A DOCUMENT REPOSITORY ARCHITECTURE FOR HETEROGENEOUS BUSINESS INFORMATION MANAGEMENT

Mohamed Mbarki [1], Chantal Soulé-Dupuy [1, 2] and Nathalie Vallès-Parlangeau [1,2]
[1] *IRIT, SIG/D2S2 Team, Paul Sabatier University 118 Route de Narbonne, 31062 Toulouse, France*
[2] *Toulouse I University, 2 rue du Doyen-Gabriel-Marty 31042 Toulouse, France*

Keywords: Repository architecture, architecture modules, document integration, document exploitation.

Abstract: As part of business memories, document repositories should bring some solutions to ensure flexible and efficient uses of dematerialized information content. While the fields of repositories modeling, document integration and interrogation have independently attracted a huge amount of attention, few works have tried to propose a general architecture of document repository management. Thus we propose a repository architecture based on the integration of different complementary modules ensuring an efficient storage of fragmented digital documents and then flexible fragments exploitation. This paper presents also an implementation of such architecture of document repository.

1 INTRODUCTION

Today's increasing of documents sources makes of their integration and exploitation a real need. The document integration provides a unified representation of heterogeneous documents owing to generic documents model. This unified modeling will facilitate documents access and exploitation. The use of documents repositories seems to be a relevant solution to achieve these goals. These repositories permit to constitute shareable spaces 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 repository must allow their processing according to several viewpoints, and via several techniques (document retrieval, multidimensional analysis, text mining, etc). Indeed, it should collect and classify the information, by means of suitable procedures, in order to produce synthesis that will be used to support decisional processes, as well as to globally control the enterprise activities. To leverage the impact of document repositories use in the handling of large sets of documents and in the decision maker process, several challenges should be addressed. These challenges include lack of flexibility to customize information storage, exploitation and presentation according to user's profiles.

The management of document repositories can be ensured by Electronic Document Management System (EDMS). An EDMS serves as an access portal to other applications. Hence, an EDMS allows to dematerialize, to classify, to store and to manage documents from computer applications in the frame of the organization activities. Our proposition presents an extension of EDMS functionalities concerning the integration and the exploitation of heterogeneous documents. Indeed, these systems use usually fixed document organizations and they are disable to extract (complex) hierarchical structures and even less to carry out comparison between such structures in order to classify them. Moreover document exploitation within EDMS is based on simple textual interrogation languages. Hence, the use of graphical interrogation might ensure a flexible and easier exploitation (especially for novice users).

Our ultimate goals are to contribute to the resolution of the above issues and facilitate information access and exploitation to any user. We propose in this paper repository architecture to face the previous challenges. We start by an overview of existing works that aim at solving problems of document modelling and interrogation. The next section is devoted to a presentation of the document repository architecture we propose. We illustrate the use of this architecture through the development of an illustration example in the last section.

2 STUDY CONTEXT

The relevant issues in repositories content are the definition of document models and querying

Mbarki M., Soulé-Dupuy C. and Vallès-Parlangeau N. (2007).
A DOCUMENT REPOSITORY ARCHITECTURE FOR HETEROGENEOUS BUSINESS INFORMATION MANAGEMENT.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - DISI*, pages 192-198
DOI: 10.5220/0002364301920198

languages which permits to manage the heterogeneity of document structure, content and associated metadata.

2.1 Modelling

A document repository organizes and structures information for content retrieval. In this context, document modelling is one of the key issues. The modelling is used to determine which information should be stored in a repository 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 document description.

The first category gathers works which aim at modelling each type of media separately. These approaches do not manage the fitting of several media in only one document. (Loisant and al., 2002) 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. Each one corresponds to only one type of media. (Moënne-Loccoz and al., 2004) provide a model to manage the specificities of video documents. This model ensures 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 all the media composing the document. They transcribe links that connect the various mono-media components of the same document. (Amous and al., 2002) 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 and al., 2002) 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 (name, keywords, duration, etc.) of these documents to index them.

All of these works suppose that 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 or aiming at the same intention of use (example co, documentary emission, etc.) have usually similar structures and/or are annotated with 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 repositories contents by focusing research only on the needed collection (Mbarki and soulé-Dupuy, 2004). Moreover, the majority of these models do not provide a clear separation between the structure and the document contents descriptions. What induces a lack of clearness, consequently documents handling becomes harder.

2.2 Documents Exploitation

With a semi-structured and especially multimedia document, research cannot be based solely on a predefined tabular schema like in classic databases. Otherwise it would not permit the exploitation of the different metadata used to annotate document content. This need gave birth to a new generation of querying languages.

The contribution of LOREL (Lightweight Object Repository Language) (Goldman and al., 1999) resides in the flexibility of semi-structured documents interrogations, even though we do not know their structure. The easiness is offered by the introduction of path expressions.
XQuery (a Query Language for XML) (W3C, 2003) permits to interrogate a XML document according to different criteria. This language is often called the SQL of the XML. It permits to elaborate complex queries. It is a hybrid language between XPath and SQL. It is capable also to browse the arborescence of a XML document, to carry up the information required by the user and to create a new document containing only the needed granules.

The power of these interrogation languages resides in the manipulation of document structure. The documents used by these languages are generally constituted by short and precise elements (title, authors, etc.). To interrogate and to manipulate documents that contain elements having important size (section, paragraph, etc.), a combination of such languages with the information retrieval techniques is necessary. To overcome this shortcoming, we propose a specific language permitting to interrogate documents according to their structures and their contents. The noticeable difference between our proposition and the previous languages concerns the complexity of queries. Indeed, in our approach the user can see the document organization (tree) while interrogating. The graphical language that we propose provides on the one hand a best management and comprehension of the document composition and on the other hand an easy querying because the user does not need to have a previous knowledge about interrogation languages.

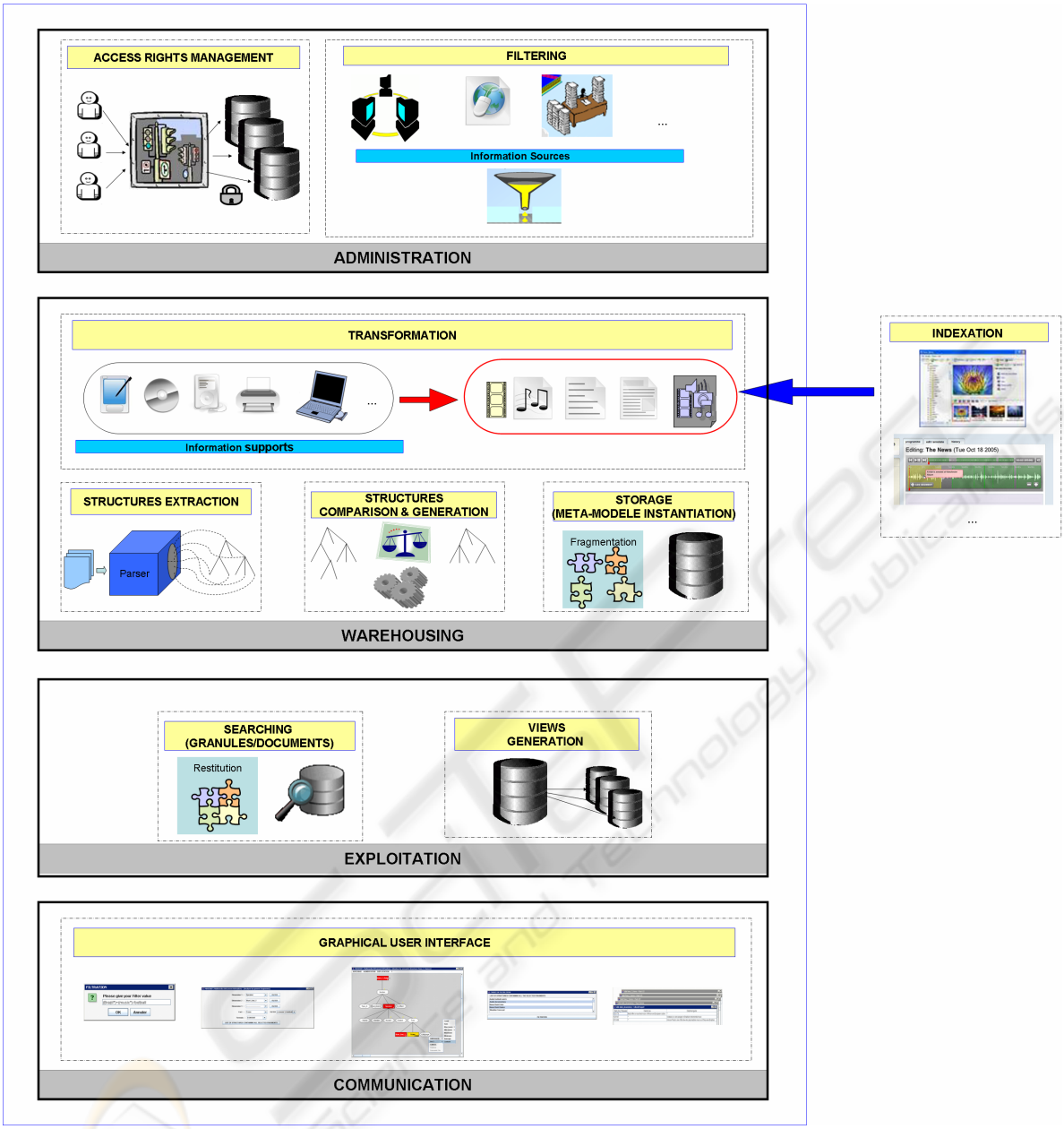

Figure1: Document Repository Architecture.

3 OUR PROPOSITION OF DOCUMENT REPOSITORIES ARCHITECTURE

The document repository architecture is composed of four modules (cf. Figure 1) that interact together in order to provide a global repository management. The first module ensures administration tasks by allowing the definition of the appropriate rules to integrate relevant information from disseminated sources. This module allows also repository administrators to fix a second set of rules concerning the access and the manipulation of the integrated information. The second parts contain a set of sub modules aiming at decomposing document into fragments, extracting and comparing structures and integrating both content and structure. The exploitation of the repository content is based on the creation of views that answer to a particular need of a specific user on presenting relevant pieces of global information. The manipulation of the

previous modules by the user is carried out through a graphical interface.

3.1 Administration

Administrators review metadata and content prior to final storage into repository. They control also different content submission and review parameters for each document collection and user groups defined within the repository. Moreover, they limit access to certain content based on the levels of user's rights.

With the overwhelming growth of multimedia information and their accessibility through disseminated sources (the Web, classic archives, enterprise networks, etc.), the information filtering become today an obvious need.

Once documents are stored, we can control two key points in the repository content manipulation: information access and information use. Access controls limit who can view, receive or download a document (according to user's rights). Use controls determine what a user can do once the document has been accessed (interrogation, update, suppression, etc.).

3.2 Document Warehousing

After their filtering from disseminated sources, digital documents must be rendered in some way to make them exploitable independently of their supports (PDA, CD, etc.). Many standards are already proposed to give unified presentation of document organization. As examples, we can quote XML (Extensible Markup Language), SGML (Standard Generalized Markup Language), MPEG7 (Motion Expert Picture Group), SMIL (Multimedia Synchronized Integration Language). Obviously, a previous step of dematerialization is needed if the selected document is on a paper support.

The sub module of structures extraction aims at extracting structures and contents of documents that will be inserted in the repository. The structure recognition is based on the identification of documentary granules (document fragments, a fragment is a small flexible and semantically homogenous unit). For example, XML annotations make easier the metadata identification. Such annotations can be generated by several tools and techniques of description: Transcriber, MPEG7, SMIL, etc. It should be noted here that our approach consists in extracting, integrating and organizing structures in the repository and not to annotate the multimedia documents.

To be able to store and handle a fragmented multimedia document under several facets and several points of view without being obliged to manage a huge amount of heterogeneous structures, it seems important to be able to classify the structures. Moreover, in order to ensure flexibility in storage it is significant to be able to model separately all the concepts related on both structural and contents aspects.

Our documents repository is composed of generic structures and specific structures (Mbarki and Soulé-Dupuy, 2004). A specific structure corresponds to only one document. A generic structure corresponds to a collection of similar specific structures. The document integration will follow this repository content organization.

Before insertion, the document content must be separated from its structure. The insertion of a document specific structure is a rather significant step because it determines the links between the fragment content and a particular generic structure. Thus, flexibility in the repository instantiation can be ensured through the flexibility of this phase of structure handling. This module provides three possibilities of structure insertion according to following cases (Mbarki and al., 2007):

- the user does not know the structure to which the document content can be attached. We then propose to compare the document structure with all repository structures. It will be merged with a similar structure or inserted as it is giving place to the creation of a new generic structure,
- the user knows a particular structure in the repository to which he wants to attach the document content. Hence, the structure of this document will be merged with this selected structure. The fusion of these two structures is ensured without comparison (directed insertion). The user requires in this case the transformation of the document structure towards a generic structure which already exists in the repository. To fulfill this fusion, we propose to use operations of suppression, substitution, displacement and modification of structure fragments (tags),
- the user judges that the document provided is representative of a new structure. This document will be then regarded as a typical document (one typical CV for example). The structure of this document will be inserted in the repository without comparison and thus without any modification.

In addition to these three approaches of structure insertion, the user can freely give the composition of a new structure to which he wants to attach

documents later. For example, it can propose the generic structure of a "thesis" or an "audio documentary".

3.3 Exploitation

We aim at managing both complex structuring of multimedia document and the heterogeneity of their content. In the same way, we want to provide an easy interrogation and pleasant result presentations. To achieve these goals, we propose to combine techniques of multi-dimensional analysis (OLAP: On Line Analytical Processing), information retrieval and structured document interrogations (SQL: Structured query language).

Several types of exploitation and analysis operations can be carried out on warehoused documents (using structure elements and content) (Mbarki and al., 2005). For example, three kinds of search and analysis can be carried out: (1) by collection or generic structures which represent a set of identical documents (news flashes), (2) by document, our analysis in this case will relate to the specific structure of a particular document (the news flash number 12) or (3) by generic fragment (any 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). The search results are generated and displayed according to intermediate views as represented thereafter.

Views are dedicated to a particular user or group of users. They must answer to specific needs or help users to analyze information content to achieve some decisions. These views are generated in two main steps:

- Step 1: Generation of an elementary view for each analysis parameter (dimensions and fact). A view is generated for every selected fragment (element of structure, attribute, component or metadata).
- Step 2: Joint and grouping of the different generated views. The final view (result) is generated by joining the different views generated previously.

Detailed examples of these two views will be given later in this paper.

3.4 Communication (Graphical User Interface)

Repositories users can manipulate the previous architecture parts through the communication module. This module provides pleasant, easy to use and yet powerful user interface to manipulate and to interrogate document repositories. A first set of interfaces are used by administrators to fix access and use rights.

Concerning the warehousing module the interface part allows:

- the selection of documents to be integrated into repository,
- the definition of personalized structure composition,
- the creation of structures comparison rules,
- visualization of both structure and content of documents already integrated into repository,
- the manipulation of synonymies dictionary.

For the exploitation module the communication layer permits to interrogate structure and content without requiring any knowledge of classic interrogation languages syntax. It provides the user an intuitive set of tools for hierarchically structured document searching, retrieved list navigation and search refinement. Particularly, these tools ensure:

- the selection of analysis type,
- the choice of analysis axes (fact and dimensions) and their criteria (dimensions order and fact formula),
- the filtering of analysis results,
- the creation and the visualization of user's views.

4 ARCHITECTURE IN USE

To illustrate our proposals, we have developed a prototype of assistance to the multimedia document integration and analysis MDOCREP (Multimedia DOCument REPository). Our experimental base contains a set of audio documents extracted from radio emissions of RFI (Radio France International) and of RM (Radio. Maroc). This collection was annotated in the framework of both projects Raives (Parlangeau-Vallès and al., 2003) and Ester (Galliano and al., 2006). In this section, we present an example of analysis concerning the exploitation module

If we what to know the number of topics expressed by each speaker in the news flashes number "310", "311" and "312" from "Ester corpus", we should flow these steps:

(1) Choice of analysis type

The first stage consists in selecting the type of analysis (by generic semantic structure in our example). Thus, the system displays the list of all the existing structures in the repository. Among these structures, we must choose the generic structure "News-F-E". (News flashes Ester) Once the choice of the structure is carried out, the system displays in an automatic way the tree of this structure as

showing in figure 2. The elements are represented by ovals, the components are represented by rectangle-ovals and the metadata are represented by polygons.

(2) Selection of components analysis

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. Figure 2). 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 language", the second is "the speaker" and the third is "the flash name". The measurement of the fact is "the count of topics".

(3) Filtering

To select only the flashes number "310", "311" and "312", we can use the filtering function offered by the prototype (cf. Figure 3).

The system displays the results in the form of multidimensional tables (cf. Figure 4). The first dimension is represented on columns, the second one is represented in lines. 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 dimensions values.

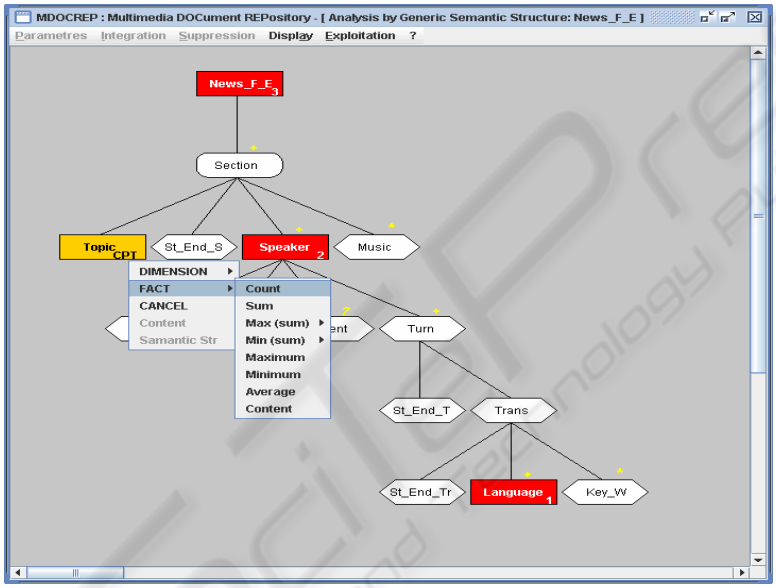

Figure 2: Selection of analysis components.

Figure 3: Filtering.

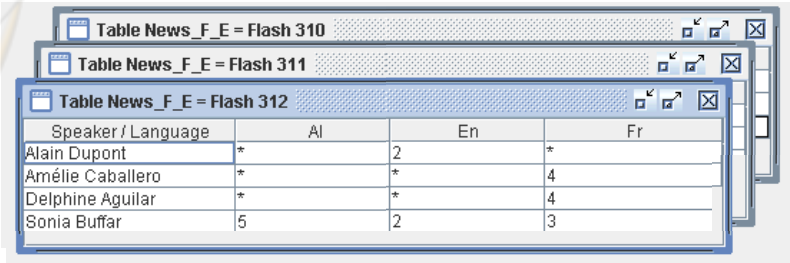

Figure 4: Results.

5 CONCLUSIONS

In this paper, global document repository architecture has been presented. Digital document warehousing is in the core of repository management. It provides information extraction, transformation and storage according to structure similarities. The document fragmentation, the partitioning in different kinds of structures and the independence of structuring mechanisms are significant since a document can be described independently of uses, of restitution mode for example. The repository administration presents also an important task. It allows to organize the integration, the update and the access to documents fragments.

The validation of our proposal is based on the realization of a prototype (MDOCREP). This prototype allows the multimedia documents repositories management by ensuring integration, interrogation and multidimensional analysis of information. We have integrated audio documents and some works are in progress to integrate video documents. The next goal is to propose an entirely automatic process for repository instantiation. We will concentrate also on quantitative and qualitative assessment of our interrogation approach.

Thus, this repository architecture is a part of a more general framework aiming at implementing a business memory management system. More than half of information used in organizations being stored in document, such repositories will permit to filter, share, have access and gain from such document contents.

REFERENCES

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

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

Goldman R., McHugh J. and Widom J., 1999. From Semistructured Data to XML: Migrating the Lore Data Model and Query Language. In *International Workshop on the Web and Databases (WebDB'99)*, p. 25-30, Philadelphia, Pennsylvania, USA.

Galliano S., Geoffrois E., Gravier G., Bonsatre J.-F., Mostefa D., Choukri K., Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In: *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006*), Genoa, Italy, May 2006, pp. 315-320.

Loisant E., Ishikawa H. and Martinez J., 2002. Designing a Model Independent Multimedia Database. In *Days of Science and Technology*, Tokyo, Japan.

Mbarki M. and Soulé-Dupuy C., 2004. A Semantic Modeling of Multimedia Document. In *the proceeding of IADIS International Conference WWW/Internet* Madrid-Espagne, vol2, pp. 1051-1056.

Mbarki M., Soulé-Dupuy, C. and Vallés-Parlangeau, N., 2005. Modeling and Flexible exploitation of Audio Documents. *In the proceeding of IEEE International Conference on Signal-Image Technology & Internet Based Systems*. Yaoundé, Cameroon, pp. 216-223.

Mbarki M., Soulé-Dupuy, C. and Vallés-Parlangeau, N., 2007. Multimedia Documents Management in a Multistructural Context. *In the proceeding of IEEE : Conference on Research Challenges in Information Science (RCIS)*, Ouarzazate, Morocco, pp. To appear.

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

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

W3C, 2003. XQuery 1.0: An XML Query Language, W3C Working draft.