# Querying and Browsing Multimedia Presentations

Augusto Celentano and Ombretta Gaggi

Dipartimento di Informatica, Università Ca' Foscari Venezia
via Torino 155, 30172 Mestre (VE), Italy
{auce,ogaggi}@dsi.unive.it

**Abstract.** Querying and searching the Web is an important research field which has drawn a number of concepts from databases and information retrieval fields, but has added its own models, requirements and techniques.
Multimedia information adds another dimension to the problem, when information is globally conveyed by different media which are archived and delivered separately, and must be coordinated and synchronized.
In this paper we discuss some issues about information retrieval in synchronized multimedia presentations. We introduce a class of multimedia presentations made of independent and synchronized media, and discuss how retrieval requirements can be defined. Then we discuss the need of a model for retrieving and browsing multimedia data in a way capable of integrating atomic media objects in coherent presentations.

## 1 Introduction

Multimedia is entering all fields of communication, and information presentation is becoming more and more rich. The Internet technology is growing and delivery of complex multimedia presentations, made of several continuous streams and static documents, does not suffer any more of technical constraints that have characterized the early stage of the World Wide Web.

Today it is possible to build and deliver video and audio information in a Web environment for domains who demand a good quality of images and audio and information organization according to complex schemas.

As the Web size grows the traditional navigation paradigm shows unbearable limits, and querying becomes a mandatory entry point for any meaningful exploration. Querying and searching the Web is an important research field which has drawn a number of concepts from databases and information retrieval fields, but has added its own models, requirements and techniques.

Multimedia information adds another dimension to the problem, when information is globally conveyed by different media which are archived and delivered separately, and must be coordinated and synchronized.

In this paper we discuss some issues about information retrieval in synchronized multimedia presentations. We introduce a class of multimedia presentations made of independent and synchronized media, and discuss how retrieval

M. Tucci (Ed.): MDIC 2001, LNCS 2184, pp. 105–116, 2001.

requirements can be defined. Then we discuss the need of a model for retrieving continuous data in a consistent way, able to reconstruct the fragments of a presentation from the atomic components returned by the query execution.

## 2 Querying Synchronized Multimedia Presentations

We consider hypermedia presentations made of one or more continuous media file, such as video or audio streams, which are presented to a user in a Web-based environment. As streams play, static documents, such as images or text pages, are sent to the browser and displayed in synchrony with them. From the user point of view the documents constitute a whole presentation which is coherent in terms of time relationships among the component media item, like a single compound stream. The user can interact with the presentation by pausing and resuming it, and by moving forward or backward along its timeline in a VCR style.

In a Web environment documents can be linked to other documents, therefore the user can follow a link, possibly reaching a completely different context. From a usability point of view the presentation should in some cases be paused, or stopped, while in other cases it could continue playing. As a general requirements, after any user action media delivery and playback must be coordinated in such a way that the user always sees a coherent presentation.

News-oriented applications (e.g., news-on-demand and Web advertising) and distance education are good representative of such presentations. The latter is a quite old scenario that at different phases of technology development has been used as a test bed for multimedia applications, due not only to its interest as a strategic field (e.g. in terms of long-life learning) but also because it demands high quality in terms of usability. The former is a relatively new field in which richness of information, variety in presentation and ease of interaction can attract the Web user. Lessons are made of text, animated presentations, slides, talks, movies, and of course of linked supplemental material that a student can access during the lesson itself. News are presented through audio or a video clips, and images and texts are displayed according to the particular argument the speaker is talking about.

### 2.1 Multimedia Retrieval

Retrieving information in such documents is a task which presents a number of peculiarities. First, information is contained in documents of different types; i.e., querying must retrieve information from different media at the same time. Second, the user is not interested in any single document but in the part of presentation in which at least one of the documents satisfy the query; i.e., the retrieved item is only a component of a part of the presentation which is the true query result. Third, being the presentation dynamic, a coherent and understandable segment must be returned to the user; i.e., the query result extent must be identified according to a context which takes into account the presentation structure.

The first issue does not present conceptual problems with respect to information retrieval state of art. Multimedia databases encompass multiple representations of data in order to answer queries on different media types. We do not suggest that such a problem is of small concern or simple, or that all technical aspects are solved. For example, the language (textual or pictorial) used to formulate the query belongs to a specific syntactic domain against which data representation of different types must be matched (see [5] for an extensive survey of combined text, audio and video retrieval in news applications).

The second and the third issues require that the static organization of a presentation, which relates media together, and the temporal aspects of its behavior, which describe how the different segments evolve, be defined according to a structured model.

### 2.2 Managing Query Results

At a first glance the result of a query on a multimedia database is a set of media objects which respond to the query criteria. Each media item can be a (section of) audio or video file, but also a non continuous medium like a text page or an image. If retrieved items are also part of a composite multimedia document, such as a complex presentation, they cannot be displayed or played alone: for example, a slide of a lesson without the accompanying audio track, which is recorded separately, is scarcely useful.

For this reason, it is not possible, in general, to return to the user only an index of retrieved media items. We need to retrieve also any other items belonging to the composite original presentation, which has to be reconstructed and shown.

The correctly retrieved information must be a list (an index) of complete fragments of different multimedia documents, in which at least one medium item responds to the query criteria. The system should retrieve all media items that belong to a coherent section of the presentation, and display them to the user in a coordinated and synchronized way.

A consequence of such a behavior is that the number of media objects retrieved can be different from the number of results shown to the user. Each media item can be part of a different presentation, or some of them can belong to a same presentation. If two media objects belong to the same multimedia document, they can be part of different sections of the document, or be close in time: in the first case, they must be proposed as two different results, in the second case they must presented as a same result.

For these reasons, the retrieval system must be integrated with knowledge about the structure of the presentations in order to show the results in a complete and coherent way. Two kinds of information are needed:

- the hierarchical structure, for deciding when two items are part of a same section, and

- the temporal relationships, for displaying different results according to the author design of the presentation.

## 3 Related Work

### 3.1 Modeling Multimedia Presentation

Many research papers have presented different ways of specifying structure and temporal scenarios of a hypermedia documents.

Amsterdam Hypermedia Model [7,8] was the first serious attempt to combine media items with temporal relations into hypertext document. AHM structures document objects into atomic and composite components. Media items are played into channels while synchronization inside composite components is described by synchronization arches and offsets that establish time relationships between two components or two anchors.

SMIL, Synchronized Multimedia Integration Language[12], is a W3C recommendation defined as an XML application. It is a very simple markup language that defines tags for presenting multimedia objects in coordinated way. Synchronization is achieved through two tags: `seq` to render two or more objects sequentially and `par` to reproduce them in parallel. Using attributes it is possible to play segments inside the time span of an object.

In [1,2] we discuss the problem of authoring and navigating hypermedia documents composed of continuous and non continuous media objects delivered separately in a Web-based environment. We have introduced a model which defines a static structure and synchronization relationships among media objects of a presentation. Our model overcomes the limitations of SMIL because it considers user interaction. The user can interact separately with each single object of the presentation and the presentation is synchronized accordingly. SMIL native features allow interactions only with the whole document.

### 3.2 Multimedia Querying and Browsing

Problems related to querying, browsing and presenting multimedia information has been largely investigated in the literature.

In [9] the authors describe an integrated query and navigation model built upon techniques from declarative query languages and navigational paradigms. The system implemented provides facilities to help not expert user in query writing activity. Navigation hierarchies are used to present to the user summary information instead of a simple listing of query results. Differently from our approach, the authors never consider timing relations between objects, while they manage to hide querying of heterogeneous data on distributed repositories with an unified user interface.

Delaunay$^{MM}$[4] is a framework for querying and presenting multimedia data stored in distributed data repositories. The user specifies a multimedia presentation spatial layout by arranging graphical icons. Then, each icon is assigned

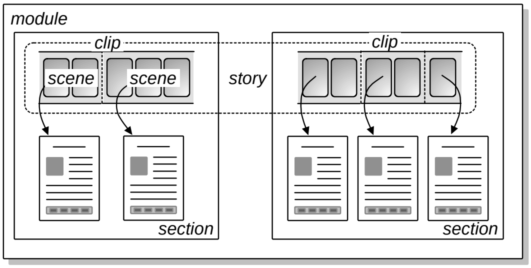


**Fig. 1.** Hierarchical Structure of Multimedia Presentation.

to a query, thus combining data selection with presentation. Delaunay$^{MM}$ uses ad hoc querying capabilities to search each type of media item in distributed database and on the Web. Also in this proposed paper, the authors did not address any solution for the specification of temporal synchronization among the objects.

In [10,11] the authors present TVQL, a multimedia visual query language for temporal analysis of video data. They consider a video of a teacher's lesson in classroom and annotate the video to identify interesting events, such as a student question or the teacher talk. TVQL (for temporal visual query language) enables user to browse for temporal relationships between two objects subsets. For example the user can search for which student speaks frequently after a teacher talk. The authors do not approach complex presentations with heterogeneous synchronized objects, but only a single video stream at a time.

In [3] a model is presented that fully integrate browsing and querying of hypermedia data capabilities. The model gives particular emphasis to structured information and combines two components: the hypermedia component and the Information Retrieval (IR) component. The hypermedia component contain information about both structure and content of a document. This integrated system allow users to compose queries which contain both kind of information in their formulation. The IR component contains information about the model and the resolution of the queries. Although it deals with composite documents, this paper doesn't consider time relationships among atomic objects of a structured documents.

## 4 A Model for Describing Hypermedia Dynamics

The model we have proposed in [1,2] describes the hierarchical structure of the document's components. A hypermedia presentation contains different kinds of media objects: static objects are referred to as *pages*; dynamic objects, video and audio clips, are hierarchically structured.

A multimedia presentation is composed of a set of *modules*, which the user can access in a completely independent way. The continuous media, an audio, or a video stream, which constitutes the main module content, is called a *story*. A

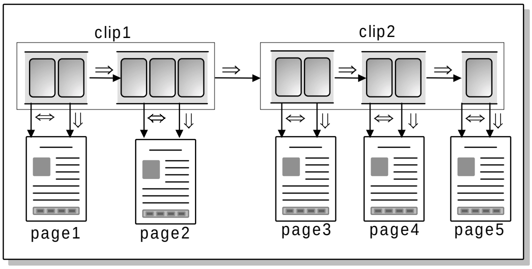


**Fig. 2.** Synchronization Relationships between Media Items

story is composed by a sequence of *clips*, each of which correspond to an audio or a video file. Clips are divided into *scenes*, each of which is associated to a static document, a *page*. A clip, with its scenes and associated pages, build up a *section*. Figure 1 pictorially shows this structure.

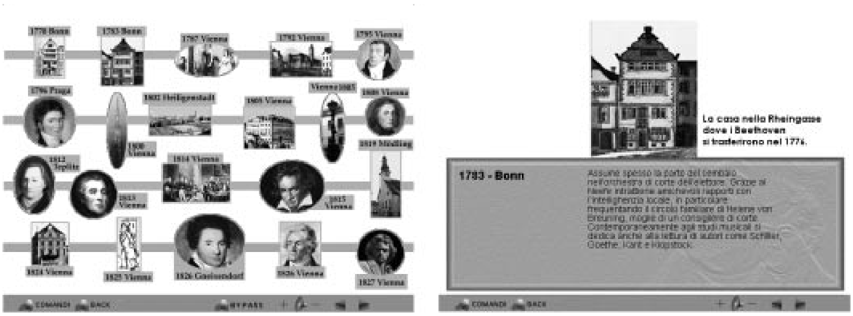
For example, in a distance education application, the video of a lesson can be divided into different *stories* according to the lesson topics. Stories are divided into *clips* according to the video segmentation into files and clips are divided into *scenes* according to the time a slide remains on the user screen. The slides of the lesson are the *pages* of the presentation which are displayed in turn as the lesson goes on. Media objects use *channels* (e.g., windows for displayable media) for their playback.

Synchronization is achieved with a set of primitives which define objects behavior during presentation's playback and channels' utilization. We have defined five synchronization relationships which describe the reaction of media objects to events. The events can be *internal*, like the beginning or termination of an object playback, or *external*, like a user action.

The synchronization primitives are:

- $A$ plays with $B$, denoted by $A \Leftrightarrow B$, to play two objects in parallel in a way such that object $A$ controls the time extent of object $B$;
- $A$ activates $B$, denoted by $A \Rightarrow B$, to play two objects in sequence;
- $A$ is terminated with $B$, denoted by $A \Downarrow B$, to terminate two objects at the same time as a consequence of an user interaction or of the forced termination of object $A$;
- $A$ is replaced by $B$, denoted by $A \rightleftharpoons B$, to force the termination of object $A$ so that object $B$ can use its channel;
- $A$ has priority over $B$ with behavior $\alpha$, denoted by $A \overset{\alpha}{>} B$, to stop (if $\alpha = s$) or pause (if $\alpha = p$) object $B$ when the user activates object $A$.

Figure 2 shows an example of such relationships for the module depicted in Figure 1. Each scene starts and ends a text page to which is associated. The whole document uses two channels, one for the clips and one for the pages. The reader is referred to [1,2] for a complete discussion of the model.

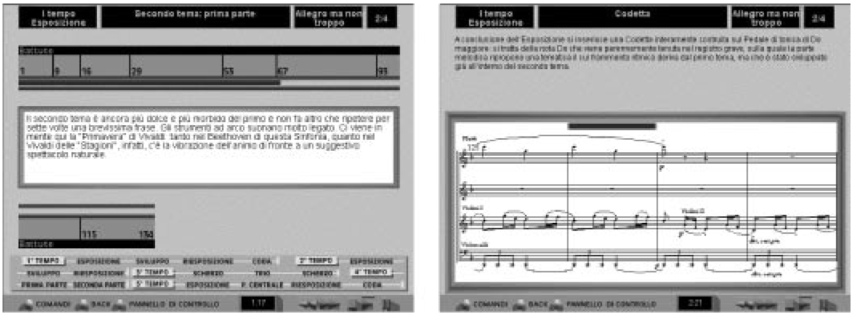


La casa nella Rheingasse
dove i Beethoven
si trasferirono nel 1776.

1783 - Bonn

Assume spesso la parte del cembalo nell'orchestra di corte dell'elettore. Grazie al Neefe intrattiene amichevoli rapporti con l'intellighenzia locale, in particolare frequentando il circolo familiare di Helene von Breuning, moglie di un consigliere di corte. Contemporaneamente agli studi musicali si dedica anche alla lettura di autori come Schiller, Goethe, Kant e Klopstock.

**Fig. 3.** Screenshots from a Presentation about Beethoven's Life.



Il secondo tema è ancora più dolce e più morbido del primo e non fa altro che ripetere per sette volte una brevissima frase. Gli strumenti ad arco suonano molto legato. Ci viene in mente qui la "Primavera" di Vivaldi: tanto nel Beethoven di questa Sinfonia, quanto nel Vivaldi delle "Stagioni", infatti, c'è la vibrazione dell'animo di fronte a un suggestivo spettacolo naturale.

A conclusione dell'Esposizione si inserisce una Codetta interamente costruita sul Pedale di tonica di Do maggiore: si tratta della nota Do che viene perennemente tenuta nel registro grave, sulla quale la parte melodica ripropone una terzina sul movimento ritmico derivato dal primo tema, ma che è stato sviluppato già all'interno del secondo tema.

**Fig. 4.** The Guide to Listening and Score Analysis of the *Pastoral* Symphony.

## 5 An Example from a Music Masterworks Collection

As a working example for the remainder of this paper we illustrate briefly the overall structure of a collection of multimedia CD-ROMs[1] featuring masterworks of famous musicians, published in Italy a few years ago[6]. Each CD-ROM follows a constant structure and contains a wide set of information about the musician's life, the historical period, the masterwork itself (as an audio file), a description of its musical structure in form of graphical score and text comments, a criticism, in form of written text, which accompanies the music play, and a set of audio files of related works. The CD-ROM collection could be viewed as a multimedia data repository for queries about biographies, musical structure and critical analysis, historical references and other kinds of information. Figures 3 and 4 show some screen shots from the CD-ROM on Beethoven Symphony no. 6 "Pastorale".

In Figure 3 a description of the life of Beethoven is illustrated. It is an animated presentation (the left image shows the final state) stepping through selected years in Beethoven's life. Each year is described by a short animation

[1] We use a CD-ROM based example because it is a complete product featuring a complex multimedia presentation, even if our target is the World Wide Web. In Section 6 we'll discuss the main differences between the two environments.

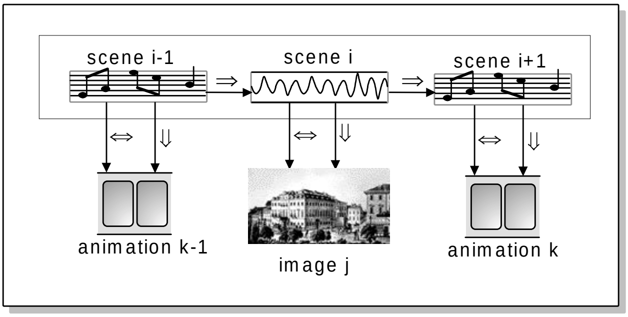


**Fig. 5.** The Synchronization Schema for the Presentation about Beethoven's Life.

which draws a segment of a timeline, displays an image and starts a spoken comment. A background music plays during the whole animation. As the narration goes on, a complete timeline is built. At the end of the presentation the user can click over an image to display a text page which reproduces the comment that was given by voice about that year, as shown in the image on the right.

The speaker's voice and the background music are integrated in the same audio file. According to our model, the audio track is a *clip*, which is divided into a sequence of *scenes* which interleave the speaker's voice and the background music. The animations which draw the timeline, between two events in Beethoven's life, are also *clips*, while the images displayed during the spoken comments can be considered *pages* since they are static media objects. Since the presentation starts with an animation, even numbered scenes correspond to voice comments and images, while odd numbered scenes correspond to the transitions from one year to the next.

Figure 5 shows the synchronization schema of this part of the presentation. The scenes play in sequence, hence $scene_i \Rightarrow scene_{i+1}$. Each scene which draws a timeline segment begins and ends with the related animation, and is described by the relationship $scene_{2i-1} \Leftrightarrow animation_i$. Similarly, each spoken comment goes in parallel with an image, as described the relationship $scene_{2i} \Leftrightarrow image_i$.

The user can stop the presentation by clicking over a "bypass" button. The audio track is terminated, and also the animation or the image displayed are terminated. This behavior is described by the relationships $scene_{2i-1} \Downarrow animation_i$ and $scene_{2i} \Downarrow image_i$.

Figure 4 shows two different but similar sections of this multimedia presentation: the analysis of the music work from the points of view of the artistic quality and the musical structure. In the left image the overall structure of the work is shown, and as the music play a bar moves showing the position in the score length. Text comments help the user to interpret the execution. In the right image the musical score is drawn and updated as the music plays. Also in this case some text comments are displayed. The regular structure of these two sections can be described, in terms of synchronization relationships, in a way similar to the example shown above.

## 6 Presentation in Multimedia Document Retrieval

We may notice that in several contexts the information which is presented to the user is composed of several media which are to be coherently harmonized in order to be meaningful. This is evident in the CD-ROM sections illustrated in Figure 4, where a text comment without the music and the context given by the visual interface is meaningless.

If we want to query or browse a multimedia database based on such an information structure we need to identify, retrieve and integrate all the components of the presentation segment which contains the relevant information.

In this paper we do not consider models and technologies for retrieving multimedia contents. Our concern is the reconstruction of a coherent multimedia document fragment given one or more of its components media objects. Therefore we assume (maybe with some imagination effort) that a system for retrieving multimedia information such as texts, voice talks, musical fragments, scores and images by content analysis is available, even if information other than text still exhibits a high degree of uncertainty in existing products or prototypes.

Let us suppose that we want to retrieve all the musicians in the data repository which have been living in Leipzig, and for each of them we want a description of such periods, or in general the information about their life in that town. We could find a set of voice comments, taken from the presentation of the musician's life, and the texts associated of them, as in Figure 3, plus other information chunks taken from different sections of the repository.

If we want to retrieve all the music passages where strings play in a *crescendo* style we could ask for such a text description in the guide to listening and in the score analysis, or perform a content analysis on the score itself in order to identify such annotations.

In both cases the retrieval system will return a set of (pointers to) multimedia data instances, but the user should receive composite parts of the original presentation in order to understand the result of the query. Browsing has similar problems: once identified a relevant information, the user needs to enlarge the scope and access a more complex information in order to understand it.

A model, such the one presented in Section 4 allows the system to fill the gap between the retrieval of separate components and the presentation of the composite document. One could argue that in many cases the presentation is built as a single file or a set of tightly coupled files, i.e. a set of *tracks*, therefore a reference from the components to the time interval in which they are played is all is needed to identify the relevant context. Even if the presentation is made of separate files integrated by a player, the execution proceeds linearly in time, therefore something similar to a time stamp could solve the problem.

In many cases this is true, but we have assumed as the enclosing scenario of our discussion the World Wide Web, and there are at least two issues that make this simplification not realistic:

- the media are delivered by a server according to a composite document structure which can be distributed in several files, and is known as they are delivered to the client;
- a WWW document usually has links that allow the user to explore its structure in a non linear way, and the user can also interact by backtracking or reloading the documents as they are displayed.

### 6.1 Building a Multimedia Query Answer

Presentation of multimedia information can rely on the static structure and the dynamic synchronization relationships defined by a model like the one described in Section 4, in order to build a query answer which contains all the presentation elements related to the media objects retrieved as results of the query execution.

Answer building is performed in two steps:

- the static structure of the presentation is used to collect all the media items that complete the fragment in which the retrieved object is located;
- the dynamic relationships are used to select the presentation unit to be delivered to the user as the query answer.

With a reference to Figure 5, if the query retrieves a passage of the voice comment in $scene_i$, the answer to be reported to the user is at least that scene, which is dynamically linked to an image $j$, which should also be displayed according to the presentation behavior.

In the same way, after retrieving a text in which a *crescendo* of strings is described it is easy to identify the corresponding part of the audio track in the guide to listening because a relationship links that segment of the audio track to the comment that describes it. In the score analysis, the text comment is related to the same segment of audio track, but also the image of the pertaining score page is related, and the whole presentation fragment made of the music, the score and the text can be delivered and displayed.

If more than one object is selected as a retrieved item, it is easy to check if they are part of a same section or of different sections of a same presentation, or of different presentation, avoiding duplicates. The hierarchical structure of a multimedia document helps to identify the scope of a module, of a clip or of a scene.

Thanks to the same hierarchical structure, it is possible to give the user different levels of access to the resulting presentation. A first attempt is to index the scenes, which are the finest level of access in our model. The user can quickly browse them without being compelled to browse longer sections. In our example, if the user is interested in a particular moment of a musician's life, he or she wants to listen only the scene regarding that year. This level of indexing gives thus the user the possibility to evaluate the accuracy of the retrieve in terms of precision.

The user could then select, according to a relevance analysis, some responses which appear more appropriate than others, and gain access to the higher level of hierarchy, which give a broader context for the retrieved items. In our example, the user could be interested in the whole life of one musician, say Beethoven, identified through a specific information about one of the narrated episodes. The higher level is in this case the whole presentation of a specific instance in the retrieved set, together with animations and images.

In all cases the presence of a synchronization model guarantees that the sections of the presentation which are displayed are coherent.

## 7 Conclusion

Information retrieval in distributed multimedia presentations requires modeling of the relationships among the media objects which build the presentations. We have illustrated a synchronization model which makes a step further, with respect to other models defined in the literature, in considering also user actions such as pausing or stopping when reading such a presentation in a hypermedia context.

In this paper we have analyzed the model in the perspective of information retrieval and browsing, showing how such model can help in building coherent answers to information retrieval queries to multimedia data repositories where complex presentations are stored. Technical issues concerning retrieval of multimedia data and index construction for complex composite documents deserve of course great attention and investigation in order to move from a modeling approach like the one described here to a prototype implementation.

## References

1. A. Celentano, O.Gaggi. Synchronization Model for Hypermedia Document Navigation. *Proceedings of the 2000 ACM Symposium on Applied Computing*, pages 585–591, Como, 2000.
2. A. Celentano, O. Gaggi. Modeling Synchronized Hypermedia Documents. *Technical Report n. 1/2001*, Department of Computer Science, Università Ca' Foscari di Venezia, Italy, January 2001, submitted for publication.
3. Y. Chiaramella. Browsing and Querying: Two Complementary Approaches for Multimedia Information Retrieval. *Proceedings of Hypertext - Information Retrieval - Multimedia'97*, pages 9–26, Dortmund, WA, USA, 1997.
4. I.F. Cruz, W.T. Lucas. A Visual Approach to Multimedia Querying and Presentation. *Proceedings of the Fifth ACM International Conference on Multimedia'97*, pages 109–120, Seattle, WA, USA November 9-13, 1997.
5. A. Del Bimbo. *Visual Information Retrieval.* Morgan Kauffmann,1999
6. Enda Multimedia. *CD-ROM Musica.* Enda Srl Milano, Italy, 1996.
7. L. Hardman. Using the Amsterdam Hypermedia Model for Abstracting Presentation Behavior. In *Electronic Proceedings of the ACM Workshop on Effective Abstractions in Multimedia.* San Francisco, CA, 4 November 1995.

8. L. Hardman, D.C.A. Bulterman and G. van Rossum. The Amsterdam Hypermedia Model: adding time and context to the Dexter Model. *Comm. of the ACM*, 37(2), pages 50–62, 1994.
9. R. J. Miller, O. G. Tsatalos and J. H. Williams. Integrating Hierarchical Navigation and Quering: A User Customizable Solution. In *Electronic Proceedings of the ACM Workshop on Effective Abstractions in Multimedia*. San Francisco, CA, 4 November 1995.
10. S. Hibino, E. A. Rundensteiner. A Visual Multimedia Query Language for Temporal Analysis of Video Data. *Multimedia Database Systems: Design and Implementation Strategies*, Kluwer Academic Publishers, ISBN 0-7923-9712-6, pages 123–159, 1996.
11. S. Hibino, E. A. Rundensteiner. User Interface Evaluation of a Direct Manipulation Temporal Visual Query Language. *Proceedings of the Fifth ACM International Conference on Multimedia'97*, pages 99–107, Seattle, WA, USA November 9-13, 1997.
12. Synchronized Multimedia Working Group of W3C. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification. *W3C Recommendation*, 15 June 1998. `http://www.w3.org/TR/REC-smil`.