JOANNEUM
RESEARCH

MPEG-7 Detailed Audiovisual Profile

Description of the Profile

DATE Oct. 31, 2007

ABSTRACT Description of the MPEG-7 Detailed Audiovisual Profile (DAVP).

AUTHOR, COMPANY Werner Bailer (JRS), Peter Schallauer (JRS), Helmut Neuschmied (JRS)

KEYWORDS MPEG-7, audiovisual, content description
RELATED ITEMS

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
0.1	2004-11-20	First Conceptual Draft	Living	Confidential
0.3	2005-04-12	Updated	Living	Confidential
0.4	2005-04-29	Reorganized profile description.	Living	Confidential
0.5	2005-05-23	Minor updates	Living	Confidential
0.51	2006-04-10	Minor corrections	Living	Confidential
0.6	2007-10-31	Clarified some definitions	Living	Public

Content

1 Introduction

1.1 About this document

This document is a supplement to the XML schema that defines the Detailed Audiovisual Profile (DAVP). It describes the context and the basic rationale for the definition of the profile and it describes the tool selection as well as the tool and semantic constraints of the profile. In [3] the constraints on semantics are the third step of the tool selection process, they specify the use of the selected description tools. The selected tools and some of the tools constraints are formalized in the XML Schema defining DAVP, some of the tool constraints and the semantic constraints are only described in textual form in this document.

1.2 Using MPEG-7

One of the strengths of MPEG-7 is its flexibility, which is provided by a high level of generality. It makes MPEG-7 usable for a broad range of application areas and does not impose too strict constraints on the metadata models of these applications. However, in the practical use of MPEG-7, two main problems arise from these features: complexity and hampered interoperability. The MPEG-7 requirements group has early recognized these issues [1].

The complexity arises from the use of generic concepts, allowing deep hierarchical structures, the high number of different descriptors and description schemes and their flexible inner structure, i.e. the variability concerning types of descriptors and their cardinalities. This complexity makes MPEG-7 difficult to learn and may thus sometimes cause hesitance in using the standard in products. It also makes it more difficult to implement tools for working with MPEG-7, and a lack of tools and implementation can contribute to the hesitance mentioned before. Moreover the complexity of the constructs increases the documents in size, especially when serialized using XML.

The interoperability problem emerges from the openness in the definitions in the standard. There can be several standard conformant ways to structure and organize descriptions which are similar or even identical in terms of content. While conformance and interoperability can be checked on a level of used description schemes and descriptors and their structure, interoperability on a semantic level is not fully guaranteed by the standard. This means, that standard conformant MPEG-7 documents can only be understood correctly with the knowledge of how the standard has been used when creating the description. This means that an additional layer of definitions is necessary to enable full interoperability between systems using MPEG-7.

1.3 MPEG-7 Profiles

1.3.1 The concept of Profiles and Levels

Recently, profiling has been proposed to partially solve these problems [3]. Based on the experience from other MPEG standards the means proposed are profiles and levels. Profiles are subsets of MPEG-7 tools which cover certain functionalities, while levels are further restrictions of profiles in order to reduce the complexity of the descriptions. However, the two means are related, and profiles will also significantly contribute to the reduction of complexity, as they will influence some of the complexity measures proposed in [1], e.g. the number of descriptors and description schemes used.

The following three main steps for defining a profile are proposed in [3]:

1. Selection of tools supported in the profile, i.e. the descriptors and descriptions schemes used.
2. Constraints on the selected tools, e.g. reduction of cardinality of some elements.
3. Constraints on the semantics, i.e. specifying the use of the tools in the profile.

If a profile is just defined according to the first two steps, conformance to the profile can only be ensured on the level of the structure of descriptors and description schemes. Without doubt this kind of conformance is important on a technical level and it is the only type of conformance for which validation tools are available today.

We think that the third step is a crucial one for defining a profile. Describing the semantic constraints of MPEG-7 profiles are required for making descriptions interoperable between application and systems. Without sufficient semantic constraints mappings between different profiles cannot be defined automatically. If semantic constraints are only defined in textual form, as it is currently done for MPEG-7 profiles, conformance to a profile in terms of semantics cannot be checked automatically and mappings between different profiles can only be defined manually.

1.3.2 Profiles and Interoperability

The introduction of profiles and levels aims at defining subsets of the standard that cover certain functionalities and reducing the complexity of descriptions. An important benefit of the definition of profiles is the support of interoperability for different systems using MPEG-7 in an application area. Interoperability is a key requirement for a widespread use of MPEG-7 and thus the most important reason for the definition of profiles.

Profiles and levels are suitable tools to increase the interoperability between systems by providing a clear definition of the semantics of the elements in the description. To achieve this goal, interoperability has to be considered in the process of profile definition, mainly by avoiding ambiguities in the description and by restricting the use of the standard, so that there is only one way to model a certain kind of description. The first step in the profile definition process, the selection of tools, is only of limited relevance in this context, as it mainly limits the functionality of the description. The definition of constraints on the selected tools is in some cases sufficient to formulate the intended restrictions, but only a limited set of constraints can be expressed with this tool.

Due to limitations of XML Schema, not all constraints on tools can be formalized in the XSD of the profile. Also, the semantic constraints are not within the scope of the schema definition. However, the semantic constraints are those that describe the use of the standard most precisely and are most powerful in avoiding ambiguities. This means that the majority of those elements of the profile definition that ensure interoperability have to be specified in textual form, due to the lack of a way of formalization in XML Schema. While the textual representation allows nearly unlimited expressive capabilities it does not allow checking for validations of these constraints by automatic validation against the profile.

1.3.3 Adopted Profiles in Part 9

Several profiles have been under consideration for standardization and in [2] a set of proposed and a set of adopted profiles have been collected. The definitions of the three adopted profiles have been revised in [4] and will constitute part 9 of the standard.

The Simple Metadata Profile (SMP) allows describing single instances of multimedia content or simple collections. The profile contains tools for global metadata in textual form only. The proposed Simple Bibliographic Profile is a subset of SMP. Mappings from ID3, 3GPP and EXIF to SMP have been defined.

The functionality of the User Description Profile (UDP) consists of tools for describing user preferences and usage history for the personalization of multimedia content delivery.

The Core Description Profile (CDP) allows describing image, audio, video and audiovisual content as well as collections of multimedia content. Tools for the description of relationships between content, media information, creation information, usage information and semantic information are included. The profile does not include the visual and audio description tools defined in parts 3 and 4.

The profiles that are defined in part 9 of the standard [4] will not be sufficient for a number of applications. If an application requires additional description tools, a new profile must be specified. It will thus be necessary to define further profiles for specific application areas. For interoperability it is crucial, that the definitions of these profiles are published, to check conformance to a certain profile and define mappings between the profiles.

1.3.4 The Need for a Profile for Comprehensive Description of Audiovisual Content

The application area we are considering is that audiovisual production, archiving, search and retrieval and media monitoring. In this area, the main functionality is that of describing image, audio, video and audiovisual content. Other functionalities can be considered out of scope. For the description of the audiovisual content types listed above, there is the requirement for having a detailed description that includes:

Structural description of the content: The description must allow arbitrary fragments of media items. The scope of a description may vary from whole media items to small spatial, temporal or spatiotemporal fragments of the media item. The definition of these fragments must be flexible enough, to allow fragments that are based on audiovisual features (such as image regions representing objects or shots of a video), any higher-level features (e.g. scenes in a video) or manually defined by an annotator. This includes descriptions of different kinds of modalities, descriptions produced with different tools, such as results from automatic content-analysis, semantic interpretation and manual annotation. The latter are mainly in textual form, but it is nonetheless beneficial to structure these instead of having simple free text annotations.

Description of visual and audio features and signal properties: In many search and retrieval systems, query by example is an important query paradigm. As a prerequisite, the content description must include visual/audio feature descriptions. These feature descriptions may also be required for semantic information extraction algorithms. Many approaches rely on the low- and mid-level features that can be extracted automatically from audiovisual content. Especially in the archive application area, the description of the condition of the audiovisual material (e.g. using the audio signal quality descriptor defined in [5]) is an important requirement.

Media, creation and usage information: These kinds of metadata, which are usually global in the sense that they refer to a complete content, are commonly used in the envisaged application areas and often the only ones available in legacy metadata information.

Summaries: Efficient browsing, visualization and sonification of descriptions of multimedia content is an important requirement in many applications. Summaries, used in connection with the full content descriptions, are a very valuable tool for this purpose.

The Core Description Profile (CDP) is the one that has most overlap with these requirements and already fulfills some of them. However, for a detailed description of audiovisual content, further tools are necessary, most prominently, the visual and audio descriptors from part 3 and 4. From this point of view, DAVP could be seen as a more complex level of CDP. An intention in the definition of DAVP is to be complementary to metadata standards and container formats that are restricted to global technical and descriptive metadata.

An important requirement is that of interoperability, as already discussed in Section 1.3.2. The current definition of CDP contains some constraints of the tools included in the profile, but no semantic constraints that describe restrictions going beyond the XML Schema definition. The definition of a detailed audiovisual profile that ensures interoperability between systems using the profile, must contain the semantic constraints that define the unambiguous use of the functionalities of the standard required in this application area.

2 DAVP Overview

2.1 Application areas

This profile supports detailed content description of image, audio, video and audiovisual content. It provides tools for describing global technical and descriptive metadata, spatiotemporal structure, semantics and audiovisual features of the content.

It is intended for a broad range of applications that deal with the analysis, description, retrieval, summarization and exchange of audiovisual content. The profile is defined to support the use of a variety of automatic content analysis tools and content-based query paradigms such as query by example.

Application areas include:

- audiovisual archives
- image and video databases
- media monitoring applications
- audiovisual content production
- educational applications

2.2 Functionality

2.2.1 Functional Overview

The Detailed Audiovisual Profile (DAVP) covers the functionalities to describe video, audio and still image content. The requirements of archiving, search and retrieval and media monitoring systems on a comprehensive description are considered in the profile.

DAVP includes tools for:

- the description of image, audio, video and audiovisual content
- the description of metadata of these descriptions
- the description of the spatial, temporal and spatiotemporal structure of the types of content listed above
- the description of media information
- the description of creation and production information
- the description of semantic information
- the description of visual and audio features
- the summarization of image, audio, video and audiovisual content

2.2.2 Parts of the Standard Included in DAVP

2.2.2.1 Visual and Audio descriptors (Part 3 and 4)

Part 3 and 4 have been fully included into DAVP, so that all visual and audio descriptors may be used, whenever defined by the MDS tools.

The reason for this decision is that we believe that a comprehensive description of audiovisual media must include low- and mid-level feature descriptions. Within the application scope of DAVP, there are two main uses for this kind of descriptions:

Similarity search: In many search and retrieval systems, query by example is an important query paradigm. The prerequisite for query by example is that the required feature descriptions have been extracted for all media items in the set being searched and for the query. If the media description shall be self-contained, it must include the feature descriptions.

Semantic reasoning: In an ideal world, audiovisual content analysis would extract high level information which is semantically meaningful to a user. It is however in many cases not possible to directly infer high-level information form the audiovisual content. Semantic information extraction algorithms are capable of inferring high-level information from a set of media descriptions and often additional external sources. Many approaches rely on the low- and mid-level features that can be extracted automatically from audiovisual content. Thus the description of content analysis results must include visual and audio feature descriptions, so that they can be used later for inferring semantic information.

The support of visual and audio descriptors is a main difference to the adopted Core Description Profile (CDP) and the proposed Bibliographical Simple Profile [2].

The inclusion of all visual and audio descriptors in DAVP has the consequence, that all MDS tools, from which visual and audio description tools are derived or which are used by them must be included in DAVP.

2.2.2.2 Multimedia Description Schemes (MDS, Part 5)

Many MDS are included in DAVP. All exclusions of tools and tool restrictions described in the following sections of this document apply to MDS tools. Sections 3.1 and 3.2 discuss the MDS tools included and excluded in DAVP, respectively and illustrate the rationale for doing so.

2.3 Structure of the Description

A basic principle that we followed is that only one audiovisual content is described per MPEG-7 description. The profile allows the description of audiovisual content and still images (mutually exclusive, as otherwise the constraint of describing one content would be violated).

The profile allows the use of all spatiotemporal structuring tools. In the case of an audiovisual media item, the description is split into the parts describing the visual and audio related features only. This also includes structuring information, which is only based on these features, e.g. shots in the visual part, audio segments delimited by silence the audio part. All structuring information and other annotation, which relates to the audiovisual information (e.g. scene information derived from analysis of both visual and audio features), is attached directly to the top-level audiovisual segment. In the visual and audio part of the description, the use of the respective low-level descriptors (defined in part 3 and 4 of the MPEG-7 standard) is allowed.

The description of image, audio, video or audiovisual content may optionally contain a summarization of this content description.

An overview diagram showing the structure of DAVP description is shown in Figure 1.

The following sections describe the tool selection and constraints and the semantic constraints of DAVP.

MPEG-7 DETAILED AUDIOVISUAL PROFILE

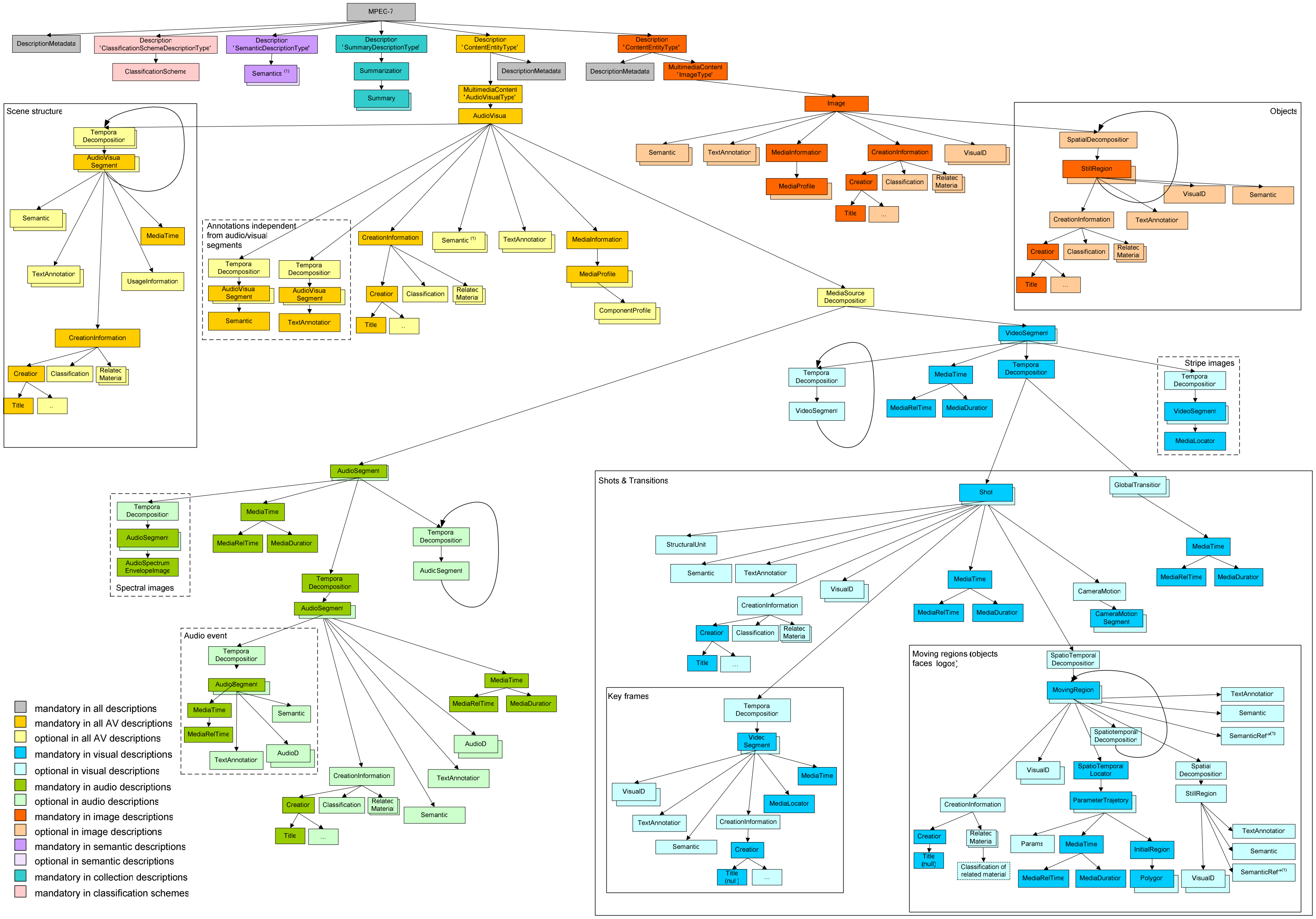

Figure 1: Structure of MPEG-7 DAVP descriptions.

3 Tool Selection and Constraints

In the following sections, we describe the tools which are included in the profile and those which have been considered out of scope of this profile. The last part of this section describes the formal constraints on the selected tools. The semantic constraints of the profile can be found in Section 3.3.4.1.

3.1 Included Tools

3.1.1 Top-level Types

3.1.1.1 Descriptions

The scope of DAVP is restricted to the description of audiovisual media items, with the intention to have a one-to-one relation between audiovisual content items and MPEG-7 documents. The main type of description is thus ContentEntityType for describing audiovisual content.

Except for ContentEntityType only three types of descriptions may be used in DAVP:

ClassificationDescription: DAVP supports the use of classification schemes (cf. Section 4.8.1) as they facilitate exchangeability and interoperability of descriptions. Applications may need to access the referenced classification schemes, e.g. to obtain a name in the user's language for a term. To allow applications supporting DAVP to parse classification schemes, this type of description has been included in the profile.

SemanticDescription: Semantic descriptions can be used to describe the set of concepts, objects or agent objects appearing in audiovisual content. DAVP follows the paradigm of having one MPEG-7 description for an audiovisual content. There may however be a set of content, that shares some semantic entities. To avoid redundancies, they can be collected in a separated semantic description, and are referenced from the description of the content in which they occur.

SummaryDescription: For efficient visualization, sonification and browsing of the content being described, a summary of the complete content description may be optionally included. The contents of the summary are restricted to summarizing one audiovisual content.

The DescriptionMetadata type is included in DAVP.

3.1.1.2 Multimedia Content Entities

For ContentEntityType descriptions, two types of multimedia content are supported in DAVP:

Image: Description of a still image.

AudioVisual: Description of audio, video and audiovisual content. We have decided not to use Audio and Video content in order to unify the description of content in which either the visual modality or the audio modality or both are present. The type of content is described in the MediaInformation of the root audiovisual segment.

3.1.2 Basic datatypes, linking, identification and localization

Except for inline media definitions, all types defined in MDS are included.

3.1.3 Basic description tools

All types of textual annotation are included with the exception of DependencyStructure. The MDS tools for agents, places, graphs and relation are fully supported, especially for the use in semantic descriptions. Affective content descriptions are included in DAVP.

Classification schemes are supported as defined in MDS with the exception of GraphicalClassificationSchemes.

3.1.4 Media description tools

The media description tools are fully supported in DAVP.

3.1.5 Creation and production description tools

The creation and production description tools are fully included in DAVP.

3.1.6 Usage description tools

The usage description tools are supported except for the Financial type.

3.1.7 Structure description tools

The structuring tools are one of the most powerful tools in MDS. All tools for the structural description of audiovisual, audio, video and image segments have been included in DAVP. This also includes tools for describing moving regions, visual text in images and video, and most of the description tools for edited video segments.

3.1.8 Navigation and Access tools

Summarization tools have been included in DAVP to allow the description of a summary aside the complete detailed content description. The summary can be used for efficient browsing and navigation of the content description. Multiple summaries of one audiovisual content entity may be described.

Both hierarchical and sequential summaries and all types of summary segments and components are included in DAVP.

3.1.9 Visual Description Tools (Part 3)

All visual description tools are included.

3.1.10 Audio Description Tools (Part 4)

All audio description tools are included.

3.2 Excluded Tools

3.2.1 Top-level Types

All types of descriptions, that are not needed to represent a description of audiovisual content (such as views, variations, user preferences) have been excluded. All media types except of audiovisual and images have been excluded.

3.2.2 Basic datatypes, linking, identification and localization

The inline media type has been excluded from DAVP, as the content being described in the application areas is usually too large to be kept in the description itself. Moreover, detailed audiovisual description tend to result in large MPEG-7 documents, so that parsing is already a time-consuming task and the problem shall not be increased by adding binary encoded data to the document.

Representational media, such as key frames, are smaller than the content being described, but they are typically numerous, so that they also should be stored separately and only be referenced from the MPEG-7 document.

3.2.3 Basic description tools

For text annotations, the DependencyStructure type is excluded, as it is from our experience a rarely used type and it may lead too far for an audiovisual content description.

GraphicalClassificationScheme has been excluded, as it may be questioned if MPEG-7 is the right tool and the right context for this kind of descriptions. We recognize the need for controlled vocabularies, which can be expressed using simple classification schemes. For any descriptions, that include further relations between the terms, we advocate the use of an ontology representation like OWL [6].

Ordering tools have been excluded, the task of defining orders based on the descriptions is left to the application.

Phonetic transcriptions have not been included in DAVP to simplify the handling of textual annotations in DAVP. If phonetic transcriptions are required to represent results of automatic speech recognition, the SpokenContent description tools (defined in part 4) can be used instead of a textual annotation.

3.2.4 Structure description tools

Point of view type has been excluded as they include implications about the use of the content.

DAVP is limited to 2D descriptions, so the StillRegion3D and the dependent types have been excluded.

Tools for the description of multimedia, ink and handwriting segments have been excluded as the scope of DAVP is limited to the description of audiovisual and image content.

From the tools for the description of edited video segments those describing composition shots and the related transitions have been excluded. The description of edited video is restricted to global editing effects.

3.2.5 Navigation and Access tools

Partitions, Views and Variations have been excluded, as they are considered out of scope of DAVP.

3.2.6 Content organization tools

All types of collections have been excluded, as the scope of DAVP is the detailed description of audiovisual content entities. Thus content management and organization functionalities have been excluded.

All models, which are not required by audio descriptors of part 4 have been excluded.

3.2.7 User interaction tools

User interaction tools have been completely excluded from DAVP as they are considered out of scope. The tools serve for the description of personalization of systems and for the collection of records of user customs, but they need not be included in exchangeable descriptions of audiovisual content in the application area envisaged for DAVP. For this purpose, the User Description Profile (UDP) [4] is complementary.

3.3 Tool Constraints

This section describes the formal tool constraints. Those constraints, which are implied by the tool selection (i.e. prohibition of child elements because the type has been removed from the profile) are not listed here.

3.3.1 Top-level Types

3.3.1.1 Root element

In DAVP, only full descriptions are allowed. DescriptionUnits are prohibited.

3.3.1.2 Content Entity Type

The ContentEntityType in DAVP allows only one MultimediaContent element.

3.3.1.3 Description Metadata

Only one description metadata element is allowed on the root element and on DSType elements.

In the DescriptionMetadataType, the package element has been removed.

3.3.2 Basic description tools

3.3.2.1 Textual Annotations

Phonetic transcriptions may not be generally used in textual annotations but only where explicitly allowed in a specialized descriptor (e.g. SpokenContentType). The phoneticTranscription and phoneticAlphabet attributes have thus been removed.

3.3.3 Structure description tools

3.3.3.1 Segments

The use of the elements point of view and relation are prohibited.

All audiovisual, video and audio segments must have IDs so that they can be referenced.

Structural unit elements are required on all audiovisual, video and audio segments. At least the `href` attribute of the structural unit elements must be set. It is recommended to use the terms defined of the appropriate classification scheme listed in Section 4.8.1.

3.3.3.2 Segment Decompositions

The criteria attribute is mandatory on all segment decompositions. As this is a string and not a reference to a classification scheme. Values of the criteria attribute for segment decompositions defined by DAVP are listed in Section 5.2.1.

The MediaSourceDecomposition element is prohibited on MovingRegions.

3.3.3.3 Edited Video Segments

No area decomposition is allowed on AnalyticEditedVideoSegments. The use of EditedMovingRegion in decompositions of these segments is prohibited.

Only VideoSegmentTemporalDecomposition may be used for ShotType elements.

3.3.4 Navigation and Access Tools

3.3.4.1 Summarization

A summary description may only contain one summarization element. The summarization element may contain any number of summaries.

4 Semantic Constraints

4.1 Introduction

The tool selection and the tool constraints can only define the scope of functionality and the complexity of the profile. The semantic constraints are much more powerful as they define the use of the tools and thus are crucial for interoperability.

While DAVP allows nearly unconstrained use of the structuring tools on the level of tool constraints, most of the semantic constraints deal with the structure of the description. The main design criteria of these semantic constraints is to keep the description as modular as possible. This means that description fragments which stem from different sources or are based on different modalities are kept in different parts of the description. The same approach is used for descriptions on different levels of abstraction, such as descriptions based on features of the audiovisual content (e.g. shot boundaries) and descriptions on a higher level (e.g. scene structuring), which may have been generated using prior knowledge or other external sources.

The aim of these constraints is not to limit the flexibility of structural descriptions, but to organize them in a way, that theirs mutual dependencies are reduced as much as possible so that single decompositions can be accessed or modified without influencing others. Especially for annotations on a higher semantic level, such as textual or semantic annotations, there is still a very high degree of flexibility. For common types of decompositions, such as shot lists of video content and description of the associated key frames, DAVP defines a description structure that shall ensure the exchangeability of these common structural descriptions, while leaving other, additional structural annotation unaffected.

Other semantic constraints deal with the use of media information on different parts of the description structure. This is a key issue, as it is the link from the metadata description to the essence being described. Some of the technical metadata contained there is crucial for using the description along with the essence. Thus a clear definition of the use of this description tool is necessary.

Summarization tools are included in DAVP but their use is constrained by the intention to provide descriptions of single multimedia content entities. Summaries may be either used in connection with a detailed content description in the same document, or as stand-alone summaries. In both cases, the summary must be restricted to one audiovisual content entity.

Many of the semantic constraints are visualized in Figure 1.

4.2 Top-level Types

4.2.1 Descriptions

A basic principle of DAVP is to have one description in one MPEG-7 document. If this description is an audiovisual content description (ContentEntityType), only one multimedia content may be described. The content entity type element has either a AudioVisualType or a ImageType element (both are of type MultimediaContentType), each having one AudioVisual or Image element respectively.

In the case that the document contains a description of image or audiovisual content, a summary of this description may be included in the same document. In any other case, only one description may be contained in a document.

4.2.2 Audiovisual Content

The root element of the description will be of AudioVisualType to allow a common description of visual and audio of a multimedia item. The AudioVisualSegment at the top has 0 or 1 a MediaSourceDecomposition (criteria “modalities”), which contains 0 or more AudioSegments and/or 0 or more VideoSegments. If the MediaSourceDecomposition element is present, at least one segment must be contained. The root Audio- and VideoSegments must have the same start time and duration as the root AudioVisualSegment element.

Annotations concerning both visual and audio information, such as a scene structure generated by combining visual and audio hints, can then be stored as a temporal or spatiotemporal decomposition of the AudioVisualSegment with any recursive structure of AudioVisual and TemporalDecomposition elements below.

4.2.3 Description Metadata

Complete description metadata elements containing identifiers, date, version, tool used for modification, etc. may only be used on the MPEG-7 root element and on the description elements. The description metadata element of the MPEG-7 root element is required.

The description metadata element of the root element of an audiovisual content description must contain at least one PrivateIdentifier, which uniquely identifies the metadata description.

A description metadata element that only specifies the confidence of a description scheme, may be used as description scheme headers whenever the description scheme does not provide any other means to specify the confidence of the description.

4.3 Linking, Identification and Localization Tools

4.3.1 Media Time Description Tools

For time points within the media description both absolute and relative time points may be used. For relative time points, if the mediaTimeBase is not set, the relative time points will implicitly refer to the start time point of the root segment. Otherwise this attribute must be used to reference to an absolute time point.

4.4 Media description tools

A MediaInformation element must be present and attached to the top level AudioVisualSegment or to the Image element of a multimedia content description. The use of a stand-alone MediaLocator on the root audiovisual or image elements is prohibited.

The media information DS may only be used on the top-level Audiovisual or Image element. The media information element must contain at least one media profile. In the case that there is a profile, where the content is spread across several media instances, this must not be described in the structure description of content. The most appropriate way would be to described this inside the media instance DS, e.g. by allowing a list of sub-instances or at least (similar to component profiles) for the different media items. This is currently not possible in the standard, but the media instance descriptor should be extended accordingly.

If the modalities of the audiovisual content being described are stored separately, the media information must also be attached to the top-level audiovisual element. It must not be attached to the respective top-most audio and video segments, but component profiles of the joint audiovisual media profile must be used in this case.

4.4.1 Media Profiles

A media profile may contain component profiles, only if it is attached to an AudioVisual segment, not if it is attached to an Image segment.

A media profile may contain as many component profiles, as there are modalities in the multimedia content. Component profiles must not contain any further profiles.

Every media profile is required to contain at least on media instance element. There are however some exceptions of this basic rule:

- A media profile that contains component profiles does not need to have a media instance element, if each of the component profiles has at least one media instance element for each of the modalities.
- A component profile does not need to have a media instance element, if the parent profile contains at least one media instance elements and each of them refers to an instance that contains the modality being described in the component profile.

4.4.2 Media Locators

Stand-alone media locators of segments (i.e. instead of media information elements) may be used to reference representational essence (meta-essence) of audiovisual, video and audio segments (e.g. the images associated with key frames). In these cases, no media information elements may be used. They must not be used on the root audiovisual or image elements of a description.

4.5 Creation and Production Description Tools

A creation information element may be used on every video, audio or audiovisual segment. The content of the CreationInformation element will depend on its context.

A CreationInformation element is mandatory on the root audiovisual segment.

4.6 Structure Description Tools

4.6.1 Concept of structure description in DAVP

The basic principle of structure description in DAVP is to organize structural descriptions of audiovisual content according to modalities. This implies that all visual descriptions are within the visual part of the media source decomposition of the root audiovisual segment, the audio descriptions within the audio part and the jointly audiovisual descriptions directly attached to the root segment.

Structure descriptions in DAVP may be nested to arbitrary depth. However, nesting structures must be justified based on the occurrence of the feature. For example, the decomposition into key frames is nested under the decomposition into shots, as the key frames represent the visual content of the shot.

In the following sections, the description of some common structural decompositions of audiovisual content is discussed. These structures are grouped into those attached to the root audio and video segment respectively, those attached to the root audiovisual segment that those that may be attached to arbitrary segments.

4.6.2 Visual Segment Based Descriptions

4.6.2.1 Shot and Transition Structure

The shot structure is a decomposition of the root video segment. The temporal decomposition containing the video segments (shots and gradual transitions) will allow neither gaps nor overlaps, the criteria is “visual shots”.

Shots of a video are represented in DAVP using ShotDS.

Cuts are implicitly represented as two subsequent shots without a gap. Gradual transitions (dissolves, fades, wipes, etc.) are described using GlobalTransitionDS.

4.6.2.2 Key Frames

Key frames are described as a decomposition of a shot. The temporal decomposition containing the key frames must allow gaps, but no overlaps (the criteria is “key frames”).

Key frames are represented as VideoSegmentType elements with a time point and no duration. The VideoSegment describing the key frame optionally has a media locator referencing the still image representing the key frame.

4.6.2.3 Dominant (Camera) Motion

The dominant motion (camera motion) of a video segment is described with CameraMotion descriptor, which contains one or more camera motion segments. One descriptor may be attached to a shot.

Camera motion descriptions are only permitted on segments representing shots.

For the description of each of the segments, both mixture and non-mixture amount of motion type may be used. If no camera motion has been detected, no camera motion segment is attached to the shot.

4.6.2.4 Visual Objects

Moving regions are used to describe visual objects. The term visual objects refers to any kind of object that is detected or recognized in the visual domain (e.g. moving objects, faces, persons, etc.). We assume that visual objects have a spatial extent and, in the case of video content, exist at least for one point in time.

Depending on the temporal extent of the object appearance, either moving or still region DS are used.

In the case of video, object appearances are described on a per shot basis, as a shot boundary is a naturally boundary of a continuous object appearance. The still and moving regions are contained in a spatiotemporal decomposition of the shot. The association between appearances of the same objects in different shots can be done by referencing the entity being represented (e.g. by using the Semantic DS).

Moving Video Objects

Moving objects will be described using MovingRegionType. The region of a moving object is not required to be contiguous.

Moving regions will be described on a shot basis, i.e. the moving regions of a shot will be found in a spatiotemporal decomposition of the video segment describing the shot. The decomposition is allowed to have gaps and overlaps; the criteria is “moving objects”.

To describe the moving region, SpatiotemporalLocatorType is used. To describe the trajectory, ParameterTrajectoryType is used.

Faces

Faces will be described using a decomposition of a shot, where each segment represents a face. The decomposition of the shot segment containing the face segments must allow both gaps and overlaps, the criteria is “faces”.

The faces will be described using MovingRegionType, similar to moving video objects.

4.6.2.5 Other Decompositions

Any other decompositions, which are only based on visual features may be attached to visual segments.

4.6.3 Audio Segment Based Descriptions

4.6.3.1 Audio Segments

The basic unit of the audio description are audio segments. Audio segments are delimited by silence.

The temporal decomposition of the root audio segment must neither have gaps nor overlaps; the criteria is called “audio segments”.

4.6.3.2 Audio Classification

Classification of audio segments (into classes like speech, music, environmental sound, silence, etc.) is done on the basic audio segments.

The temporal decomposition containing the audio classification may have both gaps and overlaps.

4.6.3.3 Speech Segments

Audio segments classified as speech may contain a speech to text transcription (using SpokenContent description tools or TextAnnotation). If the ASR system produces a more fine grained output, the speech segment is decomposed into sub-segments (representing e.g. sentences or words) using a temporal decomposition. The decomposition may have gaps, but no overlaps.

4.6.3.4 Other Decompositions

Any other decompositions, which are only based on audio features may be attached to audio segments.

4.6.4 Descriptions Based on Audiovisual Segments

The structural descriptions attached to audiovisual segments are jointly created from the audio and visual modalities of the content. These descriptions usually correspond to high-level descriptions, that possibly include prior or domain knowledge. This means, that there may exist multiple descriptions for the same structure criteria, e.g. different scene structures based on different features.

4.6.4.1 Scene Structure

Decomposing audiovisual content into semantically meaningful scene or story units is a typical high-level content analysis task.

Scenes shall be described as audiovisual segments contained in a temporal decomposition that allows neither gaps nor overlaps (criteria “scenes”). Each scene or story segment may be further decomposed into sub-segments.

4.6.4.2 Events

Events are other high-level descriptions.

Events shall be described as audiovisual segments contained in a temporal decomposition that allows both gaps and overlaps (criteria “events”). The media duration element of the audiovisual segments describing events is optional.

4.6.4.3 Other Decompositions

Any other decompositions, which are jointly based on audiovisual features may be attached to audiovisual segments.

4.6.5 Descriptions Based on Arbitrary Segments

This group of descriptions may be either attached to any existing audiovisual, audio or video segment that may be the result of one of the structural descriptions discussed above.

To attach these descriptions to temporal segments that do not correspond to any of the segments found in the description, a separate temporal decomposition for each of the descriptions may be created. This description is attached to the root audiovisual segment, will allow both gaps and overlaps and has a criteria attribute indicating the type of description contained. For example, this may be used in an application where the user can attach textual annotation to any user-defined temporal segment.

4.6.5.1 CreationInformation/RelatedMaterial

Apart from the use of creation information discussed above, CreationInformation elements containing only a RelatedMaterial element may also be used in a stand-alone structure.

In the case that the descriptions are put in a stand-alone structure, the criteria “related materials” should be used for the temporal decomposition.

4.6.5.2 Text annotations

Text annotations which describe exactly one audio, video or audiovisual segment will be attached to the segment they describe.

Any other text annotation will be contained in an own temporal decomposition of the audiovisual root element, where there is a segment for each text annotation. The criteria attribute of this temporal decomposition should be “text annotations”.

4.6.5.3 Semantic annotations

Semantic annotations which describe exactly one audio, video or audiovisual segment will be attached to the segment they describe.

Any other semantic annotation will be contained in an own temporal decomposition of the audiovisual root element, where there is a segment for each semantic annotation. The criteria attribute of this temporal decomposition should be “semantic annotations”.

4.7 Semantic Description Tools

Currently there are no semantic constraints on semantic description tools in DAVP.

4.8 Navigation and Access Tools

4.8.1 Summarization

The use of summarization is permitted in DAVP to include a summary of a content, which is being described in detail, in the same MPEG-7 document. The stand-alone use of summaries is permitted in DAVP, but the summary is also restricted to summarizing one audiovisual content entity.

The intended use of summaries in DAVP is to facilitate browsing, visualization and sonification of a multimedia content description. Thus the summary must not be used to summarize a set of multimedia content entities.

5 Controlled Vocabularies

This section describes the controlled vocabularies to be used with DAVP. This includes the classification schemes defined in MDS, the extensions of the classification schemes defined below and the recommended values for string attributes.

5.1 Classification Schemes

Whenever possible, the classification schemes defined in the standard are used. In those cases, where the classification schemes in standard were not sufficient, new classification schemes have been defined. They are listed in Table 1. In many cases, the newly defined classification schemes import a classification scheme defined in the standard. For compatibility reasons, the terms coming from the original classification scheme should be referenced with its original URI and only newly defined terms with the new URI.

The new classification schemes will be defined using the experimental URN [7] urn:x-mpeg-7-davp.

Domain	Classification scheme URI	imports
//MediaInformation/MediaProfile/MediaFormat/Content	urn:x-mpeg-7-davp:cs:ContentCS:2005	urn:mpeg:mpeg7:cs:ContentCS:2001
//MediaInformation/MediaProfile/MediaFormat/FileFormat	urn:x-mpeg-7-davp:cs:FileFormatCS:2005	urn:mpeg:mpeg7:cs:FileFormatCS:2001
//MediaInformation/MediaProfile/MediaFormat/Medium	urn:x-mpeg-7-davp:cs:MediumCS:2005	urn:mpeg:mpeg7:cs:MediumCS:2001
//MediaInformation/MediaProfile/MediaFormat/VisualCoding/Format	urn:x-mpeg-7-davp:cs:VisualCodingCS:2005	urn:mpeg:mpeg7:cs:VisualCodingCS:2001
//VideoSegment/StructuralUnit	urn:x-mpeg-7-davp:cs:StructuralUnitCS:2005	

Table 1: Classification schemes in DAVP.

5.2 String Attributes

5.2.1 Segment Decomposition Criteria

Each segment decomposition element in DAVP is required to have a criteria attribute. As this attribute is a string and not a reference to a classification scheme, a list of string values for defined decompositions is given in Table 2. Decompositions with other criteria may occur, however, for a decomposition with the semantics described below, the defined criteria must be used.

Semantics	Type of decomposition	Criteria
decompose root audiovisual segment into visual and audio part	MediaSourceDecomposition	modalities
decompose root visual segment into shots and transitions	TemporalDecomposition	visual shots
decompose root audio segment	TemporalDecomposition	audio segments
decompose shot into key frames	TemporalDecomposition	key frames
decompose shot into moving objects	TemporalDecomposition	moving objects
decompose shot into faces	TemporalDecomposition	faces

Semantics	Type of decomposition	Criteria
decompose still image into objects	SpatialDecomposition	objects
decompose still image into faces	SpatialDecomposition	faces
decompose root segment into scenes	TemporalDecomposition	scenes
decompose root segment into events	TemporalDecomposition	events
decomposition holding text annotations on arbitrary segments	TemporalDecomposition	text annotations
decomposition of audio segment into speech segments	TemporalDecomposition	speech segments
decomposition holding semantic annotations on arbitrary segments	TemporalDecomposition	semantic annotations
decomposition holding related material annotations on arbitrary segments	TemporalDecomposition	related material

Table 2: Recommended values of criteria attributes of selected decompositions.

6 References

[1] ISO/IEC JTC 1/SC 29/ WG 11 N4039: MPEG-7 Interoperability, Conformance Testing and Profiling, Mar. 2001.

[2] ISO/IEC JTC 1/SC 29/ WG 11 N6039: MPEG-7 Profiles and Levels under Consideration, Oct. 2003.

[3] ISO/IEC JTC 1/SC 29/ WG 11 N6079: Definition of MPEG-7 Description Profiling, Oct. 2003.

[4] ISO/IEC JTC 1/SC 29/ WG 11 N6263: Study of MPEG-7 Profiles Part 9 Committee Draft, Dec. 2003.

[5] ISO/IEC, Multimedia Content Description Interface, Part 4: Audio, ISO/IEC 15938-4:2002/Amd 1:2004.

[6] OWL Web Ontology Language, W3C Recommendation, URL: http://www.w3.org/TR/owl-features.

[7] L. Daigle, D. van Gulik, R. Iannella, P. Faltstrom, *URN Namespace Definition Mechanisms*, RFC 2611, June 1999.