Managing Full-indexed Audiovisual Documents: a New Perspective for the Humanities

Gwendal Auffret (gauffret@ina.fr)
INA, DRP, 4 av. de l'Europe, F-94366 Bry/Marne Cedex

Yannick Prié* (yannick.prie@insa-lyon.fr)
LISI 502, INSA-Lyon, F-69621 Villeurbanne Cedex

Abstract. The digitization of library documents and archives increasingly extends to audiovisual (AV) document repositories. As a consequence, new computer-aided techniques are being devised, providing opportunities for new uses of AV documents. As scholars work mainly by reading, annotating, reusing, and producing documents they are directly concerned by these changes. The first part of this article describes AV document use in the humanities, as well as the current and future influence computers might have on evolving practices. After establishing that "full-indexing" (indexing of the content for random access to any segment of an AV document) is a necessary condition if scholars are to develop new practices in using AV material, we will focus on the specific problems raised by AV indexing as opposed to text indexing, followed by a discussion of related AV indexing projects as well as standardization issues. The third part will propose a representation model for the description of AV material (AI-Strata) and an exchange format of AV annotations (AEDI), based on a free segmentation approach. An example of annotation is also provided. The last part is devoted to a discussion regarding potential long-term influences of digital AV indexing techniques on scholarly uses of AV documents.

Keywords: Scholarly use of audio-visual documents, Indexing, Modeling, Standardization, AI-Strata, AEDI

1. Introduction

The everyday activity of scholars is mostly based on working with documents (i.e., consulting, reading, analyzing and producing). Documents are entities organizing pieces of information into an intentional structure (André et al., 1989; Furuta, 1997) and they may represent valuable and sometimes rare information. There are textual, e.g., an ancient manuscript or an essay, as well as audiovisual (AV) documents including cinema, TV, radio and musical documents such as Verdi's "La Traviata", or yesterday's BBC 6 o'clock news program.

Scholars perform on documents what we call an active reading, as opposed to the traditional passive reading done when reading a novel

* This work is partially supported by France Télécom (through CNET/CCETT), research contract N° 96 ME 17.

for leisure or when watching TV. Active reading is a thorough analysis of the document carried out largely through annotation. Annotation is the reader's writing of his/her interpretation of the source document leading to the production of new documents (Virbel, 1995).

As part of our global cultural heritage, documents are stored in personal or institutional libraries, which are, unfortunately, often out of reach for technical or economic reasons. Nevertheless, in recent years large-scale digitization projects and the development of computer networks have allowed computer-based remote access to many document repositories. Libraries, archives, and museums throughout the world have gone on-line, digitizing their documents and even providing new tools allowing scholars to work efficiently on this digital material (see the computer-aided reading environment of the French National Library (Chahuneau et al., 1992)). Most of these projects make use of hypertext concepts defined by V. Bush (Bush, 1945), which have been widely implemented on the World Wide Web. The long-awaited worldwide digital library is even announced for the near future (Fox and Marchionini, 1998; Paepcke et al., 1998). Until now, these projects have been mostly devoted to text and still-image management, but, as technology evolves, more and more AV material is being digitized or even directly produced using digital cameras, professional editing tools, and specific databases.

Nevertheless, we will show that efficient use of such material requires more than digitization: some means of random access to any piece of a digital document[1] are needed. Moreover, truly working with AV documents in an academic context implies not only accessing but also annotating and recomposing parts of documents. Hence, one of the key questions raised by the digitization of AV material is how to define, process, and retrieve the information necessary to perform computations on such segments. We claim that this can only be achieved by a formal representation of the content of AV documents, here called *full-indexing*, and we will argue that the management of full-indexed audiovisual documents opens new perspectives for future academic studies.

This article will first describe current digital AV material user practices in the humanities, the limitations and potential changes implied by new digital environments, and then will present the specificity of AV indexing and ongoing AV indexing research projects. We will show how these projects do not take into account the practices described and we will argue for the creation of new models. A presentation of

[1] For instance, be able to retrieve and watch all close-ups of Ingrid Bergman in *Casablanca* without having to download and pay for the entire document.

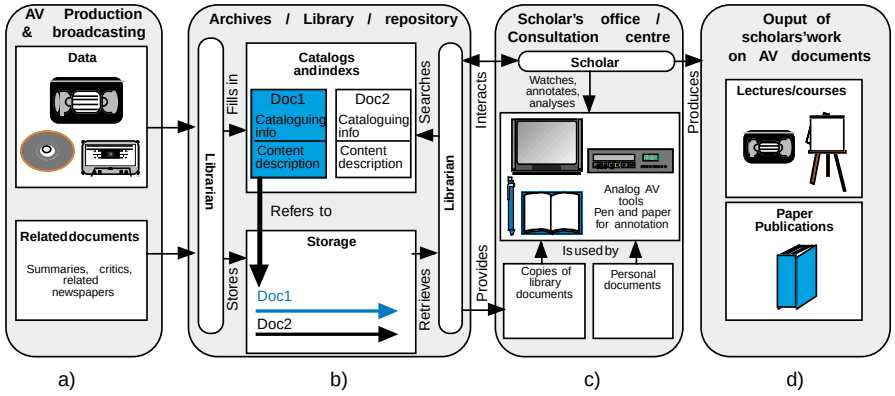


Figure 1. Traditional scholars' usages of AV material

our proposal for the indexing and the annotation of AV documents will follow as well as the document representation format used for the exchange of AV document descriptions. Finally, we will conclude after discussing the long term influences of indexing on scholarly uses of AV documents.

2. Current and future use of AV material in the humanities

Although they represent very valuable information, AV documents remain under-used in the humanities. After describing current uses and their cultural, legal, and technical limitations, we will show how recent technology improvements provide a framework for a more widespread use of AV materials in humanities research.

2.1. Traditional uses of AV material in the humanities

What we call traditional uses of AV documents in the humanities are those which do not take advantage of recent progress in information technology. The environment of such uses is summarized in Figure 1.

2.1.1. *Access to document repositories*

Today, AV documents are still produced and stored for the most part in analogue format. They are documented by hand by librarians in archiving institutions or TV networks. AV documents mainly consist of programs broadcasted as single production units by TV or radio networks. They are traditionally described following cataloguing guidelines (author, title, broadcasting channel, duration, etc.) similar to those defined in (Harrison, 1991) and are sometimes indexed using a controlled

vocabulary. These descriptions are stored on paper or in proprietary textual databases and are, of course, independent of the AV data itself. Catalogues and indexes are accessed by librarians to respond to scholars' needs or, sometimes, by scholars themselves, as in the French Legal Deposit library (Inathèque de France). Once the searched document is found and ordered, a copy is produced and provided to the academic.

In recent years, computer technology has improved to such an extent that free textual descriptions can be used for searching, as in any textual information retrieval system, and networked access to catalogues and indexes is now possible. Though important, these changes have not really modified well-established practices based on manual textual indexing of a given document considered as a whole.

2.1.2. *Working with AV documents*

We introduce here a basic typology of user practices related to AV documents in the humanities[2]. According to this typology, AV documents are mostly used by academics in the following contexts:

AV documents as pedagogical tools. It is often said that a picture is worth a thousand words. As with still images, moving image documents are used in courses, lectures, or conferences for illustrating a subject, or as specific pedagogical tools (Jacquinot, 1985) (see Figure 1d). Most of the time, documents come directly from an archive (i.e., without further editing) and are stored on tapes or film reels. VCR functions (stop, start, fast forward, etc.) can be used to provide some interaction with the public during the projection.

AV documents as a testimony of the past. Scholars in general, and historians in particular, consider AV documents as valuable testimonies of the past or as a relevant mirror of our societies[3] (Ferro and Planchais, 1997). Analyses are almost exclusively carried out with pen and paper and imply intensive use of the VCR.

AV documents as a work of art. Many AV documents are analyzed by critics as works of art. They are the testimony of an author's creativity, of his/her aesthetic and thematic choices (Bordwell, 1993). The context of user practice for such analyses is equivalent to the one described above.

[2] This typology should not be regarded as the result of an ethno-methodological analysis of scholars' work with AV documents but as a fairly large framework derived from observations and from the literature. A real large-scale study of these practices, conducted by specialists, would be of great interest for its improvement.

[3] Indeed, the way something has been broadcast as part of a news program is often as relevant as what was actually filmed.

AV documents as personal notes. For many humanities scientists such as anthropologists, sociologists, psychologists, or education specialists, recording events using a camera can be of great interest for an a posteriori analysis of human behavior. In this type of context, AV documents are mostly a means of memorizing past real-life facts. They are the scientist's notes and Aguierre-Smith (Aguierre-Smith, 1992) shows that they need to be organized and reworked, which requires some means of editing and manipulation.

AV documents as communication acts. Semiotics of AV documents — as it has been defined by C. Metz (Metz, 1968; Metz, 1972) for the cinema and then extended to other AV streams such as TV (Jost, 1992; Jacquinot, 1977) — studies AV documents as communication acts. Semioticians analyze how the AV medium is used for conveying information and emotions. Their main questions are, How does a film make sense? What are the elements of its composition that are used for interpretation? To answer these questions, scholars have to study corpora, perform discourse analysis, evaluate the production and reception conditions, and relate the form and the meaning of the AV material using some model of AV communication strategies.

Of course, each of these examples of research on audiovisual documents would need a specific thorough analysis in order to define, task by task, the requirements in terms of technical devices. However, we can already note that these very different user contexts share some common constraints. Indeed, for each of them, the users want to be able to perform basic active reading functions such as stopping the document stream whenever they want, moving backwards or forwards, selecting and directly accessing any segment of a document, annotating these segments according to specific analysis purposes, and finally browsing and/or working on the document content using these annotations. These functions are the basics of scholars' work.

2.1.3. *Output of scholars' work on AV documents*
As shown in Figure 1d, scholars working with AV documents publish academic paper articles and books in which AV documents are referenced, just as any other source, and quoted by still images. The source documents may also be projected as an illustration when presenting the result of the work in a conference. More rarely, the scholar's work may be the basis for editing a new AV document (for example, a documentary on the research done). However, directing an AV production remains a time-consuming and complicated job that often requires the participation of AV professionals to help the scholar.

2.2. Current limitations

Compared to other types of documents, scholarly uses of AV documents are rare and limited. Indeed, many cultural, legal, and technical obstacles remain to be cleared.

2.2.1. *Cultural and legal limitations*

Even though the AV cultural heritage is increasingly taken into account as a resource by scholars, it remains undervalued. There are many cultural reasons for this: books have been traditionally considered as the only "real" cultural artifacts, whereas mass media in general, and TV and radio in particular, are regarded as popular means of leisure, without any real cultural value. These cultural assumptions partly explain why there are still few centralized AV archives and legal deposit centers: for instance the Inathèque de France, the French legal deposit, was created only in 1995 (Denel et al., 1994). Before this date, from an academic and a legal point of view, AV documents produced by French TV and radio broadcasters were not officially considered as part of the cultural heritage.

Moreover, contrary to text and still images, since AV documents involve long and expensive production processes, numerous intellectual property rights are attached to them and their cost is often too high for research laboratories.

2.2.2. *Technical limitations*

Traditional tools used to interact with AV documents (film reels and projectors, TV, VCR, video tapes, audio CDs, etc.) impose limitations on their efficient use in an academic context. Indeed, it appears that the access to such material remains difficult: AV documents are often accessible only in archive libraries. Obtaining a copy is complex — catalogues are often hard to search and finding relevant documents using catalogues and indexes requires the help of a librarian (see Figure 1b) — and expensive. Moreover, the medium used might be fragile and/or damaged and copying it often requires specific high skilled professionals. Even when an AV document can be accessed and used, means for visualizing it are frequently based on the projection paradigm: an AV document is played on a screen (whether it be a TV or a movie screen) for a certain continuous amount of time. The introduction of the VCR has allowed users to stop, go back and forth in a document, but interaction opportunities remain limited. Finally, tools for editing and/or annotating AV material are rarely available outside of the professional production world. It is difficult for scholars to work with AV

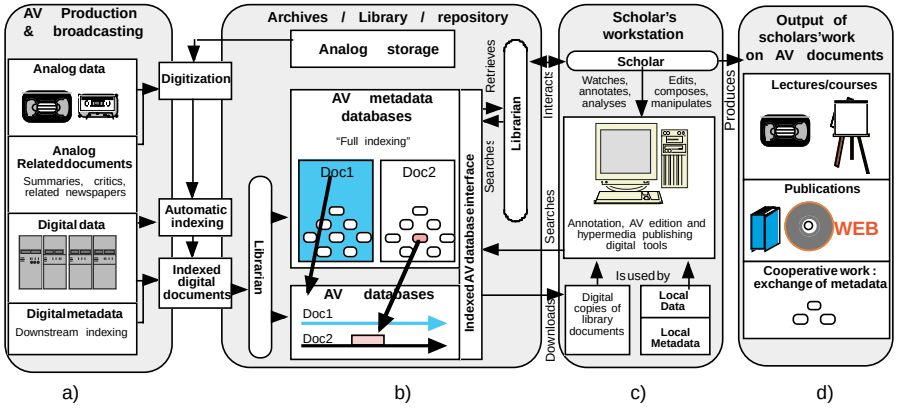


Figure 2. Computers get involved

documents as they do with text documents, i.e., annotate them, classify them, transform their structure, insert quotations, etc.

2.3. Computers get involved

For many years, the digitization of AV documents did not seem to provide any economically acceptable technical solution to the limitations described above. Indeed, the digital AV documents in databases are huge in size and require adapted compression. Even compressed, one hour of digital video requires gigabytes of storage (Yeo and Yeung, 1997). As a consequence, network transmission of AV data is confined by bandwidth limitations. Moreover, despite the development of compression norms such as MPEG (Mitchell et al., 1996), there has been a lack of sufficiently stable standards for storage.

However, recent technological development as well as the convergence of the worlds of AV production, computers, and telecommunications has encouraged the development of new environments for access to and use of AV documents by scholars. Indeed, after library catalogues and indexes, documents themselves are digitized and new computer-based environments are devised for interacting with them.

2.3.1. *From catalogues to document parts: towards full-indexing*

As described in the previous section, AV document repositories traditionally provide catalogues and global indexes (see Figure 1b): a document is referred to as a whole, there is no direct reference to pieces of it. In past years, many such indexes and catalogues have been digitized or directly produced in digital format in libraries and archive databases. This information is more and more often available on networks: users can query them and find references to the documents they

want to consult. Then the document can be ordered. If it is available in digital form, it is possible to download it from a server (see Figure 2b)[4].

However, downloading a two-hour document when only a two-minute extract is needed is inefficient and expensive. As a consequence, the real challenge of AV data digitization is the access not only to documents but also to relevant segments of these documents. Formats such as the Digital Video Disk (DVD) include functions allowing direct access to segments of the video content in the same way as one selects a specific song when playing an audio CD on a stereo. Nevertheless, such access remains limited to a fairly basic table of contents feature.

To go beyond these limitations, it is necessary to define a model for the representation of the content of AV documents. Contrary to digital texts, it is impossible to rely on fully automatic methods to create indexes that would allow a full-image search. If digitizing a text is representing it as meaningful units such as letters and words which can then be used as symbols[5] for computer processing, on the contrary, digitizing AV material does not provide any meaningful symbolic elements (i.e., elements available as symbols for machine computation and human interpretation). A pixel or even a group of pixels do not make sense by themselves and more information is needed to interpret the images in terms of objects or units that are relevant for a certain task. As a consequence, access to parts of documents implies the use of an intermediate representation, which has to be formalized to allow easy computer-aided random access.

We define full-indexing as the process of representing the content of an AV document by a structure of symbols allowing random access to any segment of this document[6]. This symbolic structure is the formalization of an interpretation of an intentional structure of the document. We claim that future full-indexed repositories should

[4] Video-on-demand projects such as the French "Services and Programs Bank" (BPS) (see `http://www.lacinquieme.fr/`), for instance, have been created for developing the use of AV material in high schools. The system allows teachers to order films through an internet-based catalogue including short extracts or previews. Selected AV documents are then downloaded onto a PC using a satellite connection and then played or projected to students.

[5] In this article, we will use the term "symbol" to refer to formal symbols used for computation. This notion is not connected with the Peircean notion of symbol. The same applies to the term "icon", which will be used to refer to small images used as pictograms in computer interfaces and not to the Peircean notion of icon.

[6] The notion of full-indexing can be considered as equivalent to the notion of encoding traditionally used for digital texts. We decided to use another term, however, since in the multimedia community, the encoding of an AV document refers to its compression mode (e.g., MPEG-1 or MPEG-2).

provide new means of working with AV documents and thus provide adequate contexts for the development of new scholarly uses.

As illustrated in Figure 2, new technical opportunities are now offered by full-indexing. It allows efficient access to remote document repositories through distance browsing of the content indexes. Users can select, download and view only segments of documents that are relevant for their needs. Moreover, at the document level, non-linear navigation such as the one performed in hypertext systems (Rosenberg, 96; Landow, 1997) becomes possible (Sawhney et al., 1997): users can browse the document from the index to the data (e.g., find all the shots related to the US President in a news program), from the data to the index (see the keywords attached to a specific shot), and from the index to the index (browse all the keywords conceptually related to each other by using formalized knowledge representation). This also provides automatic creation of new 'views' within AV documents, such as summaries and tables of contents adapted to the ongoing task.

Moreover, the representation model used for indexing can also be used to provide user-oriented annotation facilities. Scholars working with AV sources on specific tools[7] can analyze a document by directly attaching annotations to it and then navigate through this document using their own annotations as entry points.

Finally, individual editing and electronic publishing of new AV documents based on archive documents is facilitated by relevant indexing. For instance, it allows a teacher using AV material not only to show videos to the class but also to efficiently select excerpts from many relevant documents and to compose and edit them in order to produce a new document that meets the needs of the lecture. Relatively easily accessible AV or multimedia editing tools such as Premiere[8] or Director[9] are already available and user-friendly authoring environments should flourish in the next few years. If these environments include full-indexing facilities, they will allow AV documents to become "active documents" (Quint and Vatton, 1994), providing new types of electronic publications including direct AV data[10], as shown on Figure 2d. (Auffret et al., 1999; Auffret and Bachimont, 1999).

[7] Such as the French legal deposit's AV Reading Workstation named *SLAV* (Station de Lecture Audio-Visuelle), see Figure 6.

[8] See `http://www.adobe.com`.

[9] See `http://www.macromedia.com`.

[10] This process would be equivalent to what happened in the domain of digital audio documents with the MIDI format. This format, in existence for many years now, allows any computer user to create new documents using his/her own home studio, leading to new kinds of music and radio programs. Video documents might well develop along the same lines with computer users accessing on-line AV document repositories to use extracts for building their own hypermedia documents.

3. Indexing digital AV documents

The developments cited above remain only potential changes in the technological state of the art. They will evolve only if relevant and efficient access to pieces of documents (i.e., fully indexed AV repositories as opposed to document-as-a-whole indexing) are provided to users.

One question remains, however. What type of full-indexing should be provided in order to encourage an expansion of AV document use by humanities scholars?

In this section, we will discuss the notions of index and metadata applied to AV data and subsequently present current ongoing research projects to show that these projects do not take into account the user practices described above. We will then suggest a change in approach.

3.1. Indexes, metadata, and AV data

Our definition of indexing is fairly broad: Indexing a content means reformulating (by any means) this content in a given semiotic form to make use of this content for a given task. In libraries, this definition includes not only document cataloguing information (such as author's name or title) but also content description (such as keywords). These indexes are suited to a librarian's use of documents so as to provide access to them.

For nontextual documents, keywords are traditionally used as a means of retrieval, but new extraction methods allow other types of indexes to be defined (e.g., images or texture information). Study of nontextual document indexing indeed shows that every imaginable description characteristic can a priori be considered as an index for any content, and conversely that it is how the content will be used that guides the choice of the indexes. Such a change is noticeable in the computer world, where the notion of metadata seeks any datum that will allow using another datum[11].

In the digital audiovisual context, metadata can be classified according to many criteria, among others we can consider:

[11] Data corresponds to documents in their most basic form in computer memories, for instance MS Word or Postscript files for text documents, GIF or MPEG files for images or videos. Metadata was at first supplementary information about files ("Data about data"), such as date of modification or size. With the development of multimedia and web documents, the notion of metadata has been widely extended 1) to keywords and other descriptors for document retrieval and 2) to any descriptor for exploiting documents. As a consequence, the standardization of metadata is a very important current research issue (see subsection 3.3), while the notion of metadata itself is evolving (Sheth and Klas, 1998) to any (useful) supplementary information that can be attached to data, and finally appears to be similar to the definition of indexes.

- their origin: a video recording, a digitized document, the digital editing device (which can provide information on shot boundaries[12]), the digital camera itself (light conditions), the elements provided with the AV documents such as Teletext, TV magazines, studies;

- how they are elaborated: automatic calculation (color rate, duration), semi-automatic (the trajectory of an image object pointed out by hand) or manual (keywords, screenplay, etc.);

- their degree of subjectivity: from low (a shot cut, a texture analysis) to high (the description of a feeling, the quality of an actor);

- their temporal application domain: atemporal (the name of a movie concerns the entire movie) to highly temporal (the movement of a character).

As stated above, full-indexing of AV content (i.e., the creation of relevant and efficient metadata) is the key to future AV reading and writing environments. Many current research projects focus on metadata creation.

3.2. Current research projects

Until recently, research on full-indexing of AV content has been carried out for the most part in the signal processing community. Indeed, researchers from this field have been interested in using still and moving image processing for the extraction of features that can be used for retrieval. Color histograms, textures, shapes, movements, and shot cut detection led, generally, to indexing at the shot level using signal-based features, hardly directly usable by humans. For instance, a shot can be indexed with its color histogram and main texture measures calculated on a still image extracted from its frames and considered as representative of the whole shot. A specific metric is associated with automatically extracted indexes and calculated similarity measures can be used in sample-based or sketch-based queries: the computer analyzes the sample in terms of features and tries to match it in a metadata database. Browsing and video playing capabilities are then added to such systems (Zhang et al., 1995; Taniguchi et al., 1995; Gupta et al., 1997). The main problem of such an approach is to define the features to be extracted. In fact, many of the projects described above are often

[12] A shot is a stream of contiguous frames continuously recorded by a single camera.

ad-hoc ones: their models of AV content are frequently driven by the available technology more than by real-life users' needs.

Many of these systems have been presented in the last few years, and have even led to industrial products [13] but as has been acknowledged elsewhere, simple feature-based indexing is neither sufficient nor practical for real research in huge image and video databases, enhancements have been proposed, for instance learning from users' interactions (Minka, 1996) or use relevance feedback (Nastar et al., 1998). On the other hand, some projects have always taken into account the need for textual metadata along with image processing and features. The Informedia system (Christel, 1995) uses automatic processing of the audio stream to search for sequences associated with words. Yeo and Young (Yeo and Yeung, 1997) propose to describe documents with classical text description, while a more refined search should be based on similarity features. WebSEEK (Chang et al., 1997) considers hierarchical ontologies of terms to describe documents.

At the same time, database researchers have taken an interest in video and proposed several adaptations from classical database models to cope with audiovisual and multimedia documents (Hjesvold and Midtstraum, 1994; Oomoto and Tanaka, 1993), but there is still a great deal of work to be done, as Grosky recognizes (Grosky, 1997).

Nevertheless, current trends in automatic processing of AV material are concerned with scene extraction[14], not an easy task, as scenes can be highly subjective units, and because semantic considerations and domain knowledge have to be taken into account (Yeung et al., 1996). Synthetic presentation of AV material is also an active research field, both to present shots of AV documents (extraction of images, construction of summary images such as Salient stills (Massey and Bender, 1996), trying to represent movements as temporal signatures) and to organize them into document views.

There is clearly a need for international standards which would serve the purpose of organizing these developments. The next section will present the standardization initiatives currently addressing AV indexing issues.

3.3. Standardization issues

Standardizing the way AV documents can be fully indexed is a major challenge and an unequivocal prerequisite to the efficient exchange and manipulation of AV documents. Moreover, it is a fundamental economic

[13] See QBIC (Flickner et al., 1995) or VIRAGE `http://www.virage.com/`.

[14] Scene: collection of one or more adjoining shots, that has the characteristic of perceptual continuity and semantic homogeneity.

issue for the companies emerging from the convergence of the telecommunications, broadcasting, and computer industries (Internet-enabled TV sets, multimedia computers, video on demand, cable TV, etc.). New tools indexing AV content from the source and integrating this indexing into the stream (e.g., shot cuts) are also being devised. Providing a way to extract and/or represent this information up-stream would save a great deal of effort and money currently spent in downstream analysis of video content.

As a consequence, many initiatives such as the EBU/SMPTE task force[15], MPEG-7[16], W3C standards (XML, RDF, SMIL, etc.)[17] and ISAN [18] are attempting to define the future norm for the full-indexing of AV documents.

Most of the work is currently carried out on defining representation formats and relevant descriptors for search and retrieval purposes, but other concerns, such as the exchange of indexing among members of a community or the conversion of indexed documents for electronic publishing purposes are emerging. One could say that the AV community is currently going through a process quite similar to the one that led text specialists to SGML[19] more than 10 years ago, and then to the TEI[20].

However, standardizing representations of AV content raises many specific thorny issues such as defining what the content of an AV document actually is. Indeed, this content of an AV document is potentially infinite: it always depends on the user's interpretation, which, in turn,

[15] The joint European Broadcasters' Union / Society of Motion Picture & Television Engineers Task Force is an initiative which aims at standardizing the exchange of television program material as bitstreams. It has published an ontology of AV document content. See `http://www.ebu.ch/pmc_es_tf.html`.

[16] MPEG, Moving Pictures Experts Group, is the body within ISO/IEC responsible for recommending a set of standards for compression, decompression, processing, and coded representation of moving pictures, audio, and their combination, in order to satisfy a wide variety of applications. MPEG-7 (Multimedia Content Description Interface) aims at providing a standard for representation of the content of AV documents. See `http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm`.

[17] W3C, World Wide Web Consortium, is a consortium whose goal is to develop standards for the World Wide Web. See `http://www.w3c.org`.

[18] ISAN, International Standard Audiovisual Number, is an ISO standard under development which might become an equivalent to the ISBN (International Standard Book Number), currently used for text publication worldwide.

[19] SGML (Standard Generalized Markup Language)(Goldfarb, 1990) is an ISO standard for the markup of textual documents.

[20] The TEI (Text Encoding Initiative)(Burnard and Sperberg-McQueen, 1994) is an international project to develop guidelines for the preparation and interchange of electronic texts for scholarly research, and designed to satisfy a wide range of uses by the language industries. It provides a very broad set of descriptors (tags) organized into an SGML DTD (Document Type Definition).

depends on the task the user is performing. It seems impossible to define any generic description scheme that would be independent of any context of use. This is particularly true of humanities scholars working on AV documents: an historian, a semiotician and a specialist of the mass-media will not interpret AV documents in the same way.

3.4. A need for a new model

We claim that AV document representation and systems built upon it should now be designed keeping a global framework for use in mind. As a consequence, full-indexing models should not restrict the expression of interpretations of AV documents but, on the contrary, should focus on flexible AV annotation tools allowing users to write down their own interpretations of AV documents (whatever they may be) and to use this interpretation as a navigation tool.

These are the assumptions of recent projects such as SESAME[21] or DELPHES[22]. Following this idea, we believe that re-introducing real-life scholarly uses of documents as a central issue in AV indexing models is crucial (Simpson-Young and Yap, 1995). We respond to this problem with our own model.

4. Representating and exchanging AV material: our proposal

The representation model used to index AV documents should be as flexible as possible in order to allow any domain-dependent interpretation to be expressed and used as a means of accessing the content. However, once stated that this type of model is based on annotation, that is to say on the reformulation of the AV content in a semiotic form which will make it usable, one question remains, Which semiotic form should be chosen? Should AV documents be annotated using text, numbers, still images (keyframes), moving images (extracts or recomposed images), or any other means of representation?

As stated in previous sections, AV material differs from text in that it is not directly computable and manageable in a symbolic form. Though

[21] SESAME (System for Multimedia and Audiovisual Sequences Exploration enriched by Experience) is a French project supported by France Télécom (through CNET/CCETT) to propose and study a global approach to make use of the potentially huge repositories of audiovisual documents.

[22] European project for the creation of hypermedia history courses allowing students as well as teachers to annotate, manipulate and compose AV data using as a basis a full indexing method provided by INA.

textual data can be extracted from AV material (for instance, transcription of audio, or close caption automatic detection), it remains mostly non-textual. As a consequence, indexes will have to represent nontextual data and may be nontextual data themselves, and this can result in two principal forms for metadata: symbolically and nonsymbolically expressible metadata.

Nonsymbolically expressible metadata are the results of computer calculations such as color histograms and images extracted or constructed from shots. Although it seems obvious that signal processing techniques for feature extraction and similarity search are an important field for development of nontextual retrieval (i.e., query by example) from AV data, we should nevertheless note that this cannot be totally sufficient unless computers truly understand AV data as human-beings do. For instance, if a computer is given an image as a prototype for similarity retrieval, it must define what the relevant and important objects or colors are as a human would. This remains an interpretative task requiring synthesis, and thus a tricky problem for a computer essentially skilled in thorough analysis. It appears immediately that low-level metadata features cannot be enough for the AV representation scheme needed.

Symbolically expressible metadata such as date of creation, cast, keywords, etc. remain mainly in the scope of human interpretation, even if computers can analyze shot cuts and people detection, for example. It seems at the present time that the most practical form of computer representation is the textual one (as opposed, for instance, to visual language form), resulting from a semiotization of AV content elements into linguistic terms. Most of AV annotations will naturally belong to this form[23].

[23] We would like to say a word on the notion of the visual thesaurus. If we choose linguistic terms as a means of computer representation for metadata, some other approaches such as those suggested by Davis (Davis, 1993), use icons instead. This type of annotation could raise the idea of developing a visual thesaurus for description of AV material. However, we believe that if such a thesaurus is not correlated with AV raw data (that is, if these icons act as symbols and are in no way linked to images), its specificity would consist only in replacing a term by an icon. On the other hand, if icons used in the thesaurus have visual similarity with AV objects, then "thesaurus" would not be the right term. Indeed, a thesaurus is a tree relating types of objects (e.g., the term "car" relates all the instances of the term "car" such as "car", "cars", etc.). If the icons are directly derived from the AV data, they cannot be anything but instances. This means that a car in a movie will not be represented by the same icon as a car in another movie since the source images where not the same. This type of thesaurus is merely a catalogue of objects available in the corpus and it can only act, at best, as a visual media access to easily recognizable image objects (cars, faces). These latter having names, they can be described in linguistic

In the end, it appears that both types of metadata have to be used in AV descriptions, but that the textual form is to remain essential for our representation needs, especially in the scholarly research fields. To show how this can be implemented, we will first present AI-STRATA, a model for describing AV content, and then AEDI, a format for exchanging AV descriptions, followed by an example illustrating the possibilities provided by such approaches.

4.1. A formal model for representing and indexing AV documents

For a more detailed presentation of the concepts addressed here, the interested reader should refer to (Prié et al., 1998).

4.1.1. *The problem of partitioning a document*

Annotation is the fundamental concept when considering digital sequential document representation and modeling, . In this context, it consists in attaching an annotation (a description) to a piece of the document, each piece being delimited by two boundaries. For temporal media, these boundaries are obviously two instants in the AV stream.

To characterize annotation of AV data, we defined several criteria. Time granularity is concerned with the level and the regularity of the partitioning of documents into AV pieces: document level, shot or scene level with full decomposition, or pieces as simple strata without constraints. The kind of metadata used to annotate the pieces is the second criterion: from the low-level features mentioned above to higher conceptual level characteristics such as shots, keywords, or texts; everything is possible. The third criterion is the degree of complexity of the organization of characteristics into annotations: simple or atomic when a term or a numerical feature is attached to a piece, it can reach higher complexity with attributed structures or even semantic networks.

According to these criteria, several ways exist to describe, at different levels and with different complexities, AV material pieces that are cut following different granularity schemes. Our last and fourth criterion for annotation characterization deals with the structuring of pieces of documents into documents, and is strongly related to granularity choices. Two principal approaches for structuring AV documents are currently in use: segmentation and stratification.

Segmenting an AV document consists in cutting it up into predefined pieces (mainly shots) which will be annotated later. An arborescent structural organization (related to C. Metz's research as reported

terms. We therefore think that our linguistic approach remains the most adequate in the current state of our knowledge and of the technology.

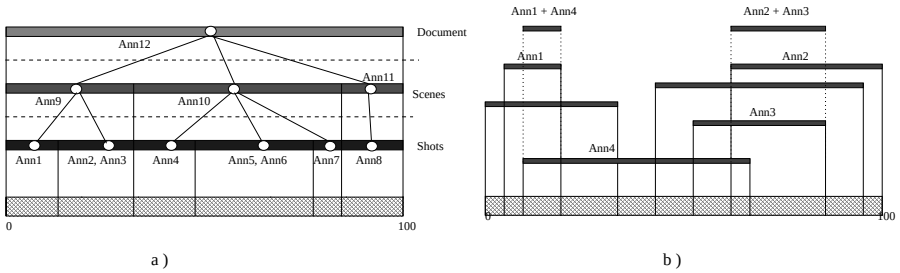


Figure 3. Segmentation and stratification

in (Corridoni et al., 1996)) is then often set up to express document structure (see Figure 3-a).

The stratification approach (Aguierre Smith and Davenport, 1992) allows the annotator to freely define *strata* (pieces) when they are needed. Davis (Davis, 1993) uses icons to annotate characteristics, possibly organized into sentences, and the a posteriori useful partitioning can be derived from strata intersection (see Figure 3-b).

The essential and important difference between these two approaches lies in fact in the definition of the temporally situated and annotated pieces of documents. In one case partitioning (and structural assumptions on the document) exists before an annotation, which itself can be considered as second to segmentation, while in the other case it is dynamically created by the annotation process, annotation and partitioning being tightly linked. In our model, we favored the stratification scheme because it seemed better adapted to representation of the sequential and dynamic aspects of AV material and because we consider strata and atomicity of the annotation as paramount. Another reason for this choice is that stratification subsumes segmentation, in other words, that relating strata is a way to structure a document. In the next part we will present the annotation interconnected strata approach for AV representation.

4.1.2. *AI-Strata*

An object of interest is defined as any object (in the general sense of the term) that can be spotted when watching/listening to an AV stream. Objects of interest can refer to any kind of characteristic, at any level of abstraction, and there are as many of them as analyses of the stream. We group these analyses into analysis dimensions that spot the same kinds of objects. For instance, an analysis dimension can be related to a detection of a shot, faces, people, movements, or President Clinton.

As soon as an object of interest is detected, it defines a temporally extended audiovisual unit (AVU) that represents a stratum, and at least one annotation element (AE) such as a term, the symbolic expression of

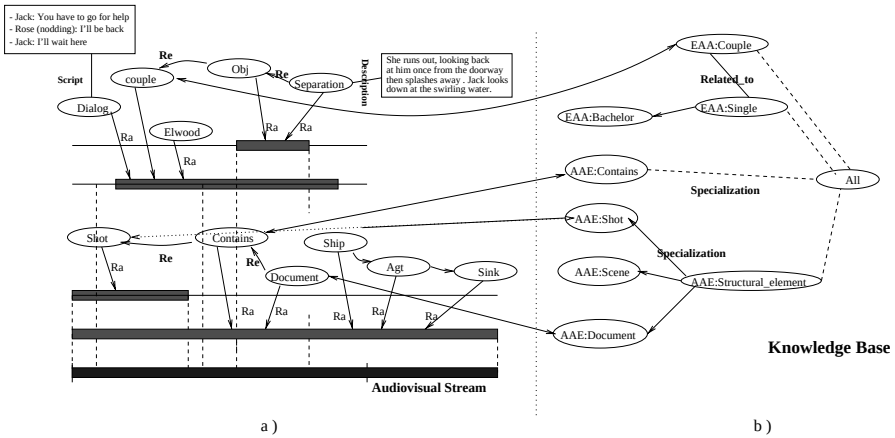


Figure 4. General overview of AI-STRATA

its meaning. The annotation element annotates (is in relation R_a to) the audiovisual unit it has defined (still within the stratification approach). For instance, spotting a shot leads to defining an AVU annotated by the AE $\langle Shot \rangle$, spotting a well-known face leads to the creation of another AVU annotated by $\langle Clinton \rangle$, and so on for any object of interest: $\langle Zoom \rangle$, $\langle Round\_shape \rangle$, $\langle Sad \rangle$ etc. A second description level is possible when considering AE attributes, for example, the numerical value of the histogram for $\langle Color\_histo \rangle$, a text for $\langle Script \rangle$ or an image for $\langle Key\_frame \rangle$.

To complete the primitive annotation that defines AVU, it is possible to add as many AEs as necessary. The first way to do so consists in adding AEs that have the same temporal range, for example adding to an AVU defined by $\langle Document \rangle$ an AE regarding its author. The second way results from structuring of the annotation: in order to express more complex information than simply terms with a temporal extension, we allow relations between them. For instance, to express that "this shot is included in that scene," or that "this shot is re-used in that document," or that "this character is doing this action," we define relations between annotations in the same way we already annotated. Indeed, we first define annotation elements that act as relation terms, and then connect them with a simple elementary relation R_e. For instance, supposing an AVU annotated by $\langle Ship \rangle$ and $\langle Sink \rangle$, then using a relation AE like $\langle Agt \rangle$, it is then possible to express "A ship is sinking" (see Figure 4).

As representation units and indexes, annotation elements are the metadata that support all access to AV material (because of the AVU[24]). The annotation process results in a graph composed of annotation ele-

[24] This is a solution to the annotation location problem, i.e., the determination of the elements used to link the annotation (the metadata used for description) and the

ments and audiovisual units (see Figure 4), and is led by a professional (an archivist) or nonprofessional user. In order to facilitate and monitor later access, it is necessary to consider AEs as terms issued from a controlled vocabulary in a knowledge base. An AE is then issued from an abstract annotation element (AAE). The knowledge base is in fact a network of AAEs with classical thesaurus relations (hierarchical or otherwise), plus information regarding possible attributes of AEs, or privileged relations.

4.1.3. *Use of AI-STRATA*

An annotation task through AI-STRATA follows guidelines linked to the goals of the on-going AV material use task. Indeed, the vocabulary that is used, the analysis dimensions that are considered, and the precision of the annotation depend on the precision of content description that is needed. Moreover, it is always possible to go deeper into the annotation and complete it.

The annotation from AV material following the AI-STRATA model finally leads to a set of AVUs and AEs. This set is not independent of the system, it is part of it, the document not only appearing per se in the base as a set of annotated AVUs, but it can also be linked to other documents via AE relations. The deconstruction of a document based on its annotation gives a foundation for

- retrieval of AV material and access to it through AEs and their attributes, reuse of AV material;
- individual annotation as work both on the document and on its primary annotation, as a kind of AV stream writing;
- browsing and navigating along AE relations to explore context from the AVUs, construction of views of a document through contextual filters, contextualization acting as a way to provide new meaning by relating annotations[25], etc.

data itself (the AV stream). This determination is not trivial since such AV segments, composing the material upon which useful descriptions can be constructed, convey no other semantic meaning than "this segment was cut-up because it was considered as useful for some purpose." All the semantics belongs to the description of these segments, in other words, to the annotations attached to them. This description can therefore evolve, be completed or changed depending on the context in which it is used.

[25] For deeper insight into context modeling studies, see (Prié et al., 1999).

4.2. A format for exchanging descriptions of AV documents

Once descriptions of AV documents are produced using models such as AI-STRATA, scholars need a way to exchange them. Indeed, active reading and academic publishing are often collaborative tasks requiring efficient exchange of annotations. Moreover, scholars working in archive libraries might want to stop the analysis of a document, store their work in progress, keep it with them, and use it at another moment. Although the document may be the property of the archives, the annotations are the property of the researcher. There are also many other reasons why it should be possible to exchange descriptions of AV content using a common format (inter-library loan, export of computer analysis results, etc.).

Of course, this exchange format should be application and platform independent so as to last longer than the current constantly evolving technological context.

In this section, we will introduce Audiovisual Event Description Interface (AEDI). AEDI is a language developed by INA for the expression, validation, storage and exchange of descriptions of AV documents. It was not originally intended for the exchange of AI-STRATA annotations. AI-STRATA relies on concepts and models from the knowledge representation community, whereas AEDI is an attempt to extend the concepts and technology of the electronic publishing community to specific AV documentation problems. However, we believe that the current trends in document management systems require the combination of these two paradigms in order to provide more accurate and intelligent, but still effective, processing of documents.

The next section will attempt to show how AI-STRATA can be expressed in the context of AEDI.

4.2.1. *Overview of the AEDI model*

The AEDI model defines three main elements: the description scheme, the document description, and the media description.

- Description scheme. This defines the classes of elements which must or can appear in a specific type of description as well as the constraints applying to the combination of these elements (cardinality constraints, instantiation constraints, grammatical constraints, etc.). AEDI Description schemes have a role similar to Document Type Definitions (DTDs) in SGML or XML, but they allow the expression of more complex constraints. In particular, they define and validate three-dimensional document grammars based on coordinate systems quite similar to the ones developed

for HyTime (ISO, 1992). Just as the syntax of SGML defines core objects such as elements and attributes, we defined the core classes as the following:

- data types: AEDI provides some basic data types such as string, integer, float, boolean and time references;
- value container: attribute-value pairs, where the value can be a stand-alone object (ex: title:string), a list, or a structure of objects (for example, Filmography:film+);
- descriptor: element of a description which can hold a name and a set of attributes. For instance, an actor description element, defined by the first name, last name and filmography, is a descriptor;
- axial descriptor: subclass of a descriptor characterized by a content model defined on the axis of the descriptions. For instance, a shot description element, defined on a time axis is an axial descriptor.

– *Document description.* This is composed of the instances of the classes defined in the description scheme which must conform to the grammar of their class. These instances are the objects describing the actual content of the AV document. They are organized in a tree structure of axial descriptors for which an API equivalent to the Document Object Model (DOM) in XML has been defined. It is possible to validate the description against the description scheme, as an SGML instance can be validated against its DTD, using a Java parser developed for the ACTS-DICEMAN project[26].

– *Media description.* In AEDI, we distinguish the description of the document from the description of the media. The document description describes the content of the document (for example, today's 6 o'clock news program), whereas the media description contains the information necessary to associate this document with one or more pieces of media. In particular, all the axes of the document (defined in the description scheme) are virtual axes and they can be related to actual axes of the media by what we call a projection (see Figure 5). This projection mechanism provides a way to use the description to play documents that are available on more than one chunk of media (e.g., a single document can be divided into several files) without forcing users to get involved with files and media management systems. Moreover, it enables archives to change the media used for storage (by digitizing a

[26] See `http://www.teltec.dcu.ie/diceman/`

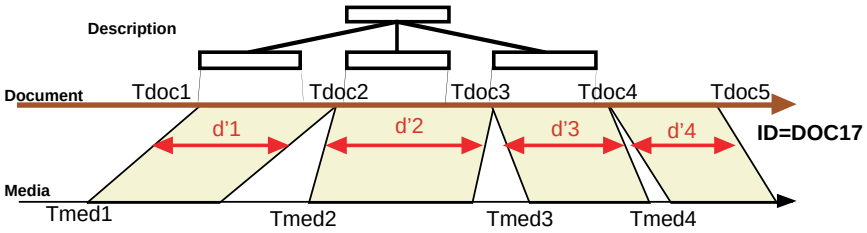


Figure 5. Separation between metadata and data using an indirection in AEDI

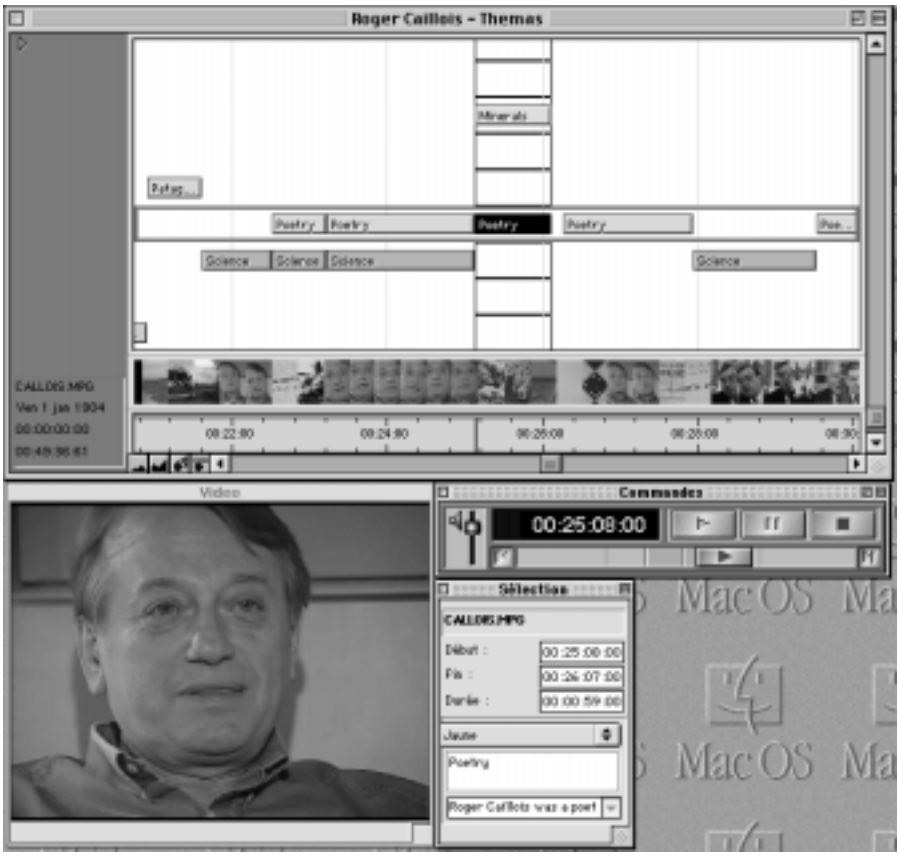


Figure 6. Representing thematic annotations with AEDI in Médiascope

tape for instance) without any impact on user interaction with the library system.

4.2.2. *Current use of AEDI*

AEDI is a general model for the representation of description structures and it provides an XML-based exchange syntax. It is used in INA's AV indexing projects. Some prototypes have been built to allow scholars

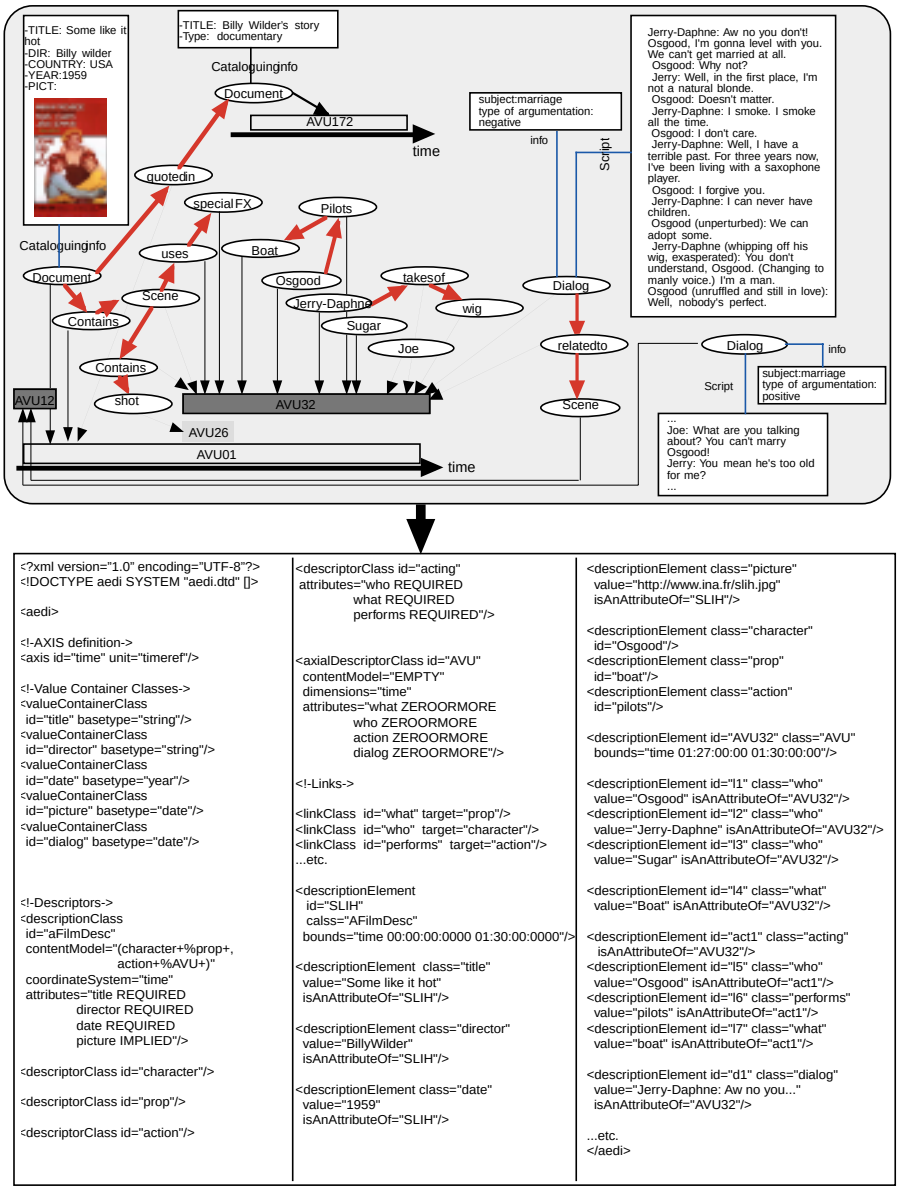
to produce, exchange and reuse annotations on AV documents, which are based on AEDI. For instance, *Médiascope*, which is used in the French Legal Deposit's Computer-Aided AV Reading Station, proposes a graphical interface for the annotation of AV documents, automatic segmentation facilities, and an export function which outputs descriptions as AEDI text files. Figure 6 shows how *Médiascope* can be used to annotate and browse thematic strata on a documentary. When clicking on a segment related to $\langle poetry \rangle$, the user visualizes the inclusion of this particular segment in a "theme = poetry" annotation dimension (horizontal scale) and other segments available on the same timespan (vertical scale): the theme of the selected temporal segment is related to $\langle minerals \rangle$ as well.

4.3. An example of collaboration between AI-STRATA and AEDI

As emphasized above, AI-STRATA is a model for expressing knowledge about the content of AV documents whereas AEDI is a model and a format to express and control the structure of descriptions. However, it seems possible to combine the two approaches by expressing the concepts conveyed by AI-STRATA annotations in a document-structuring approach.

Figure 7 presents an example of an AI-STRATA annotation-graph of the last scene from *Some like it hot* (Wilder, 1959). A $\langle Document \rangle$, the movie, contains a $\langle Scene \rangle$ (temporally included in it), which is annotated by its main characters and objects, and supplementary relations between them. The annotation concerning, for instance, the $\langle Dialog \rangle$ taking place in the scene is added, with `info` and `script` attributes. Inter-audiovisual unit relations are used, both inside the document (between $\langle Dialog \rangle$ and $\langle Scene \rangle$) and outside of it, as for the $\langle Quoted\_in \rangle$ relation). These two important features (attributes and inter-AVU relations, including inter-document level) show the expressive power of the model.

At the bottom of the figure a (partial) transcription from this annotation into the AEDI model and syntax is provided. As these two models are originally different, there are multiple ways to map one onto the other. For this example, we have chosen to represent AI-STRATA's audiovisual units by AEDI axial descriptors, annotation elements by attributes, abstract annotation elements by descriptor classes, and analysis dimensions by axial descriptors whose boundaries are implied by their sub-descriptors in the description structure.




```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE aedi SYSTEM "aedi.dtd" []>

<aedi>

<!-AXIS definition->
<axis id="time" unit="timeref"/>

<!-Value Container Classes->
<valueContainerClass
 id="title" basetype="string"/>
<valueContainerClass
 id="director" basetype="string"/>
<valueContainerClass
 id="date" basetype="year"/>
<valueContainerClass
 id="picture" basetype="date"/>
<valueContainerClass
 id="dialog" basetype="date"/>



<!-Descriptors->
<descriptionClass
 id="aFilmDesc"
 contentModel="(character+%prop+,
                action+%AVU+)"
 coordinateSystem="time"
 attributes="title REQUIRED
             director REQUIRED
             date REQUIRED
             picture IMPLIED"/>

<descriptorClass id="character"/>

<descriptorClass id="prop"/>

<descriptorClass id="action"/>
```

```
<descriptorClass id="acting"
 attributes="who REQUIRED
             what REQUIRED
             performs REQUIRED"/>


<axialDescriptorClass id="AVU"
 contentModel="EMPTY"
 dimensions="time"
 attributes="what ZEROORMORE
             who ZEROORMORE
             action ZEROORMORE
             dialog ZEROORMORE"/>

<!-Links->

<linkClass  id="what" target="prop"/>
<linkClass  id="who"  target="character"/>
<linkClass  id="performs"  target="action"/>
...etc.

<descriptionElement
  id="SLIH"
  calss="AFilmDesc"
 bounds="time 00:00:00:0000 01:30:00:0000"/>

<descriptionElement  class="title"
 value="Some like it hot"
 isAnAttributeOf="SLIH"/>

<descriptionElement class="director"
 value="BillyWilder"
 isAnAttributeOf="SLIH"/>

<descriptionElement class="date"
 value="1959"
 isAnAttributeOf="SLIH"/>
```

```
<descriptionElement class="picture"
 value="http://www.ina.fr/slih.jpg"
 isAnAttributeOf="SLIH"/>

<descriptionElement class="character"
 id="Osgood"/>
<descriptionElement class="prop"
 id="boat"/>
<descriptionElement class="action"
 id="pilots"/>

<descriptionElement id="AVU32" class="AVU"
 bounds="time 01:27:00:00 01:30:00:00"/>

<descriptionElement id="l1" class="who"
 value="Osgood" isAnAttributeOf="AVU32"/>
<descriptionElement id="l2" class="who"
 value="Jerry-Daphne" isAnAttributeOf="AVU32"/>
<descriptionElement id="l3" class="who"
 value="Sugar" isAnAttributeOf="AVU32"/>

<descriptionElement id="l4" class="what"
 value="Boat" isAnAttributeOf="AVU32"/>

<descriptionElement id="act1" class="acting"
  isAnAttributeOf="AVU32"/>
<descriptionElement id="l5" class="who"
 value="Osgood" isAnAttributeOf="act1"/>
<descriptionElement id="l6" class="performs"
 value="pilots" isAnAttributeOf="act1"/>
<descriptionElement id="l7" class="what"
 value="boat" isAnAttributeOf="act1"/>

<descriptionElement id="d1" class="dialog"
 value="Jerry-Daphne: Aw no you..."
 isAnAttributeOf="AVU32"/>

...etc.
</aedi>
```

Figure 7. An example of annotation

5. Discussion: long term trends in the use of AV documents

In a recent conference[27], Allen Renear stated that "The practice of computer text encoding has turned out to be rather effective at generating insights into the nature of textuality." In fact, text encoding has modified our vision of what a text is by providing new access, new ways of reading textual documents. Goody showed in (Goody, 1977) that writing, as a technology allowing human communication to overcome the temporal constraints of simple speech, was the origin for specific thought categories such as tables and lists[28]. These categories in turn became the paradigm used by grammarians when they analyzed and described natural language, leading to the development of the philosophy of language (Auroux, 1994).

In a similar way, we can imagine that new AV reading environments allowing viewers to escape the continuous timeline of the document may change our perception of what AV documents are, and thus stimulate new theories on AV communication.

For instance, AV semiotics theories may evolve. Some semioticians consider AV documents as communication acts and study how the AV medium is used to convey a certain content or to perform certain communicative functions. As stated above, the problem is to define what the content of a document is. Apart from shot cuts, current AV documents contain no explicit layout marks of their editing which could be translated into a logical representation of the document, as is commonly done for texts[29]. In the future, however, a standard description scheme for AV content may be provided (see section 3.3) and scholars may be able to use it as a basis for their research. It should be possible, for instance, to compare how two similar contents — from the point of view of a description standard — are expressed in terms of AV editing and recording. Conversely, as individual ontologies can also be defined,

[27] Oxford HCU Lecture Series: Text Ontology and Computer Text Encoding, Allen Renear, "Text Ontology from Below: The Contribution of Computing Practice to New Theories of Textuality", July 1998.

[28] In a more general view, (Stiegler, 1994) and (Bachimont, 1996) show that the technological environment is constitutive (in the Husserlian phenomenology sense) of categories and structures of human thought: there is a constitutive double bind between technology and human cognition. Moreover, psychological and ethnomethodological studies such as (Vygotsky, 1978) and (Suchman, 1987) show that new technical environments, which were originally built to help users to perform a task, modify in turn users' cognitive capabilities and therefore modify the task itself.

[29] This process is called up-translation. It requires an interpretation (i.e., a mapping) of the external appearance of a text (e.g., the section titles are in bold, 12 points, underlined) into a logical structure of the document such as an SGML Document Type Definition. This logical structure can then be used as a basis for electronic publication processes.

idiosyncratic annotation will appear. Work done in this area, as well as exchanges among scholars, would certainly be of interest and would lead in turn to definition of new logical representations of documents. At the same time, image processing tools will evolve and produce more and more metadata which will be incorporated into scholarly analyses and used as a basis for annotation. As a consequence, a new approach to AV documents, based on emerging links between low-level analysis and high-level conceptual annotation to AV content might appear and provide new up-translation possibilities.

6. Conclusion

In this article, we analyzed how audiovisual (AV) documents are used by humanities scholars and we described how these practices were limited by the current technological environment. Because of the nature of AV media, digitization of catalogues, indexes, and documents do not suffice to ensure adequate technological conditions for the emergence of new practices. Full indexing, the representation of the AV content with efficient random access to any piece of a document, remains a crucial issue and a major challenge to newly constructed digital AV archives. Starting from real-life scholarly uses of AV documents (rather than from technological constraints), we introduced a model for the annotation of AV documents (AI-STRATA) and a format for the exchange of such descriptions among users (AEDI).

We presented here the advantages and the general framework of the AI-STRATA approach. Current and future work in this field include computer-aided annotation models design, context and annotation-graph handling, and knowledge- and experience-based learning for and with AV material annotation. AEDI is also research in progress which is likely to develop greatly in the future. In particular, its ability to represent and validate n-dimensional grammars should be assessed on large scale corpora.

Finally, we believe that the management of the tremendous amount of digital AV material that we will have to face in the future calls not only for standardization but also for the development of computer-aided AV semiotics studies that should detail audio-visual media communication more thoroughly.

References

Aguierre-Smith, T. G.: 1992, 'If you could see what I mean... Descriptions of video in an Anthropologist's video notebook'. Master's thesis, University of California, Berkeley.

Aguierre Smith, T. G. and G. Davenport: 1992, 'The Stratification System, a Design Environment for Random Access Video'. In: *Proc. Network and Operating System Support for Digital Audio and Video - 3th International Workshop*. La Jolla.

André, J., R. Furuta, and V. Quint: 1989, *Structured Documents*. Cambridge University Press.

Auffret, G. and B. Bachimont: 1999, 'Audiovisual Cultural Heritage: from TV and radio archiving to hypermedia publishing'. In: *European Conference on Research and Advanced Technology for Digital Libraries*. Paris, France.

Auffret, G., J. Carrive, O. Chevet, T. Dechilly, R. Ronfard, and B. Bachimont: 1999, 'Audiovisual-based Hypermedia Authoring: using structured representations for efficient access to AV documents'. In: K. Tochtermann, J. Westbomke, U. K. Wiil, and J. J. Leggett (eds.): *ACM Hypertext '99*. Darmstadt, pp. 169–178.

Auroux, S.: 1994, *La révolution technologique de la grammatisation*. Liege, Belgium: Mardaga. In French.

Bachimont, B.: 1996, 'Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser'. Ph.D. thesis, Ecole Polytechnique, France. In French.

Bordwell, D.: 1993, *The cinema of Eisenstein*. Harvard university press.

Burnard, L. and C. Sperberg-McQueen: 1994, *TEI P3 : Guidelines for Electronic Text Encoding for Interchange*. Lou Burnard and C.M. Sperberg-McQueen.

Bush, V.: 1945, 'As We May Think'. *The Atlantic* **176**(1), 101–108.

Chahuneau, F., C. Lécluse, B. Stiegler, and J. Virbel: 1992, 'Prototyping the ultimate tool for scholarly qualitative research on texts'. In: *8th Annual conference of the UW Centre for the New Oxford English Dictionary and Text Research*. Waterloo.

Chang, S., J. Smith, M. Beigi, and A. Benitez: 1997, 'Visual Information Retrieval from Large Distributed Online Repositories'. *Communications of the ACM* **40**(12), 63–71.

Christel, M.: 1995, 'Addressing the Contents of Video in a Digital Library'. In: *Electronic Proceedings of the ACM Workshop on Effective Abstraction in Multimedia*. San Francisco, California.

Corridoni, J. M., A. Del Bimbo, D. Lucarella, and H. Wenxue: 1996, 'Multi-perspective Navigation of movies'. *Journal of Visual Languages and Computing* **7**, 445–466.

Davis, M.: 1993, 'Media Streams: An Iconic Visual Language for Video Annotation'. In: *Proceedings of the 1993 IEEE Symposium on Visual Languages*. Bergen, Norway, pp. 196–203.

Denel, F., G. Pijut, and J.-M. Rodes: 1994, *Le dépot légal de la radio et de la télévision*. Paris: INA-publications. In French.

Ferro, M. and J. Planchais: 1997, *Les médias et l'histoire*. Paris: CFPJ ed. In French.

Flickner, M., H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, G. M., H. J., L. D., P. D., S. D., and Y. P.: 1995, 'Query by Image and Video Content : the QBIC System'. *IEEE Computer* **28**(9).

Fox, E. A. and G. Marchionini: 1998, 'Toward a worldwide digital library'. *Communications of the ACM* **41**(4), 28–32.

Furuta, R.: 1997, 'What can Digital Libraries teach us about hypertext?'. *SIGLINK newsletter* **6/3**, 7–9.

Goldfarb, C. F.: 1990, *The SGML Handbook*. Oxford: Clarendon Press.

Goody, J.: 1977, *The domestication of the savage mind*. Cambridge University Press.

Grosky, W.: 1997, 'Managing Multimedia Information In Database Systems'. *Communications of the ACM* **40**(12), 73–80.

Gupta, A., S. Santini, and R. Jain: 1997, 'In Search of Information in Visual Media'. *Communications of the ACM* **40**(12), 35–42.

Harrison, H. W.: 1991, *The FIAF Cataloguing rules for film archives*. FIAF.

Hjesvold, R. and R. Midtstraum: 1994, 'Modelling and Querying Video Data'. In: *20th VLDB Conference*. Santiago, Chile, pp. 686–694.

ISO: 1992, *ISO 10744:1992, Information technology – Hypermedia/Time based Structuring Language (HyTime)*. Geneva: ISO.

Jacquinot, G.: 1977, *Image et pédagogie : analyse sémiologique du film à intention didactique*. Paris: Presses universitaires de France. In French.

Jacquinot, G.: 1985, *L'école devant les écrans*. Paris: E.S.F. In French.

Jost, F.: 1992, *Un monde à notre image : énonciation, cinéma, télévision*. Paris: Meridiens Klincksiek. In French.

Landow, G. P.: 1997, *Hypertext 2.0, The Convergence of Contemporary Critical Theory and Technology*. Baltimore + London: The Johns Hopkins University Press.

Massey, M. and W. Bender: 1996, 'Salient stills: Process and practice'. *IBM Systems Journal* **35**(3–4), 557–573.

Metz, C.: 1968, *Essais sur la signification au cinéma I*. Paris: Klincksieck. In French.

Metz, C.: 1972, *Essais sur la signification au cinéma II*. Paris: Klincksieck. In French.

Minka, T.: 1996, 'An Image Database System that Learns From User Interaction'. Master's thesis, MIT, Cambridge, MA.

Mitchell, J., W.Pennebaker, C. Foog, and D. L. Gall: 1996, *MPEG Video Compression Standard*. New York: Chapman and Hall.

Nastar, C., M. Mitschke, C. Meilhac, and N. Boujemaa: 1998, 'Surfimage: a Flexible Content-Based Image Retrieval System'. In: *ACM Multimedia 98*. Bristol.

Oomoto, E. and K. Tanaka: 1993, 'OVID: Design and Implementation of a Video-Object Database System'. *IEEE Transactions on Knowledge and Data Engineering* **5**(4), 629–643.

Paepcke, A., C.-C. K. Chang, T. Winograd, and H. García-Molina: 1998, 'Interoperability for digital libraries worldwide'. *Communications of the ACM* **41/4**, 33–42.

Prié, Y., A. Mille, and J.-M. Pinon: 1998, 'AI-STRATA: A User-centered Model for Content-based description and Retrieval of Audiovisual Sequences'. In: *Int. Advanced Multimedia Content Processing Conf.*, Vol. 1554 of *LNCS, Springer-Verlag*. Osaka.

Prié, Y., A. Mille, and J.-M. Pinon: 1999, 'A Context-Based Audiovisual Representation Model for Audiovisual Information Systems'. In: *International and Interdisciplinary Conference on Modeling and Using Context*, Vol. 1688 of *LNAI, Springer-Verlag*. Trento.

Quint, V. and I. Vatton: 1994, 'Making structured documents active'. *Electronic Publishing* **7**.

Rosenberg, J.: 96, 'The structure of hypertext activity'. In: *ACM Hypertext'96*.

Sawhney, N., D. Balcom, and I. Smith: 1997, 'Authoring and Navigating Video in Space and Time'. *IEEE Multimedia* pp. 30–39.

Sheth, A. and W. Klas: 1998, *Multimedia Data Management — Using Metadata to Integrate and Apply Digital Media*, Chapt. Overview on Using Metadata to Manage Multimedia Data. McGraw-Hill.

Simpson-Young, B. and K. Yap: 1995, 'Work Process of Film and Television Researchers'. Technical Report 95/11, CSIRO Division of Information Technology.

Stiegler, B.: 1994, *La technique et le temps I, La faute d'Epiméthée*. Galilée. In French.

Suchman, L.: 1987, *Plans and Situated Action, the problem of human-machine communication*. Cambridge University Press.

Taniguchi, Y., A. Akutsu, Y. Tonomura, and H. Hamada: 1995, 'Efficient Acces Interface to Real-Time Incoming Video Based on Automatic Indexing'. In: *ACM Multimedia'95*. San Fransisco, pp. 25–33.

Virbel, J.: 1995, 'Annotation dynamique et lecture expérimentale : vers une nouvelle glose?'. *Littérature* **96**, 91—105. In French.

Vygotsky, L.: 1978, *Mind in Society*. Harvard Press.

Yeo, B. and M. Yeung: 1997, 'Retrieving and Visualizing Video'. *Communications of the ACM* **40**(12), 43–52.

Yeung, M., B. Yeo, and B. Liu: 1996, 'Extracting Story Unit from Long Programs for Video Browsing and Navigation'. In: *Int. Conf. on Multimedia Computing and Systems*. Vienna, Austria.

Zhang, H., C. S. S. Low, and J. Wu: 1995, 'Video parsing, retrieval and browsing : an integrated and content-based solution'. In: *ACM Multimedia'95*. San Fransisco, pp. 15–24.