A graph based audio-visual document annotation and browsing system

Elöd Egyed-Zsigmond[*], Yannick Prié[*], Alain Mille[**] & Jean-Marie Pinon[*]
Laboratoire d'Ingénierie des Systèmes d'Information
[*]INSA-Lyon, Bat 501
20 AV. EINSTEIN
69621 Villeurbanne Cedex, France
[**]UCB Lyon1, Bat. 710
43, boulevard du 11 Novembre 1918
69622 VILLEURBANNE CEDEX
{elod.egyed-zsigmond, yannick.prie, jean-marie.pinon}@insa-lyon.fr, amille@bat710.univ-lyon1.fr

Abstract

In this paper we present an audiovisual (AV) data annotation and exploitation system, which was developed as a feasibility demonstration of several theoretical models and algorithms. These models provide a graph based AV document annotation model, a rapid search method to exploit these annotations and the representation of connected, distributed graphs in XML. First we present the theoretical background underlining the new approach of our model to the audiovisual document annotation and exploitation. In the second part we describe our demonstration application illustrating the use cases related to the theoretical model. At the end we will present our plans to enhance as well the theoretical models as the application.

1. Introduction

In this paper we present a system which supports the annotation and exploitation of audiovisual (AV) documents. This system was originally created as a feasibility demonstration of several theoretical models, presented later.

Indeed, with the increasing power of computers, AV documents are more and more frequently used, and their size and number is constantly increasing. In order to be able to manage and exploit this resource, which is "rough" on its own[1], we have to provide a method to structure it.

What we mean by exploiting AV documents is the ability to launch complex research requests and to have relevant answers to them. For example, when a television director wants to get on his computer all the film sequences where Jacques Chirac is speaking to somebody, covered by his station, he should be able to formulate easily his request and to have the right movie sequences quickly.

As complex searching directly in the binary AV data is not really tractable today, due to its limits on time consuming and capacity to extract high abstraction level characteristics, it remains necessary to attach explicit descriptors to the binary stream. This allows the efficient exploitation of huge video repositories. In this case the search is mainly done through structured descriptors and not directly in the binary data. According to this point of view, the structuring and exploiting techniques of the descriptors are the two key issues of the research in AV documents. Indeed, the ISO prepares a standard (MPEG7-Req, 1999) which is intended to bring a solution to how video descriptors (of any type) should be organized.

Our system, built on a graph based annotation model (Prié & al., 1998), implements a new algorithm (Prié & al., 2000), enables the creation and exploitation of descriptor structures through a user friendly, graphical interface. It demonstrates the underlying theoretical models capability to support the automatic as well as the manual annotation of AV streams. The representation of the annotations in XML enables the distribution of the annotations over a network as well as their publication.

[1] we consider AV documents as a binary stream of video and audio data

Firstly, we present the research background of the system: a graph based annotation model, called AI-Strata (Prié & al., 98) (Annotation Interconnected Strata), the search techniques in this model combined with the XML representation of data. In the second part the actual framework is described.

2. Requirements for an annotating system

In this section we enumerate a few issues we consider important for a AV document annotation and exploitation system.

The key concepts of a audiovisual annotation system are : the *storage* and *access* to AV documents, in a video on demand like service (Mostefaoui & al.,1999). The *indexing* of AV streams, their *search* and *presentation*. As soon as the use of a AV document exceeds simple visualization, precisely when the document is analyzed, reused or generated, we need a non sequential access to different parts of the document. So we need a *navigation* mechanism following predefined or on need calculated links. The tasks related to the AV document manipulation are complex, the user needs *assistance* in his work.

The *annotation* can be the action of associating a descriptor to an AV document fragment, or the result of this action. The descriptor, often considered as a text describing something in the document, should be able to be any locatable resource (words, text, different structures, images, another AV document, etc.). In order to search the annotated AV documents the descriptors has to be *structured*. A stratified structure is proposed in (Aguierre Smith & al., 1992).

The annotation process can be automatic, or manual. The automatic annotation is usually based on numeric image and sound processing routines extracting mainly low abstraction level characteristics from an AV stream. (like color histograms, movement detection, camera effects, …). The manual annotation consists in visualizing the AV stream and associating descriptors manually to it. This can be assisted by the system (providing adapted views of the available annotating thesaurus, analysing the stream and proposing higher abstraction level descriptors, …).

From a user point of view, the system has to provide an easy way of formulating requests combining formal syntax based queries (like SQL), queries given in a natural language, describing or example based requests. The annotated AV document base should be able to respond to queries including high and low abstraction characteristics. The response to this query should be ideally instant, in any case, given in less than a few minutes. The resulting set of video segments has to be presented in a way that the user can easily find among them what he is really looking for. See for example (Tonomura, 1997) for a study of new AV stream visualizing methods.

3. Theoretical background: the AI-Strata model

As mentioned, the system's theoretical background is the AI-Strata model. It is a general graph based annotation model, specially designed for AV documents and sequential and temporal streams.

In this section we present the basic concepts of the AI-Strata model in a nutshell. For further details please refer to (Prié & al., 1998; Prié, 1999).

The annotations are structured in a graph, whose nodes are of three basic types partitioning the graph in three sets:

- The first set contains the link elements between the annotations and the binary AV data. These nodes, are *audiovisual units* (AVU) materialize the AV segments in the graph. They specify the URI (Universal Resource Identifier) of the stream, as well as two time-codes representing the beginning and the ending instants of the strata on the stream's time scale. In Figure 1 the AVU_1 represents the sequence of the AV stream beginning at time code 2 minutes 30 seconds and ending at 3 minutes 53 seconds on the time scale of the AV stream.

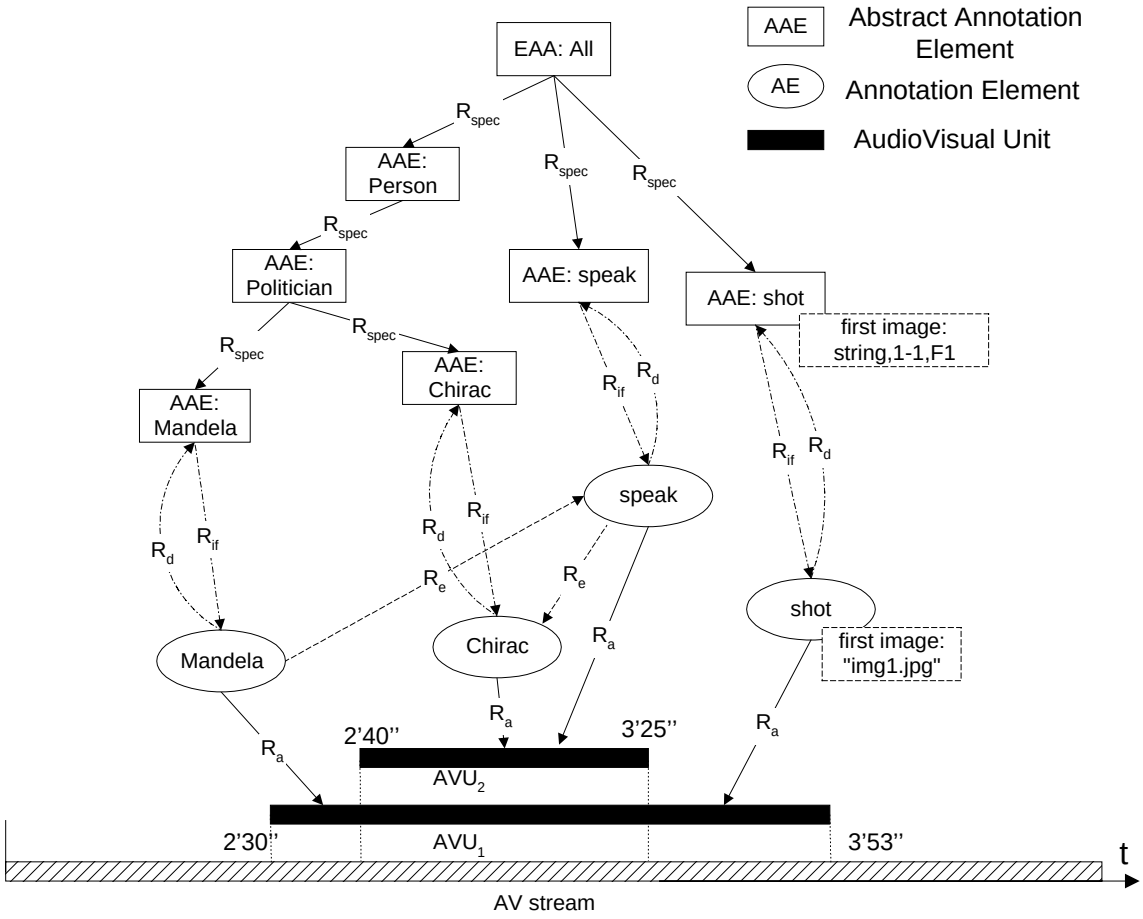

Figure 1: The annotation graph containing the 3 types of AI-STRATA elements: AAE-s forming the thesaurus, AE-s instantiating AAE-s and linking them to AV stream fragments materialized by AVU-s.

- The second set is made up from *annotation elements* (AE) forming the actual descriptors. In fact an AVU exists because an AE is attached to it. An AE is characterized by a *name* and a set of *attribute-value* pairs. Relations, creating semantic contexts can link the AE-s (*elementary relations*, R_e). For

example (Figure 1) the AE <Mandela> can be connected to AE <speak> to emphasize that in this AV segment Mandela is speaking to somebody. The AVU_1 and AVU_2 are in the same context this way, they form annotation interconnected strata (AI-Strata).

- The AE-s derive from a "Knowledge base" containing nodes of a third type, called *abstract annotation elements* (AAE). The AAE-s define the attributes of AE-s. For example in Figure 1, the AAE <shot> defines the atribute *first image* as an URI. This URI points to the image file representing the first image of the shot. The AE <shot> which instantiates this AAE, has an attribute named *first image* with the value: “img1.jpg”. The AAE specifies the minimal and maximal cardinalities for each attribute, as well as a set of comparing functions. The AAE-s form a knowledge base which is at least a thesaurus containing the terms with which AVU-s, and consequently the AV stream, can be annotated. The knowledge base is structured in a specialization/abstraction tree (following *specialization/abstraction relations*, R_{spec}/R_{abs}), but other relations between AAE-s are possible, like those what can be found in a thesaurus (for example: the *see also*, R_{sa}).

These nodes all together form a connected graph, whose labeled edges represent a set of relations, which can connect the elements we mentioned

The exploitation of this structure is based on a sub-graph-matching algorithm described in (Prié & al., 2000). In fact every request is expressed through a so-called Potential Graph, a graph made up from several nodes partially filled. For example the request: “find every video sequence containing Chirac speaking to somebody and having its first image like “example.jpg” according to the comparing function F1” will give the Potential Graph shown in Figure 2. A request is solved by searching the matching of this Potential Graph in the global graph.

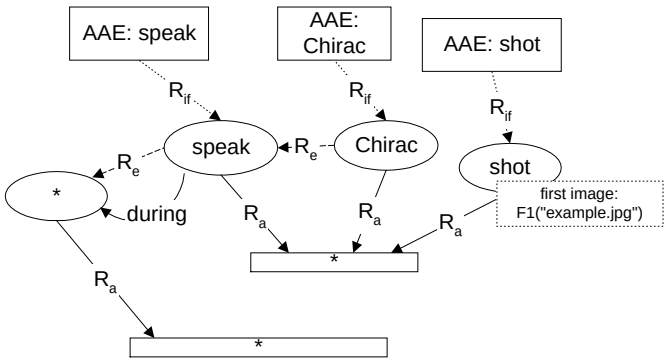

Figure 2: Example of a Potential Graph

The graph elements assigned with a “*” design unspecified AVU-s and AE-s, these are in fact what we search. The isomorphism of this Potential Graph will contain real AVU-s in place of those specified with “*”, and these AVU-s have their first image like “example.jpg” and they present Chirac talking to somebody.

The comparison functions can for instance verify for images color histogram similarity. Of course in some cases, where only one comparison function is specified for an attribute, we can omit the “according to the function F1” from the request to have less complicated formulation. This can be useful for applications created for a wide public, to facilitate the construction of predefined requests, hiding the complexity of potential graphs.

On the other hand comparison functions may require more than one parameter, like threshold values to decide whether an attribute is to be said alike to another.

In potential graphs we can use calculated time relations, like *during, before, after, …* (all of Allen's temporal relations are allowed (Allen, 1983)) between AE-s. These time relations refer to AVU-s annotated by the AE-s. These relations are not explicitly marked in the annotation graph, but calculated during the search process.

The searching algorithm allows to find isomorphisms of a graph in an other one. It is an *any-time* algorithm, which means that it delivers results as soon as they are found, not only when it terminates. The algorithm is based on the multi-propagation from known correspondences (<AAE:Chirac>, <AAE:speak> and <AAE:shot> in the example in Figure 2). This algorithm gives satisfactory results even for bases having many annotations and is controlled by a simple and parametric heuristic.

The Potential Graph in Figure 2 illustrates the capacity of the AI-Strata model to process requests combining high (Chirac, speak) and low (shot, first image) abstraction level descriptors. A large variety of interfaces can be implemented to help users build queries, from a list of predefined form like fill in queries for novice users (hiding the Potential Graph editing), up to graph editing tools for experts to exploit fully the complexity of the querying structure.

As annotations of audiovisual documents are structured in a graph, searching these documents is mainly reduced to search the graph. In fact the major part of the automatic and all of the manual processing of the binary AV data is done while annotating, so during the search we will mainly have to deal with structured representations (graphs and character strings).

In order to serialize the annotation graph we elaborated several XML based models representing graphs(Egyed & al., 1999). The search will be done in these XML based documents.

During the annotation process, the knowledge base may evolve and its size overflows the capacity of what a human can perceive at once. In order to get restricted and yet flexible views of the thesaurus, we can use so called Analysis dimensions (AD).

An AD is a set of AAE-s. The AAE-s can be selected manually or automatically by *designation methods*. The *designation methods* are materialized by Potential Graphs searching AAE-s dynamically. So even in an evolving thesaurus the view defined by the AD is complete comparing to the defined constraints. For example we could build a Potential Graph to select the derived AAE-s from <AAE:Politician>, in this case even if we add or remove politicians, the AD will contain them all.

AD-s can be created in order to group together terms for a specific domain, ore used by a certain community to annotate AV documents. On the other hand this feature helps users to choose terms for build queries, knowing the AD-s used to annotate the AV documents.

4. Framework description and use cases

The system presented here implements some aspects of the AI-Strata model. It is developed to illustrate the large utility of this theoretical model, and to serve as experimental framework in elaborating case-based user helping and assisting methods on multimedia annotation and exploitation.

In a first time we say a few words about the systems technical properties, after which we will provide a functional description presenting in the same time the AI-Strata based annotation use cases.

The application is developed in MS Visual C++[2], based on XML[3] and LEDA[4] for the graph representation. It provides a graphical user interface Figure 3 for manipulating AI-Strata based audiovisual document annotations.

The internal data structure is based on the LEDA Graph structure and the XML DOM, implementing the nodes of the graph in XML. This structure consists of a set of objects corresponding to the AI-Strata elements (AVU, AE, AAE, Potential Graph, Global Graph, …), represented in XML and the manipulating methods and search algorithms.

Through the description of the annotation and browsing system, we will present the different ways of putting down annotations, editing requests, visualizing their results and navigating in the graph.

First of all, we have to create the knowledge base, the thesaurus that will serve to create annotations. The sub-graph forming the knowledge base can be projected on a tree structure following the specialization/abstraction relations. In fact we suppose that every new AAE is the specialization of an already existing one (AAE). The application provides standard tree manipulating functions to create and manage this thesaurus as well as the other relations between the AAE-s.

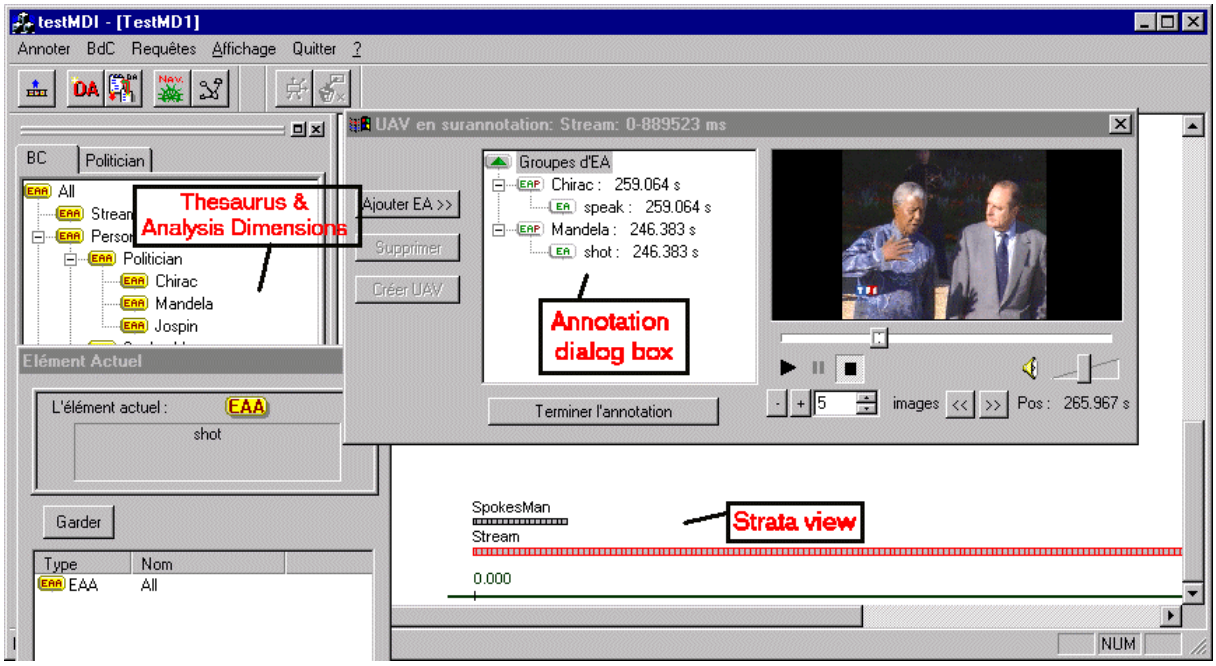

Figure 3: The application's main interface. A view of the thesaurus and selected AD-s at left, the manual annotation dialog box and Strata view at the right side. The Strata view appeals Potential Graphs to select AVU-s temporally included.

Once we have a knowledge base, we can begin putting down the annotations. The annotation consists basically in selecting terms (AAE-s) creating annotation elements (AE-s) from them, and assigning them to AVU-s. This can be done manually or automatically.

The manual annotation begins with the selection of some Analysis Dimensions (AD). A view of the thesaurus and the selected AD-s is presented as the tree controls in the left side of the image in Figure 3.

[2] <http://msdn.microsoft.com/visualc/>
[3] Extensible Markup Language < http://www.w3.org/XML/ >
[4] General C++ data type and algorithm library < http://www.mpi-sb.mpg.de/LEDA/leda.html >

Once the AD selected, we pick one or several AAE-s from it. The annotation continues with the specification of an audiovisual stream. If the stream has already been annotated, a view of the AVU-s resulting this previous annotation is shown, if not, an AVU referencing the whole stream is created. This way every stream is annotated at least once by the <AAE :stream>. We can than add new annotations to a selected AVU, or we can create new AVU-s, sub-segments of the selected one, by marking their beginning and ending time codes and selecting the AAE-s to annotate it. In the right side of Figure 3 the AVU-s in creation are presented: one, having <AE:Chirac> as primary descriptor but annotated by <AE:speak> too, the other described by <AE:Mandela> and <AE:shot>.
The automatic annotation is based on “annotation assistants”. These are external programs (typically image processing tools) which, after a binary analyze of the AV document, create annotation-graph segments and bind them to the global graph. For example an annotation assistant could segment a video in shots detecting the cuts, then create the AVU-s corresponding to each shot and link them with the AAE <shot>, can detect movements, camera effects, etc .

The system implements two interfaces to navigate in the annotation graph (Figure 4). A dialog based and a graphical one.

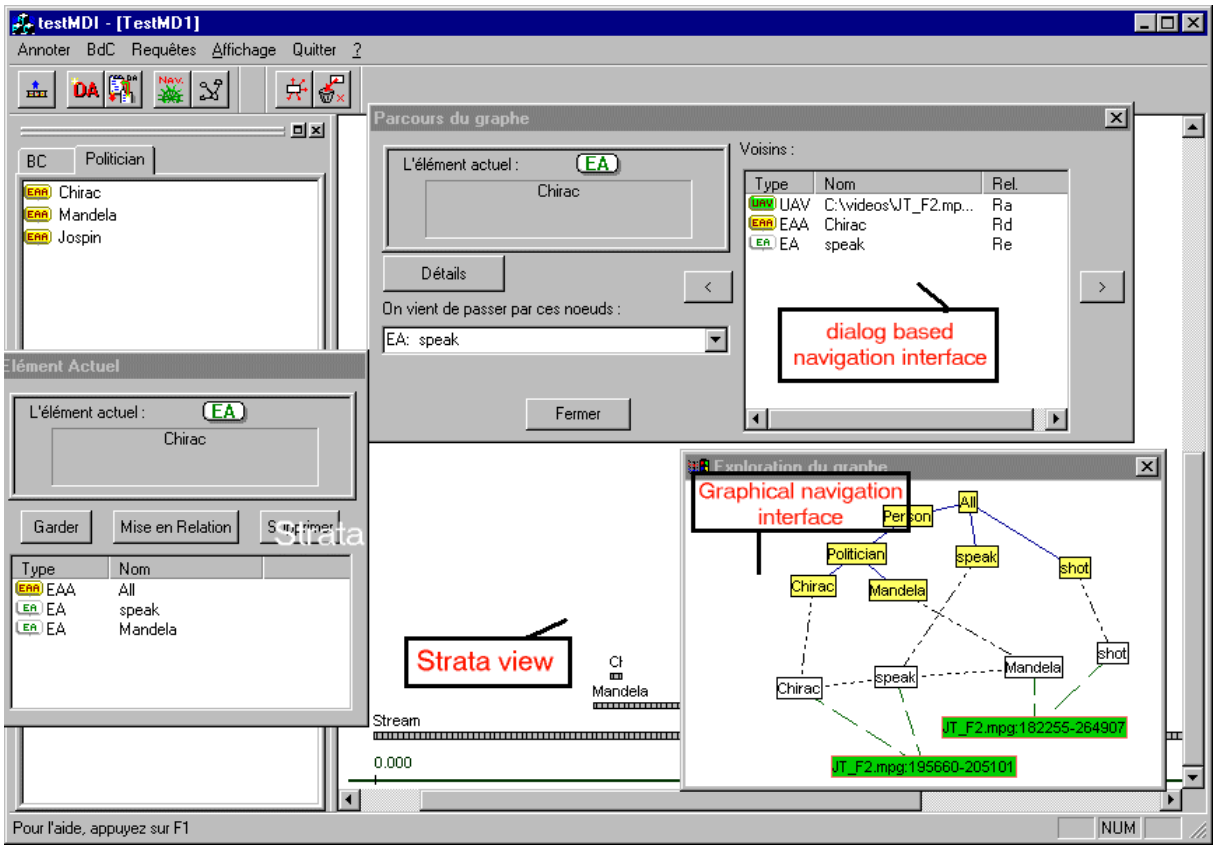

Figure 4: Navigation interfaces: dialog based (upper right) and graphical (bottom right)

The edition of requests begins with a phase of potential graph creation then the request can be launched and the resulting matching sub-graphs can be visualized. The edition of request graphs can be avoided by selecting one of the pre-defined requests. The Strata-view of AVU-s () uses view Potential Graphs to select AVU-s temporally included

Figure 5 represents a basic task model for the annotation as a decomposition diagram as described before. For example the task *Assign the annotation to the stream* in Figure 5 can be decomposed in the following sequential sub tasks: *Create an AVU* or *Select an existing AVU*, then *Link the selected AAE to the AVU creating the AE* and finally *Fill the attributes of the AE*.

Formalizing user tasks enables to capitalize user actions in a structured form and to compare them to other scenarios in order to reuse its experience. The formal representation of the tasks will serve the construction of intelligent case based user help systems (Prié & al. 1999).

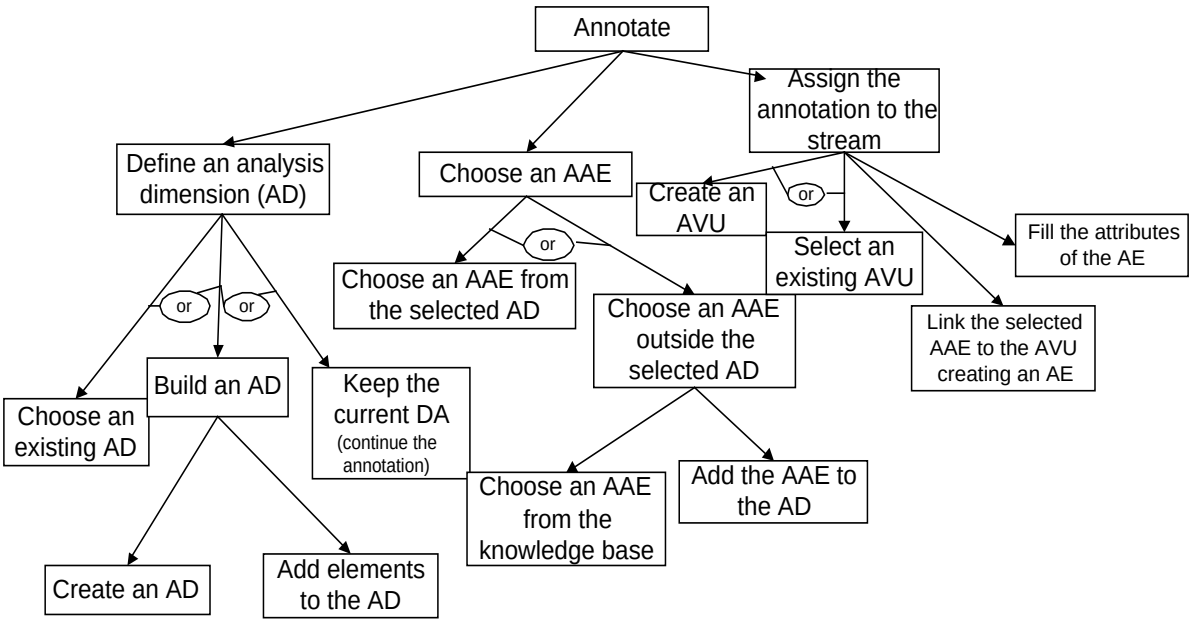

Figure 5: The basic formal model of the manual annotation task

5. Conclusion and future work

The system presented in this paper is developed as a part of the RECIS[5] project, in which we are partners with the CNET[6] and the INRIA[7]. It is a feasibility demonstration of the AI-Strata model implementing an original, graph based organization of descriptors attached to audiovisual documents. It allows the annotation of this type of data (AV), and the construction and execution of search requests. Our attention in focuses now near the on the capitalization of user experience in order to help him in the annotation and exploitation process.

In order to increase the relevance of the answers, we will implement the *description scheme* (DS). This concept is materialized as extension of the PG, which may have AD-s as nodes. An AV document can be annotated following a DS, so a minimal structure of its annotation graph is known. In fact using a DS to annotate allows us not only to know which DA-s were used to describe the document but also to have information about how the annotations are structured. An AV document can be annotated following a given DS. Knowing the DS we can construct thus more precise requests and consequently obtain better results.

The XML representation framework for graphs exists and currently the graph nodes representing the AI-Strata elements are implemented in XML. We will extend our data structure in order to handle the complete annotation graph and other structures in XML, facilitating the distribution and the portability of the annotations.

[5] RNRT project <http://www.telecom.gouv.fr/rnrt/projets/precis.htm>
[6] <http://www.cnet.francetelecom.fr/>
[7] <http://www-rocq.inria.fr/>

We will also enrich the annotation assistant palette, including other image and sound processing routines developed by our partners.

Another interesting research field remains the elaboration of user interfaces to visualize audiovisual data for annotation, to show the already existing annotations, to present search results in a well perceivable shape, to help the user build queries, etc.

6. Bibliographical References

Aguierre Smith, T. G. & Davenport, G..(1992) *The stratification system, a design environment for random access video.* In Proc. Network and Operating System Support for Digital Audio and Video - 3rd International Workshop, pages 250-261, San Diego, CA, 1992.

Allen J. (1983): *Maintaining temporal knowledge about temporal intervals.*Communications of the ACM, 1983, vol. 26, n°11, pp.832-843.

Egyed, E. & Prié, Y. & Mille, A. & Pinon, J.-M.(1999) : *Représenter un graphe d'annotations dedocuments multimédia réparti sur plusieurs sites à l'aide d'un langage dérivé de XML*, Journée jeunes chercheursGDR-PRC I3, the 9th september 1999, Tours, France (in French)

Mostefaoui, A. & Perez, C. & Brunie, L. *Serveur de séquences audiovisuelles parallèle sur réseau haut débit: concepts et expérimentations.* 11th French Meeting on Architecture & Systems Parallelism, june 1999, pp. 127-132. (in French)

MPEG7-Req (1999) *Mpeg-7 requirements v.10 iso/iec jtc1/sc29/wg11, N2996,* October 1999 Melbourne Australia, (online) < http://drogo.cselt.stet.it/mpeg/public/mpeg-7_requirements.zip>

Prié, Y. & Limane, T. & Mille, A.(2000) *Isomorphisme de sous-graphe pour la recherche d'information audiovisuelle contextuelle*, 12ème congrès Reconnaissance de Formes et Intelligence Artificielle, RFIA2000, Paris, feb. 2000, vol. I, pp. 277-286.

Prié, Y.(1999) *Modélisation de documents audiovisuels en Strates Interconnectées par les annotations pour l'exploitation contextuelle*, Computer Science PhD dissertation, INSA de Lyon, dec.1999, 270 p.

Prié, Y. & Mille, A. & Pinon, J.-M. (1999) *Modèle d'utilisation et modèles de tâches pour l'assistance à l'utilisateur basée sur l'expérience : le cas d'un système d'information audiovisuelle*, Ingénierie des Connaissances 1999, Palaiseau, jun. 1999, pp. 21-30. (in French)

Prié, Y. & Mille, A. & Pinon, J.-M. (1998) AI-STRATA*: A User-centered Model for Content-based description and Retrieval of Audiovisual Sequences* First International Conference on Advanced Multimedia Content Processing (AMCP'98), Osaka, Japan, November 1998, LNCS 1554, Springer-Verlag

Tonomura, Y.(1997) *Handbook of Multimedia Information Management, chapter Multimedia Interfaces* -- Multimedia Content Indication, pages 189-209.Prentice-Hall, Upper Saddle River, New Jersey, 1997..

Acknowledgements

We would like to thank the INA <www.ina.fr> for the provided AV streams