INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11/ M4586

MPEG 99

MARCH 99/SEOUL

Title: **The TOCAI DS for audio-visual documents. Structure and concepts.**

Source: **Nicola Adami, Alessandro Bugatti, Riccardo Leonardi,**
Pierangelo Migliorati, Lorenzo Rossi
Department of Electronics
University of Brescia, Italy.

Status: **Proposal/discussion**

ABSTRACT

This document complements the description of the audio-visual (AV) description scheme (DS) called Table of Content-Analytical Index (TOCAI) proposed in MPEG-7 CFP that was evaluated in Lancaster (February 1999). This DS provides a hierarchical description of the time sequential structure of a multimedia document (suitable for browsing) together with an "analytical index" of AV objects of the document (suitable for retrieval). The TOCAI purposes and general characteristics are explained. The detailed structure of the DS is presented by means of UML notation as well, to clarify some issues that were not included in the original proposal. Some examples of XML instantiation are enclosed as well. Then an application example is shown. For an indication on how the TOCAI DS matches MPEG-7 requirements and evaluation criteria, refer to the original proposal submission.

INTRODUCTION

Nowadays more and more AV material arises from a variety of digital sources. Thus, there is the need to provide frameworks for an efficient navigation or browsing through the large amount of material being made available and to retrieve relevant information it contains according to a specific user. We are proposing a DS for multimedia documents (the TOCAI) which the aim to address the following functionalities:

- Characterisation of the temporal structure of a multimedia document from a semantic point of view at multiple level of abstraction. This may allow for a meaningful outlook to the multimedia document and at the same time offer the possibility to go, if necessary, deep into its temporal structure thanks to different degrees of detail in the presentation.
- Provide an efficient framework for retrieving selected objects of the document. These objects can be ordered according to different criteria, so as to allow fast search throughout the AI.
- Provide a categorisation of an AV document according to several attributes (authors, date of production, subjects, etc.).
- Provide useful informations about the document description itself like, e.g., the size of the description and the type of involved extraction methods with their reliability factor.

The original idea for such a DS comes out from the structure used for technical books. One may easily understand a book sequential organisation by looking at its table of content (generally located in the first pages) while one may quickly retrieve elements of interest by means of the analytical index (typically located at the end of the book). In the first case, the chronological order of presentation is preserved, which in the last case, an alphabetical order exists to facilitate the retrieval. The TOCAI allows a similar mechanism to address multimedia material, with one extension: It allows to retrieve information at any given level of abstraction, which is not normally the case in a book (each keyword in the index points normally to the page numbers only, not the sections or paragraphs where the topic of interest can be found).

This DS is created by the aggregation of four main description schemes: the *Table of Contents* (TOC), the Analytical Index (AI), the *Context* and the *Meta-descriptors* description schemes.

TOC DS

The TOC is organised in different hierarchical levels where the lower levels provide a detailed characterization of the sequential structure of the AV document while the higher ones have the role to offer a more compact description with associated semantics. A key aspect is that the items at each level are kept in **chronological order**.

Example

For a broadcast news document, we can have a TOC description could be the following:

- Summary (0:2'30'')

- Internal affairs (2'31':7')
  - Speaker presentation (…)
  - First reportage (…)
    - Shot 1 of the first reportage (…)
    - Shot 2 of the first reportage (…)
    - …
  - Second reportage (…)
    - …
  - Speaker presentation (…)
  - ….
- International affairs (…)
  - ….
- Sport news (…)
  - ….
- …

Where the data within parenthesis specify the temporal location of the segment while the label indicates its semantics. Every TOC item may be for clarity summarised by K-frames and audio segments.

The TOC DS is very useful for browsing and navigation, since it provides summaries of the document at several levels of details. Besides the meaningful characterisation of the temporal structure of the document, provided by the TOC DS, it may also be used for retrieval tasks as it can restrict the search field for a particular query, given the hierarchical structure which is created. As a summary two words are essential in the TOC concept:

1. hierarchy
2. chronological order

Analytical Index DS

The AI allows to create an **ordered** set of audio-visual objects. An item in the AI can point at different levels of detail according to the hierarchy provided in the TOC, or according to some other criterion. It is important to notice that thanks to the AI, more than one shot or more than one scene can be referenced by the same AI item. This gives the possibility to navigate along the audio-visual material not in sequential order, rather through scenes/shots containing similar objects.

AI objects can be semantic entities (like an AV scene belonging to a particular category, e.g. a dialogue), particular kind of images (backgrounds, foreground objects, etc.) but audio objects as well (like the musical motif and/or some keywords from a speech to text transcription). These objects can be ordered according to various criteria, which are listed in the DS. As a summary two words are essential in the AI concept:

1. order
2. reference pointer

Context DS

The TOCAI, which refers to the structure of an AV document, should be considered together with a DS describing the category of the audio-visual material. This contextual DS includes descriptors such as title of programme, actors, director, language, country of origin, etc. Indeed these informations are necessary for retrieving purposes to restrict the search domain.

Meta-descriptors DS

This DS has the role to incorporate in the TOCAI DS a set of descriptors carrying information

about how accurate is the description and by which means it has been obtained. The objective is to describe not the content but to give an indication of the reliability with which descriptor values have been assigned throughout the TOCAI DS.

DETAILED STRUCTURE OF THE PROPOSED DS

We describe now the TOCAI structure by presenting the hierarchical organization of its sub-description schemes and involved descriptors. As we said, the TOCAI is organized in four main DSs (see Figure 1). We explain the structure of each DS in the following subsections.

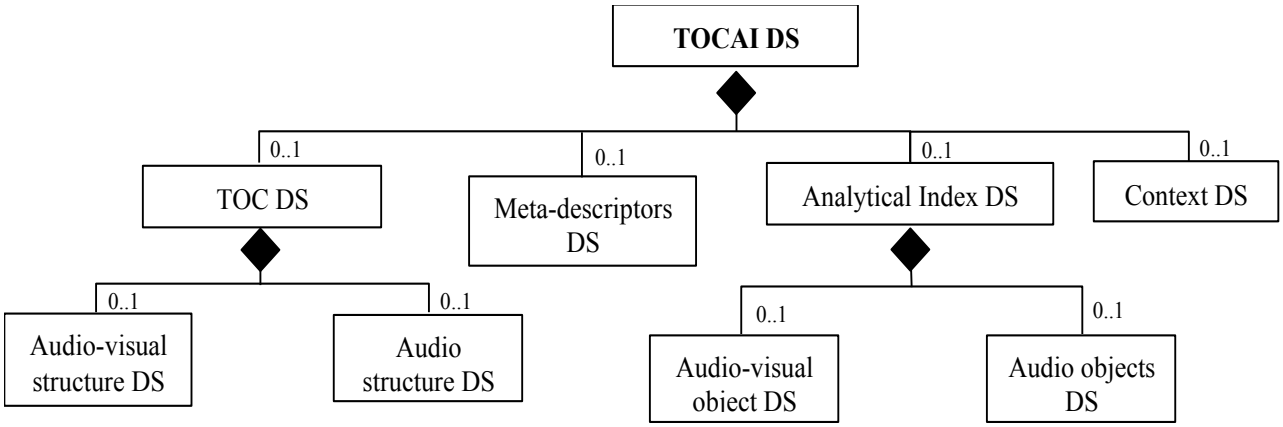

Figure 1: High level structure of the TOCAI DS

TOC DS

It describes the temporal structure of the AV document at multiple level of abstraction. It contains two DSs, explained below, namely *Audio-visual Structure* and *Audio Structure*. We proposed an *Audio-Visual Structure* DS rather than a simple visual DS because, from the semantic point of view, it is often necessary to consider the information carried by video together with the one provided by associated audio to recover reliable intermediate semantic levels for the description.

Audio-visual structure DS

This DS is represented in Figure 2. The two *Time-code Ds* specify the start and the end position of the AV document. The core of this DS is the *Scene DS*. A scene is a temporal segment having a coherent semantics at a certain hierarchical level. It is formed by a various number of sub-scenes, a time reference (2 time-code Ds) and a *type of scene D* (a string and, if useful, a characteristic icon). The elementary component of a scene is the shot[1]. The *Shot DS* indicates the type (cut, dissolve, fade in, etc.) of editing effects and their temporal location (*Editing effects D*). It includes a set of DSs for K-frames mosaic and outlier images of the shot. [2]

[1] A shot is defined by a sequence of frames captured from a unique and continuous record of camera.

[2] A mosaic represents the background in a shot. An outlier represents a foreground object in motion with respect to the background. These are typically extracted thanks to mosaicing techniques, which allows to register regions at different layers moving differently throughout the image sequence.

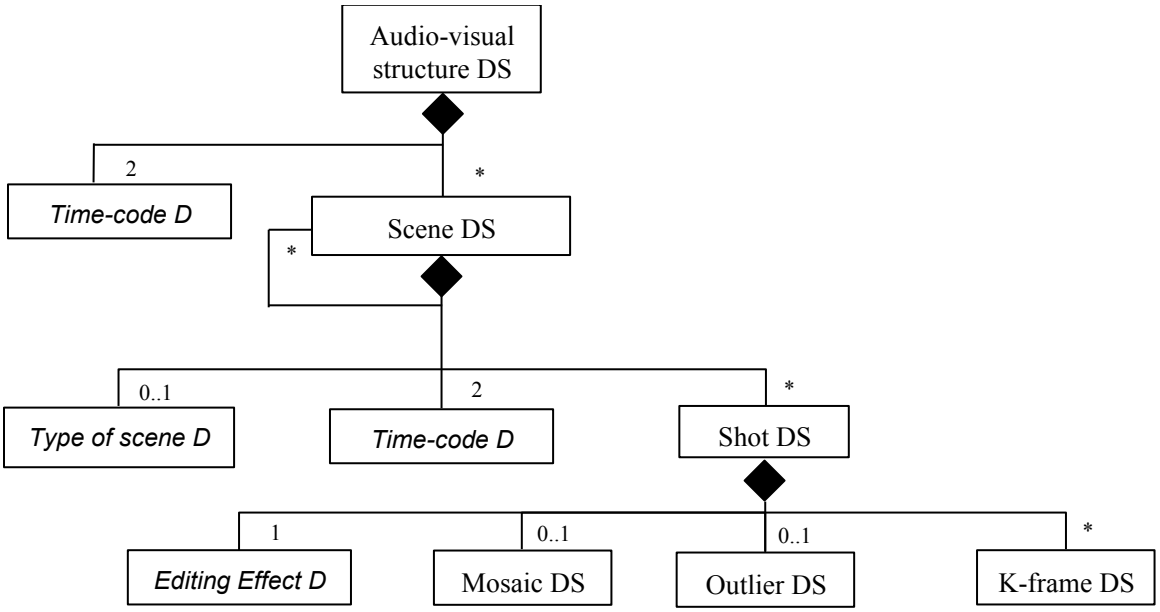

Figure 2: The Audio-visual structure DS.

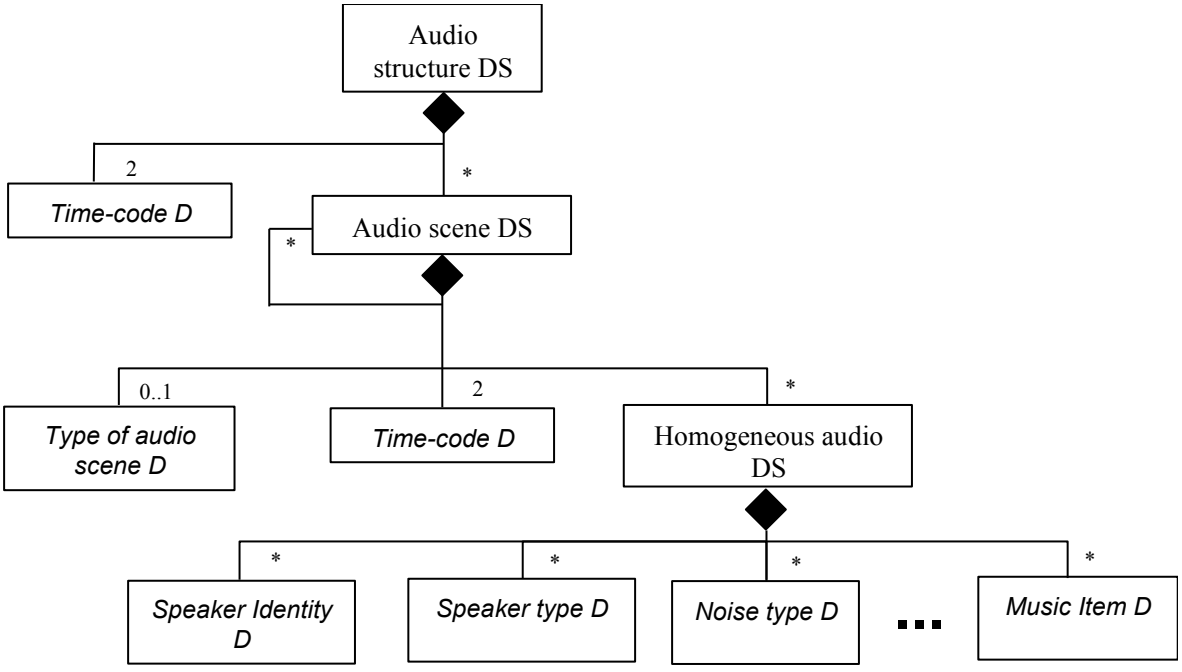

Figure 3: The Audio structure DS.

Audio-structure DS

This DS reflects the structure of Audio-visual DS (see Figure 3). Thus we can have various layers of audio scene. The Ds belonging to the *Homogeneous audio DS* represent the leaves of the tree, i.e. audio segments having an homogeneous audio source (for example a particular speaker, a particular noise, a defined music etc.). Each one of these Ds is constituted by an appropriate label and a time reference.

Analytical Index DS

As we said the AI allows to create an ordered set of document audio-visual objects pointing at different locations and different level of abstraction. Then this DS has the main role to support

retrieval of selected objects within the AV document. It is formed by two DSs: the *Audio-visual object DS* and the *Audio object DS*.

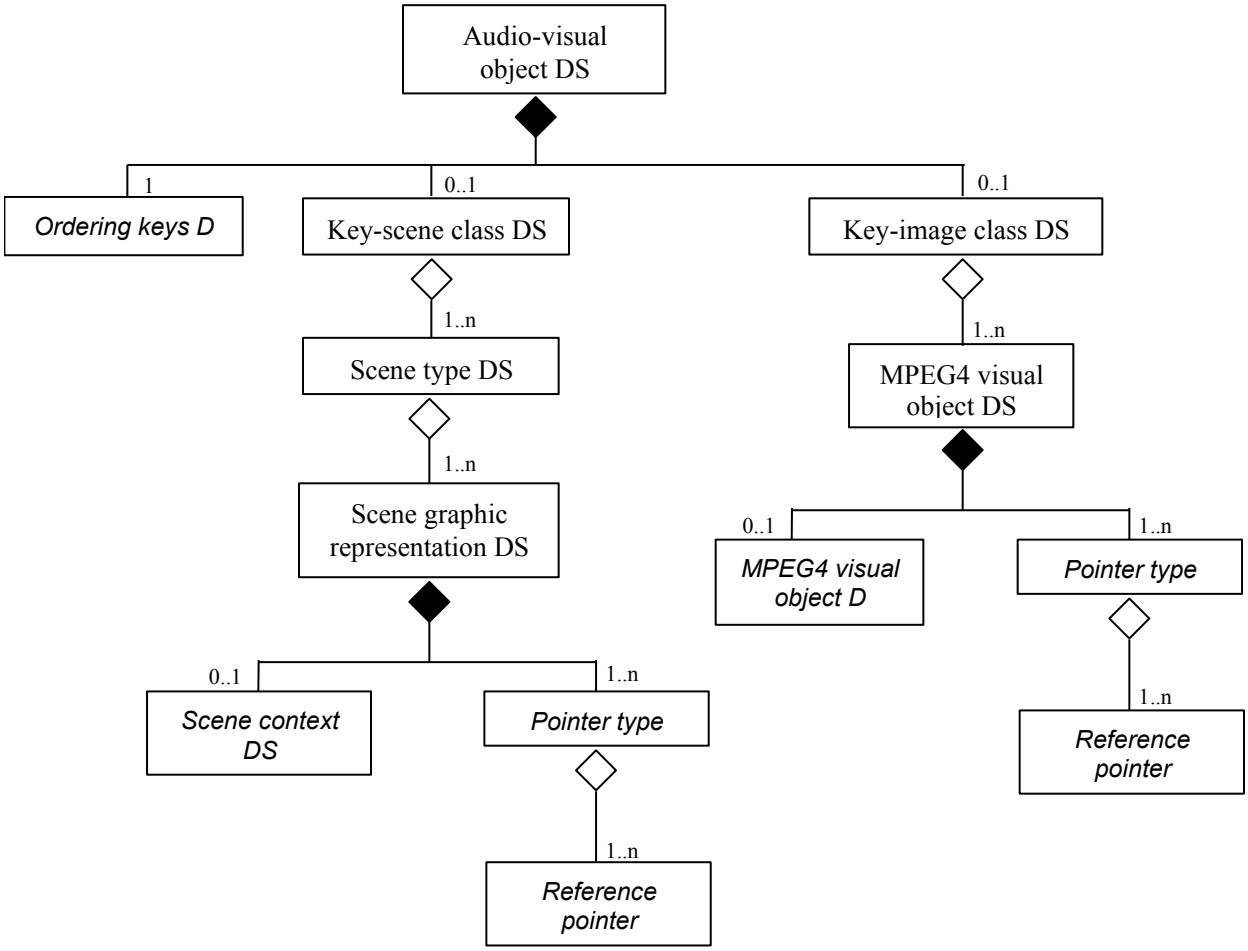

Figure 4: The Audio-visual object DS.

Audio-visual objects DS

The structure of this DS is shown in Figure 4. The *ordering keys D* is set of possible key for the ordering of AI items, e.g. colour or texture for images.

Audio- objects DS

The structure of this DS is shown in Figure 5. The *ordering keys D* is set of possible key for the ordering of AI items, e.g. name of musical instruments, time duration of an audio segment.

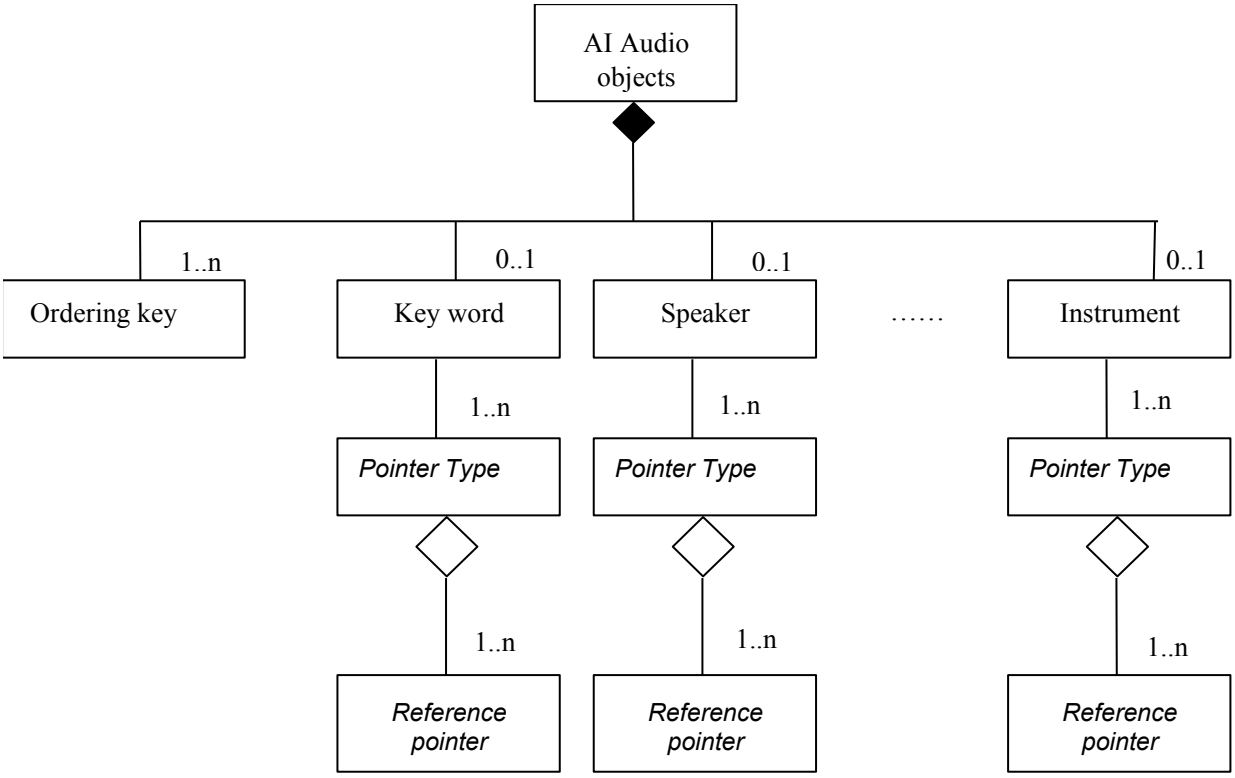

Figure 5: The structure of the Audio object DS.

Meta-descriptors DS

This DS consists of descriptors describing the other descriptors. These metadescriptors should be chosen so that they provide indirect but useful information about the content of a multimedia document. For example the reliability level descriptor gives users an idea about how much they can trust a given description for answering their query. Other import informations consist in the type of involved extraction methods or in the size of the description itself.

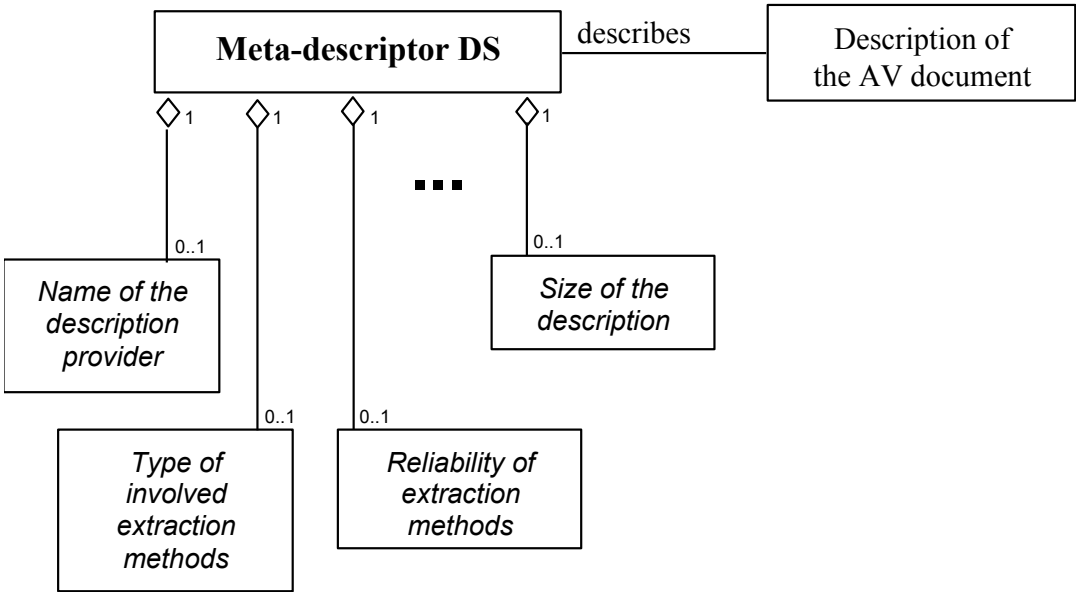

Figure 6: Structure of the Meta-descriptors DS.

Context DS

This DS is the set of the typical programme descriptors that are available, e.g., in a Radio-TV programme guide (see Figure 7) like, e.g. the title of the programme, the country of origin, the year of production etc.

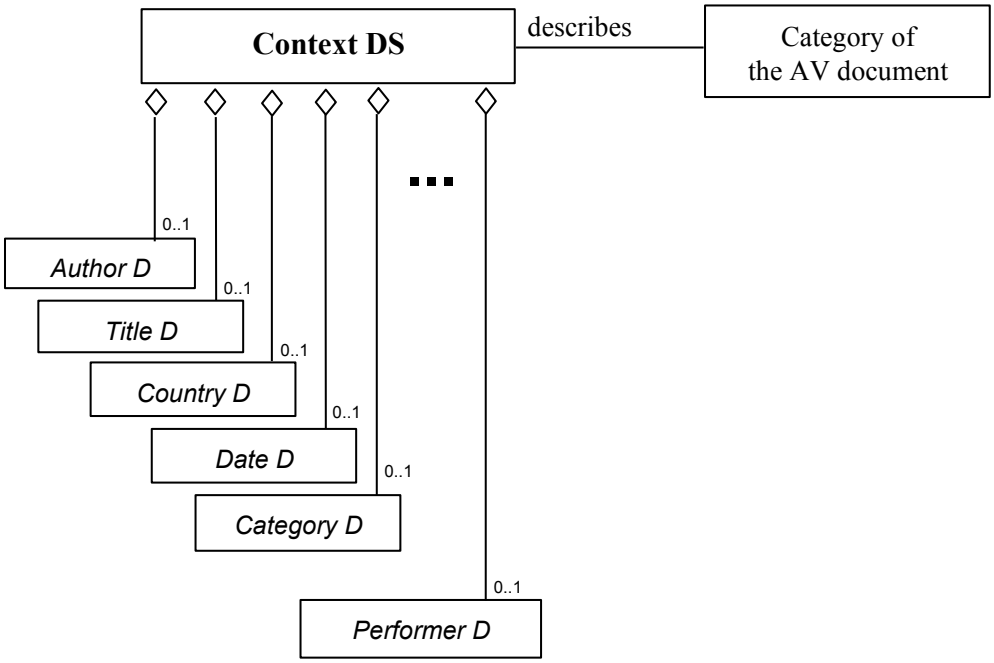

Figure 7: ***Structure of the Contextual DS***

EXAMPLE OF XML IMPLEMENTATION

```
<!--The  TOCAI DS  -->

<tocai_ds>
      <contextual_ds> ... </contextual_ds>
      <table_of_context_ds> ... </table_of_context_ds>
      <analytical_index_ds> ... </analytical_index_ds>
      <meta_descriptors_ds> ... </meta_descriptors_ds>
</tocai_ds>


<contextual_ds>
      <author> ... </author>
      <title> ... </title>
      <country> ... </country>
      <date> ... </date>
       ...
      <performer> ... </performer>
</contextual_ds>

<table_of_context_ds>
      <audio-visual_structure_ds> ... </audio-visual_structure_ds>
      <audio_structure_ds> ... <audio_structure_ds/>
</table_of_context_ds>


<audio-visual_structure_ds>
      <time-code> ... </time-code>
      <scene_ds> ...
            <scene_ds> ...
                <shot_ds> ... </shot_ds>
                <shot_ds> ... </shot_ds>
                 ...
                <shot_ds> ... </shot_ds>
            </scene_ds>
             ...
            <scene_ds> ...
                <shot_ds> ... </shot_ds>
            </scene_ds>
      </scene_ds>
</audio-visual_structure_ds>



<scene_ds>
      <scene_type> ... </scene_type>
      <time_interval> ... </time_interval>
      <iconographic_scheme> ... </iconographic_scheme>
       ...
      <visual_object> ... </visual_object>
</scene_ds>
```

```
<shot_ds>
      <editing_effect_ds> ... </editing_effect_ds>
      <k-frame_ds> ... </k-frame_ds>
      <mosaic_ds> ... </mosaic_ds>
       ...
      <outlier_ds> ... </outlier_ds>
</shot_ds>
```

```
<audio-visual_object_ds>
      <ordering_keys> ... </ordering_key>
       ...
      <ordering_keys> ... </ordering_key>
      <key-scene_class_ds> ...
            <scene_type> ...
          <scene_graph_representation>
               <scene_context_ds> ... </scene_context_ds>
     <pointer_type> ...
                              <reference_pointer> ... </reference_pointer>
                              <reference_pointer> ... </reference_pointer>
                        </pointer_type>
                 ...
                        <pointer_type> ...
                              <reference_pointer> ... </reference_pointer>
                               ...
                              <reference_pointer> ... </reference_pointer>
                        </pointer_type>
                 </scene_graph_representation>
           ...
            </scene_type>
             ...
      </key-scene_class_ds>
       ...
      <key-image_class_ds> ...
            <mpeg4_visual_object> ...
          <mpeg4_visual_object_ds> ... </mpeg4_visual_object_ds>
          <pointer_type> ...
                        <reference_pointer> ... </reference_pointer>
                 </pointer_type> ...
           ...
                 <pointer_type> ...
                        <reference_pointer> ... </reference_pointer>
                         ...
                        <reference_pointer> ... </reference_pointer>
                 </pointer_type>
           ...
            </scene_type>
             ...
      </key-image_class_ds>
</audio-visual_object_ds>
```

```
<meta-descriptor_ds>
      <name_of_dsc_provider> ... </name_of_dsc_provider>
      <type_of_involved_extaction_method> ... </type_of_involved_extaction_method>
      <reliability_of_extaction_method> ... </reliability_of_extaction_method>
       ...
      <size_of_description> ... </size_of_description>
</meta-descriptor_ds>
```

APPLICATION EXAMPLE

The TOCAI seems very adequate to describe the content of a large AV programme such as a movie. The TOC allows to navigate at different levels of details (scene or shot), while the AI gives the possibility to retrieve individuals or specific backgrounds present in the movie.

The TOCAI DS can satisfy queries like the followings:

- "I want to have a quick (and/or more detailed) idea about the content of such AV document viewing a frame (and/or listening to an audio sample) for each of the most representative scenes".
- "I want to see a list of the main objects of the AV document and select the scenes where they occur".

Description of TOCAI demo

In order to show how the TOCAI DS can be used, a demo application has been developed. The application consists of an interactive console that allows to efficiently browsing an audio/video document. During the navigation of the *Table of Content* (TOC console) the browser is always associated either to a shot or to a scene. The user can move through contiguous elements or can go from the shot layer to the scene layer and back. Moreover, by a simple mouse-click, the element currently selected is played. The developed application allows changing from the TOC-view to the AI-view, where the main objects of the document are ordered according to a certain criteria, e.g. the apparition order, the mean hue of the object and so on. In the AI console each object is connected to all the shots in which it appears. Through these connections the user can switch to the TOC-view.

In the following, it is described how the demonstration matches some evaluation criteria.

Effectiveness: the demonstration highlights the two main purposes of the TOCAI DS: time sequential characterization and object indexing.

Application domain: the application shows how the proposed DS allows an efficient browsing (and indexing) of any AV document.

Abstraction at multiple hierarchically levels: the demo has been developed according to show the hierarchical nature of the DS.

Flexibility: the demo exploits the proposed descriptors but it can be eventually extended if any new descriptor will be added. Moreover the demo would work in similar manner with different Ds.

Scalability: the size of the AV documents does not affect the performance of the application.

Simplicity: The demo will show the performance of the DS implemented with a limited number of Ds.

REQUIREMENTS

General Requirements

1. This DS addresses the following features: annotations, spatio-temporal structure, production features and composition information.
2. **Abstraction levels for multimedia material -** Yes. The TOC is naturally based on a hierarchical representation of the information.
3. **Cross-modality -** Yes. Audio and visual data are used to identify scenes (AV data).
4. **Multiple descriptions -** Yes: this DS can be applied to multimedia material at several stages of its production.
5. **Description scheme relationships -** Yes: the Ds involved in TOCAI may be used in other description schemes as well.
6. **Feature priorities –** They can be defined, but the priorities should be listed according to a specific application context.
7. **Feature hierarchy -** Yes: the features dealing with temporal structure of document are hierarchically represented (e.g. shot, scene).
8. **Descriptor scalability –** In the sense of the requirement document (N2461), the TOCAI DS involves Ds at different levels which enable refinement of the description, thus giving access to information from abstract to fine levels.
9. **Description schemes with multiple levels of abstraction -** Yes: e.g. the scene level presents an higher level of abstraction with respect to the shot level.
10. **Description of temporal range -** Yes. This DS gives a precise characterization of the temporal structure of the AV material.
11. **Direct data manipulation -** Yes. This point may be addressed by the AI by means of mosaicing.
12. **Language of text-based descriptions -** Yes.
13. **Translations in text descriptions -** Yes.

Functional Requirements

1. **Content based retrieval –** Indeed, the AI being contructed according to content. The ordered set of indices contained in the AI can easily be used so as to select a content of interest.
2. **Similarity-based retrieval –** Yes, this can be achieved both by using descriptors available from the TOC part and from the AI part of the proposed DS. Any query-by-example where each example contains a TOCAI description, can be used to retrieve similar multimedia material, at least according to some feature. For example, the relative frequency of certain scene types (which can be extracted from Ds contained in the TOC) is useful to identify a certain programme category, and this can be used to answer queries which try to identify programme using a query-by-example approach. On another hand, a query related to the presence of a specific object may be answered by comparison with audio-visual objects found in the AI.
3. **Associated information -** Yes the TOCAI can incorporate any type of associated information provided it is in digital format.
4. **Streamed and stored descriptions -** Yes: some kinds of involved Ds are streamed (e.g. editing effects) while others are not (e.g. reliability level).
5. **Distributed multimedia databases -** Why not.
6. **Referencing analog data -** Yes: provided that hand-made extraction methods are adopted.
7. **Interactive queries-** This can be handled provided an adequate interface is designed.
8. **Linking -** Yes. For example, scenes can be located in time and main objects can be located in

space and time.

9. **Priorisation of related informations –**
10. **Browsing -** Yes. This DS is very attractive for efficient multimedia browsing, especially if one considers the TOC framework.
11. **Associate relations -** Yes: there are relations between different descriptors
12. **Interactivity support -** Not foreseen.

EVALUATION CRITERIA

Effectiveness: The TOCAI DS produces the intended result: time sequential characterization OF AV documents, together with an index of the main objects.

Application domain: It is clearly applicable for a wide range of applications since every AV document can be viewed like a time-sequential event.

Comprehensiveness: it is immediate; just consider the proposed analogy with the current organization of a book.

Abstraction at multiple hierarchical levels: This DS is naturally based on various hierarchical levels.

Flexibility: Yes: it is possible to add new descriptors or substitute some of the current descriptors.

Extensibility: Yes: we have shown examples of the extension of this DS to other DSs.

Scalability: A larger number of data implies a larger number of shots (scenes) and objects. Thus except for the larger amount of required storage memory, the DS is scalable. Moreover, most descriptor values can be assigned independently of the initial AV document resolution.

Simplicity: Yes, because the number of involved Ds is not large.