We present also an instantiation example of these two parts.

3.1.1 Structural Description. The structural part of metamodel dissociates the warehouse document structures of their contents. This dissociation ensures thereafter an easier storage and exploitation (interrogation and analysis) of documentary information.

This part includes two structure types:

- the generic logical structure is the common structure of a document group (scientific publications, news flash, management reports, etc.). A generic logical structure is defined by a set of generic elements, which can be composed of other generic elements. Each of them can also be described by generic attributes,
- the specific logical structure is the structure of one document. A document is characterized by a set of declarations. It contains specific elements, which are eventually associated to specific attributes.

Compared to existing works, the structural description presents the following interests: (1) for the document storage, the regrouping and the classification of the warehouse documents according to common similar structures, (2) for the document manipulation, the description of the document granularity ensures an access more localized and retrieval more relevant of the needed information.

3.1.2 Metadata Description. To exploit easily and perfectly multimedia document contents and their semantic richness [10] the metadata structure splits up the warehouse generic elements into smaller components. These components can be annotated by an unfixed set of metadata for each type of media.

The metadata part includes also two structure types:

- the metadata generic structure details the composition of the document content. This structure is characterized by generic components of the different generic elements. These components can be described by metadata. This structure corresponds so to a finer decomposition of the multimedia generic elements.
- the specific metadata structure corresponds to a specialization of the metadata generic structure. It details the description of the document content.

This metadata description ensures the following objectives: (1) a better analysis of multimedia components (audio, video, and image) by offering a detailed composition of them, (2) a better management of semantic richness of multimdedia documents.

To describe the document contents, we propose to use metadata. These metadata are extracted by a set of specific tools for each type of media. They are defined as data on the data. More precisely, they constitute a structured set of information describing an unspecified resource. The metadata are particularly important for the audio and visual resources that, without them, can remain practically not exploitable. The users depend indeed on the information added to the image, audio or video to carry out relevant and precise retrieval.

3.1.2.1 Audio annotation. At the moment, our works are focused in audio documents. So, in what follows, we present the set of components that can be extracted from this media as well as a non-exhaustive list of metadata, which can be used to annotate these components. We give also some bibliographical references proposing methods and tools that ensure the metadata extraction.

Nowadays, more and more digital audio data are used, either stand-alone (music, radio broadcasts) or combined with other media (visual and/or textual) into multimedia documents. However, most of the audio data are not indexed, which makes the contained information difficult to reuse. The audio content indexing is likely to facilitate the management of audio data and support various multimedia applications where this data is involved.

For an audio element, we can find different descriptors according to signal nature. First information, which can be used to provide a temporal segmentation of the signal, is the Speech/Music decomposition. Each segment is characterized as being speech, music, speech and music or noise (neither speech, nor music) [16]. It seems also very interesting to locate the zones of silence which ensure to mark for example the changes of speakers, topic, etc. Lu L. and al. in [2] present schemes to classify audio signals into four classes, including speech, music, noise, and silence.

On music segments, we can extract for example:

- instruments and melody contour descriptors [14],
- jingles: a jingle is a few seconds of music that characterizes an emission or a part of it. A tool for automatic detection is proposed by [16],

Other information related to speech segments can be used:

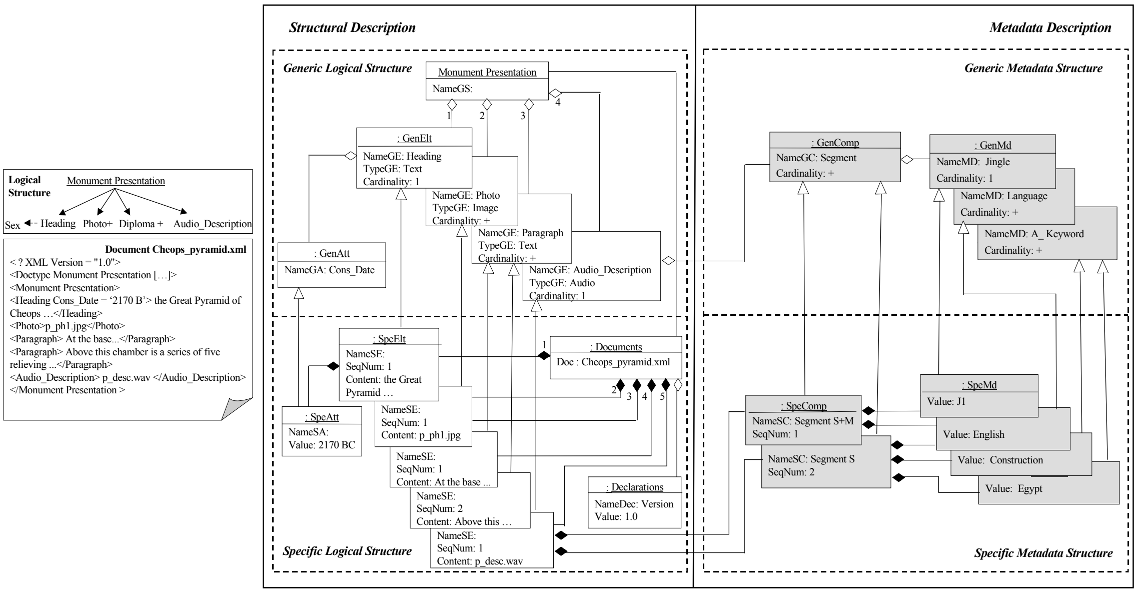
- speakers identification and tracking [2], which consists in seeking all the segments which were pronounced by a particular speaker. We can also extract specific information about the speakers like the gender (man, woman) and the age,
- textual transcription [4] and audio key words [13],
- topic which determines the general subject of a speech segment by employing a taxonomy of audio key words [18],
- language. In the same document, several languages can be spoken by one or more speakers. This metadata ensure also the best choice of the techniques that will be applied to extract topic and audio keywords [15].

The metamodel of multimedia document warehouses we propose could use these metadata to carry out a better content-based annotation of the warehouse contents. Our concern does not consist in extracting this metadata but rather proposing a description able to handle it. We do not impose the use of fixed structures but we allow the user to employ its own structures and metadata. Thus, our approach ensures a better flexibility to the multimedia document exploitation.

In what follows (cf. Fig 2), we present the structural description of the document "*Cheops_pyramid.xml*" and the metadata description of the generic element "Audio_Description" only instead of a complete example for reasons of clearness.

This figure shows that the metamodel is able to describe the multimedia document contents. Thus, for the element "Audio_Description", we can use specific tools [13] to compose it in two audio segments ("speech and music" and "speech"). These tools affect also the metadata "jingle", "language" and "keywords" for these segments.

This additional information is then instantiated in the warehouse for next retrievals or exploitations. The capacity to integrate this description ensures a more flexible management of the multimedia documents. For example, we can interrogate the warehouse content to find monuments having an "Audio_Description" marked in "English" and that contains the keyword "Egypt". In the following section, we describe the document integration in the warehouse according to the proposed metamodel (structural and metadata parts).



```
< ? XML Version = "1.0">
<Doctype Monument Presentation […]>
<Monument Presentation>
<Heading Cons_Date = ‘2170 B’> the Great Pyramid of
Cheops …</Heading>
<Photo>p_ph1.jpg</Photo>
<Paragraph> At the base...</Paragraph>
<Paragraph> Above this chamber is a series of five
relieving ...</Paragraph>
<Audio_Description> p_desc.wav </Audio_Description>
</Monument Presentation >
```

Fig 2. Example of the Metamodel Instantiation

3.2 Implementation and experiments

In order to validate our proposals, we have made a prototype of assistance to the multimedia document integration and analysis MDOCWARE (Multimedia DOCument WAREhouse). The instantiation of metamodel structural part was carried out in an automatic way through a Perl parsor [5]. However, the automatisation of the instantiation of the semantic part is under development. We present in this section our experimental base. Then we describe some taxonomies used to instantiate the proposed metamodel. We present finally the exploitation approach as well as an example of analysis offered by our prototype.

3.2.1 Experimental base. Our experimental base is made of a set of radio broadcast news from Radio France International (RFI).
These documents were annotated within the framework of Raives project [13]. The following figure shows an example of this annotation.

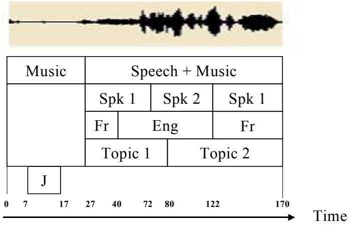


Fig3. Audio document annotation

3.2.2 Document integration. For the audio elements, the first level of composition must be the temporal axis around which we can articulate (in hierarchical way or not) the set of metadata. The components of this first level must cover entirely the document otherwise we could not represent the totality of it. For example, the segmentation of an audio document according to the decomposition Speech/Music/Noise ensures this blanket. While segmentation according to spoken languages is unable to ensure it because it does not represent the segments without speech transcription. We thus chose a decomposition of audio document in section: "speech", "Music", "speech+Music" and "Noise" segments.
In order to synchronize the various fragments (components and metadata) of an audio element, it is necessary to manage their beginning and finish moments. To ensure this synchronization without losing the homogeneity of our metamodel, we present these information as metadata, which will be linked, to each fragment. Indeed, this choice avoids the addition of a field st_end (start_end) in the classes CompSpe and MtdSpe which will be filled only for the fragments which make an audio element. In addition, we do not impose the use of fixed structures or predefined metadata but we allow the user to identify his own classification and level of granularity. Thus, annotations can handle either components, or metadata according to the decomposition wanted by the user. In our works, we have used a set of taxonomies to integrate the audio documents into the warehouse. A taxonomy reflects a

particular organization of the metadata generic structure. In what follow, we give two examples of these taxonomies (cf. Fig 4).

In the first one we use the "speaker" as being metadata associated with the components "Speech" and "Speech+Music". We also use the "language" like metadata linked to the "speaker" and the "topic" like a metadata linked to the "language". The "topic" will be annotated with the audio transcription "A_Trans". "Music" and "Noise" will be annotated with "jingle" and "special effect" (applause, laughter, etc.).

In the second taxonomy, the audio element is composed of sections which can be annotated by the metadata "topic", start and end time "S_St_End", speaker" and "S_Eff". The metadata "name", "gender", "accent" and "sequence" are linked to the "Speaker". The "sequence" is annotated by the "Language", the audio transcription "A_trans" and its start and end time "A_St_End".

The difference between these two taxonomies is due to (1) the richness level of the information that can be given by the document annotation and (2) the semantic relations that linked the taxonomy fragments.

Indeed, to perform efficient analyses, we must present in the taxonomy leaves the fragment that has more values.

For example, a section of a "documentary emission" can be interested by a unique topic ("Music in Brazil") using several languages ("French", "English" and "Spanish"). However in a "news flash" we can use one language ("English") to talk about several topics ("political", "sport", etc.). Among these taxonomies, the first one gives a deeper hierarchy level to the "topic". So it is more efficient to use it to instantiate a "news flash". In the same way, the second taxonomy is useful for the instantiation of a "documentary emission".

To ensure the integration of the audio documents in the warehouse, we start by analyzing XML files provided by the Raives project. We extract the various annotation layers (Speech/Music layer, speakers layer, languages layer, etc). We use then the appropriate taxonomy to create hierarchical and synchronization links between these layers (the automation of the choice of the suitable ontology is under development). These links allow us to have additional information providing a better management of document contents. For example from the annotation presented in figure 3, we can deduce that "Topic 1" is approached between the moments 27 and 80. We can also deduce that it is marked in French from moment 27 to moment 40 and in English from moment 40 until moment 80. This topic is approached by the speaker "spk 1" from moment 27 until moment 72 and by speaker 2 from moment 72 until moment 80.

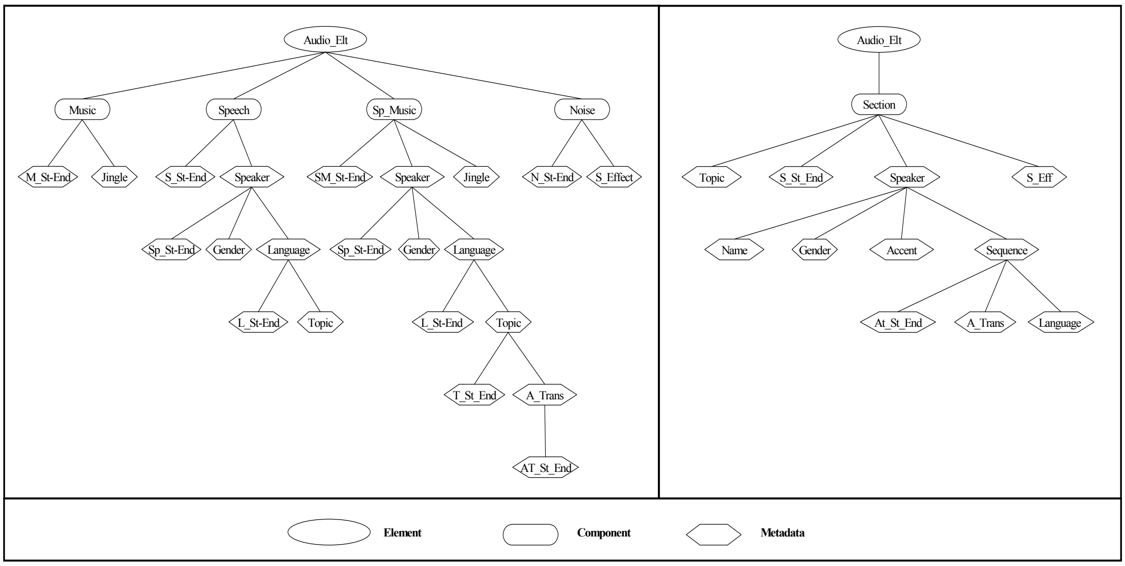


Fig 4. Examples of Metamodel instantiation taxonomies

Once the document description is stored in our warehouse according to the hierarchy suggested by the suitable taxonomy, several types of exploitation and analysis may be carried out.

3.2.3 Document exploitation. Our prototype offers three type of analysis: (1) by generic structure (generic logical structure and generic structure of metadata) which represent a set of identical documents (news flash), (2) by document, our analyses in this case will relate to the specific structure of a particular document (the news flash number 12) or (3) by generic fragment (element, component or metadata), we use generic fragments which can belong to different structures (the speaker John Smith who presents both news flashes and documentary emissions).

We give in what follows an example of analysis.

Example: In the news flashes, we want to find topics expressed by the speakers "Alain Dupont", "Amélie Cabaliero" or "Sonia Buffar" for more than 15 minutes. We want also that results show the list of retained speakers and topics as well as the time spent by each speaker to talk about each topic.

To carry out this analysis we should flow these steps:

(1) Choice of the analysis type

The first step consists in selecting the type of analysis (by generic structure in our case). Thus, the system displays the list of all existing structures in the warehouse. Among these structures, we must select the generic structure "news_flash". Once the structure was chosen, the system displays it in an automatic way (cf. Fig 5). Each fragment is preceded by its cardinality ("": one occurrence, "?": zero or one occurrence, "+": one or more occurrences, "*": zero or several occurrences).

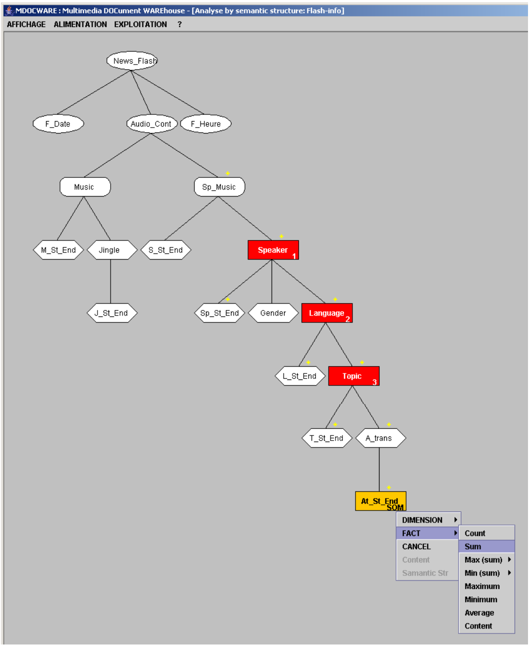


Fig 5. Selection of analysis fragments (example1)

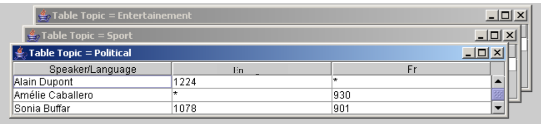
(2) Selection of the analysis fragments

In this level, we must select and define the analysis fragments: i.e. To specify the fact (subject of analysis) and dimensions. The assignment of these roles is done through contextual menus (cf. Fig 5). We must point the desired fragment and fix our choice (fact or dimension, by a click on the right button) as well as the attributes, namely: the order for dimensions and the formula for the fact (Count, Sum, Maximum, Minimum, Average, etc). In our example, the first dimension is "the speaker", the second is "language" and the third is "the topic". The fact is "the sum of the duration of intervention".

(3) Filtration

We wish to restrain our research only in three speakers. The filtration system displays all values of the element "Speaker". Thus, it provides the user to choose the corresponding names among these values. A second constraint indicates that the sum of the duration must be higher than 900 seconds. We must then apply a second filter to the fact value.

The system displays the results in the form of multidimensional tables. The first dimension is represented on lines, second is represented in columns. Each table represents a value for the third dimension. The results of the fact are represented into the tables in the form of interrelationships between the various dimension values. For example, "the speaker" "Sonia buffar" talk in "English" about "Political" during 1078 seconds.



Speaker/Language	En	Fr
Alain Dupont	1224	*
Amélie Caballero	*	930
Sonia Buffar	1078	901

Fig 6. Analysis results (example1)

4. Conclusion

Document dematerialization and increases in data storage capacity have made massive amounts of audio, video, and images available from disseminated sources. However, digital media documents are rich in content and lack generally structured and descriptive metadata. These metadata would allow indexing and easy access to required information. Consequently, modeling and indexation of multimedia components are clearly needed for storage and searching of such documents. Indeed, representation and semantic annotation of multimedia content have been identified as important steps towards more efficient manipulation and retrieval of multimedia documents.

Our research works are interested in the identification and the representation of logical structures as well as semantic ones through the metadata associated with multimedia documents. Our goal is to handle and exploit information contained in relevant documents extracted from heterogeneous and disseminated sources. This paper presents a solution based on the creation of a multimedia document warehouse. More precisely, we propose a warehouse metamodel able on the one hand to integrate and gather any type of documents and on the other hand to enrich the integrated information by the extracted metadata for each type of media. This metamodel gives more flexibility in the document manipulation.

The validation of our proposal is based on the conception of a prototype (MDOCWARE) which ensures the management of multimedia documents warehouses.

We see the possibilities to further improve the proposed work mainly by (1) the extension of our prototype to make it able to handle the video components (2) the implementation of techniques that ensure the classification of semantic structures and (3) the carrying out of more thorough experiments in order to evaluate in a quantitative and qualitative way the realized tool MDOCWARE.

5. REFERENCES

[1] I. Amous, I. Jedidi, and F. Sèdes, "A contribution to multimedia document modeling and organizing", in *8Th International conference on Object Oriented Information Systems, OOIS'02*, Montpelier, France, 2002, Springer LNCS n° 2425, pp. 434-444.

[2] C. Barras, X. Zhu, S. Meignier, and J-L Gauvain, "Improving Speaker Diarization", in *DARPA RT04. Palisades*, New York, USA, 2004.

[3] J. Darmont, O. Boussaid, and F. Bentayeb, "Warehousing Web Data", in *4th International Conference on Information Integration and Web-based Applications and Services (iiWAS 02),* Bandung, Indonesia, 2002, pp. 148-152.

[4] J-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System", in *Speech Communication*, vol. 37, no. 1-2, 2002, pp. 89-108.

[5] K. Khrouf, and C. Soulé-Dupuy, "A Textual Warehouse Approach: a Web Data Repository", *Intelligent Agents for Data Mining and Information Retrieval*, Idea Group Publishing, 2004, Chapter VII, pp. 101-124.

[6] H. Kosch, "MPEG-7 and Multimedia Database Systems", *SIGMOD Records ACM Press*, 2002, pp. 34-39.

[7] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," in *J. VLSI Signal Processing Sys. Signal, Image, Video Technology*, 1998, Vol. 20, pp. 61-79.

[8] E. Loisant, H. Ishikawa, and J. Martinez, "Designing a Model Independent Multimedia Database", in *Days of Science and Technology,* Tokyo, Japan, 2002.

[9] L. Lu, H-J. Zhang, and S. Li, "Content-based Audio Classification and Segmentation by Using Vector Machines", in *ACM Multimedia Systems Journal 8 (6)*, 2003, pp 482-492.

[10] M. Mbarki, and C. Soulé-Dupuy, "A Conceptual Modeling of Multimedia Documents", in *IADIS International Conference www/Internet*, Madrid, Spain, October, 2004, IADIS - ISBN 972-99353-0-0, vol 2, pp. 1051-1056.

[11] N. Moënne-Loccoz, B. Janvier, S. Marchand-Maillet, and E. Bruno, "Managing Video Collections at Large", in *First International Workshop on Computer Vision meets Databases (CVDB 2004)*, Paris, 2004.

[12] S. Moncrieff, C. Dorai, and S. Venkatesh, "Detecting Indexical Signs in Film Audio for Scene Interpretation", in *ICME*, 2001, pp. 1192-1195.

[13] N. Parlangeau-Vallès, J. Farinas, D. Fohr, I. Illina, I. Magrin-Chagnolleau, O. Mella, J. Pinquier, J-L Rouas, and C. Sénac, "Audio Indexing on the Web.: A Preliminary Study of Some Audio Descriptors", in *SCI* 2003, Orlando, Florida, USA July, 2003.

[14] G. Peeters, St McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7", in *International Computer Music Conference*, San Francisco, 2000.

[15] F. Pellegrino, J. Farinas, and J-L Rouas, "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech", In *International Conference on Speech Prosody Nara*, Japon, 2004, pp. 517-520.

[16] J Pinquier, and R. Obrecht,: "Jingle detection and identification in audio documents", in *ICASSP'2004*, Montréal, Canada, May (2004).

[17] Y. Rui, A. Gupta, and A. Acero, *Automatically Extracting Highlights for TV Baseball Programs*, in *8th ACM Multimedia*, 2000, pp. 105-115.

[18] JP. Yamron, S. Knecht, and P. Van Mulbregt, "Dragon's Tracking and Detection Systems for the TDT2000 Evaluation", in *Topic Detection and Tracking Workshop,* 2000, pp. 75-79.