**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N13869**

**August 2013, Vienna, Austria**

Title: Text of white paper on MPEG-7 AudioVisual Description Profile (AVDP)

Status: Final

Source: Communication

Editor: Werner Bailer (JOANNEUM RESEARCH, AT)

White Paper on the MPEG-7 AudioVisual Description Profile (AVDP)

# Summary

The intention of MPEG-7 AudioVisual Description Profile (AVDP) is to facilitate the introduction of automatic information extraction tools in media production, e.g. as web services in Service Oriented Architectures (SOA), by providing a common format for the exchange of the metadata they generate. AVDP is a profile (i.e., subset) of the MPEG-7 Multimedia Description Interface standard, targeting applications in media production and archiving, and includes all tools provided by Part 3 and Part 4 of MPEG-7 [4][6].

As a result of an extended requirements analysis mainly conducted inside EBU (European Broadcasting Union), the description tools in this profile can be used to describe the results of various kinds of media analysis such as shot/scene detection, face recognition/tracking, speech recognition, copy detection and summarization, etc. in a way that these data can be usefully integrated in media production processes. The AVDP profile supports temporal and spatial analysis of audiovisual material, including low-level audio and video descriptions. The profile defines a set of semantic constraints in order to facilitate interoperability.

# Overview

This section gives an overview of AVDP in the context of MPEG-7, of which it is a profile, and describes the scope and functionality of the profile.

## MPEG-7

MPEG-7, formally named “Multimedia Content Description Interface”, is a standard for describing the multimedia content data that supports a comprehensive range of descriptors and descriptor schemes ranging from low level audio and video features to high level semantic features. Due to its intrinsic complexity, MPEG-7 is allowing some degree of not unambiguous interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code, as the same abstract information structure could be represented in several – all compliant - manners.

To overcome this issue, in specific application domains MPEG-7 profiles can be defined, which constraint the meaning and interpretation of the MPEG-7 standard descriptions in order to achieve better (ideally perfect) interoperability. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes support as broad a range of applications as possible. For more details on MPEG-7, please refer to the MPEG-7 White Paper at <http://www.chiariglione.org/mpeg>.

## AVDP in the context of MPEG-7

AVDP is included in MPEG-7 part 9, called “Profiles and Levels” [1]. The XML schema specifying AVDP is included in part 11 “Profile Schemas” [2]. Unlike the other MPEG-7 profiles, i.e. SMP (Simple Metadata Profile), UDP (User Description Profile) and CDP (Core Description Profile), AVDP is based on version 2 (2004) of MPEG-7, and includes all low-level visual and audio descriptors defined in parts 3 (visual [4,5]) and 4 (audio [6,7]) of the standard. The constraints on description tools in AVDP concern those defined in part 5 (Multimedia Description Schemes, MDS [8,9]) of the standard, restricting AVDP documents only to complete content descriptions and summaries. A number of constraints are aimed at improving interoperability, by limiting the degree of freedom in choosing and combining description tools, and enforcing the use of elements and attributes that fix the semantics of elements in the description.

## Scope and functionality

The main scope of AVDP is describing the results of automatic media analysis with low-, mid- and high-level features for audiovisual content. Thus, the profile includes functionalities for representing results of several – (semi-) automatic or manual – feature extraction processes.

AVDP provides tools for describing:

* feature extraction tool, version , contributors, and the date/time the tool applied
* several results in multiple timelines
* several results with confidence level
* various results of multimedia analysis, such as segmentation, text recognition, face detection, person identification, format detection, genre detection, keyword extraction, and speech recognition.
* results of multimedia analysis having several related descriptions of audio-visual contents, such as copy detection and summarization

# AVDP Structure and Semantics

Like any other MPEG-7 profile, AVDP defines a subset of the complete MPEG-7 standard for a specific application domain by selecting a subset of description tools (description schemes and descriptors) and possibly limiting the cardinalities of elements. In addition, AVDP is the first profile to define a set of semantic constraints on the description tools in the profile. The constraints aim at improving interoperability by enforcing clearly declared semantics of the elements in the description, by requiring type identifiers in all structural components and a modular structure of the descriptions. The modularity is based on

* separating metadata produced by different tools,
* separating metadata on different abstraction level (from low-level extracted content to inferred by external knowledge),
* separating metadata specific to one modality or valid for all, and
* separating content segmentation and representative elements (e.g. shots and key frames).

The top-level structure for content description supported by AVDP considers three cases. The most basic case is used to describe a single audiovisual content (Fig. 1a). The second is used to describe several contents and their relations such as similar segments or copies (Fig. 1b). The third describes the result of summarization and the related audiovisual contents (Fig. 1c). Other description types, such as collections or individual description schemes, have been excluded in order to have a clear and simple top level structure, and in order to fulfil the majority of identified use cases.

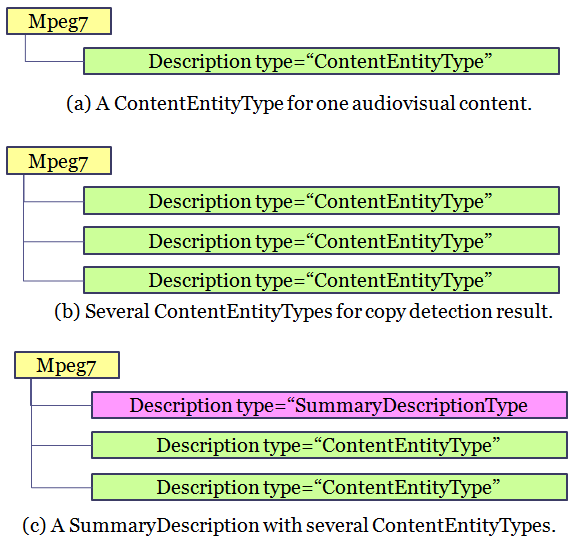


Figure 1: Top-level structure of AVDP documents.

Fig. 2 visualizes the high level structure for content description supported by AVDP. The root audiovisual segment (AudioVisual) expresses a whole content to which automatic media analysis is applied, and it is decomposed into AudioVisualSegments (AVS) by temporal decompositions (TD). Each temporal decomposition corresponds to the result of an analysis/annotation process. If specific audio and/or video descriptors are needed, the AVS segments can be decomposed by means of media source decomposition (MSD) into video and audio segments (VS and AS), to which the descriptors are attached. These VS/AS segments must have the same duration as the AVS. There can be as many such video/audio segments (VS/AS) as there are video/audio channels. Further recursive temporal decompositions of the AVS can be added. There can be several MSD of an audiovisual segment, e.g. one into modalities and one into key elements (key frames, key audio clips). Results of feature extraction from the complete segment may be added to the video or audio segment. Results of feature extraction from single frames within the segment or regions in single frames must be described using StillRegionType elements contained in the MSD of the segment and optionally further spatial decompositions of the still region. A still region element directly in the MSD must represent a full frame, those in the decompositions below may refer to regions (e.g. face regions, image text). Results of feature extraction from the complete segment or a subsegment, which create region information with a temporal extent of more than one frame must be described as moving regions in a spatiotemporal decomposition of the video segment.

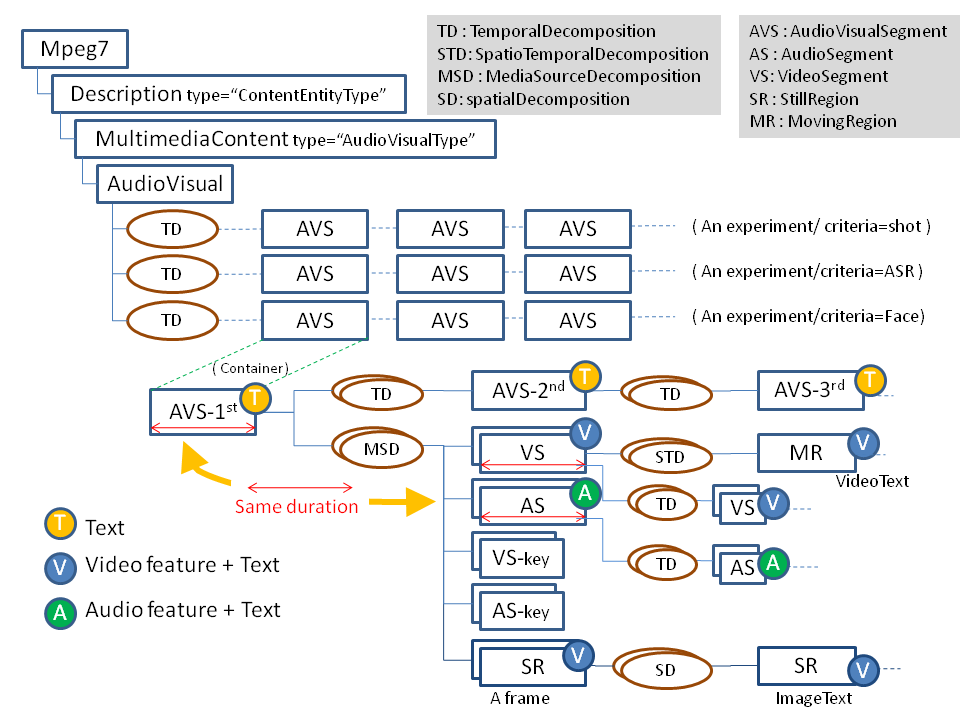


Figure 2: Basic structure of AVDP descriptions.

# Controlled terms

AVDP makes use of reference data also known as controlled terms, represented e.g. as MPEG‑7 classification schemes, whenever possible. While ClassificationSchemeType is not included in the profile, descriptions conforming to the profile can make reference to terms defined in MPEG‑7 classification schemes. In order to simplify descriptions, AVDP also excludes the MPEG-7 SemanticDS, recommending references from TextAnnotationType to controlled terms rather than defining the concepts inside the description document.

In addition to the classification schemes provided with the MPEG-7 standard, the EBU has defined a large set of openly available classification schemes, which can be found at http://www.ebu.ch/metadata/cs. In particular, the profile enforces the use of controlled terms for SegmentType/StructuralUnit and SegmentDecompositionType/@criteria in order to ensure well-defined semantics of the elements in a hierarchical content decomposition.

# Benefits

In summary, the use of AVDP provides the following benefits:

* Subset of MPEG-7 tailored to the needs of applications in media production and archiving
* Reduced complexity and improved interoperability compared to using the complete MPEG-7 standard by
  + clear separation of metadata extracted from different modalities, abstraction level and structural elements and their representation elements (e.g. key frames)
  + constraints on description tools and clear definition of their semantics (allowing formalisation and automated validation of semantic constraints)
  + mandatory use of identifiers for structural elements (decompositions, segments) and provision of classification schemes for these identifiers
* Full support for video and audio descriptors

# Related resources

## Validation service

Validation of MPEG-7 documents is an important issue whenever documents are produced, exported or imported. On a syntactic level, standard tools are available for this problem, most notably XML schema validators. However, the semantic expressivity of XML schema is limited, and thus validations on higher levels cannot be done with standard tools, but need specific application logic.



VAMP (<http://vamp.joanneum.at/>) is a service that formalises the semantic constraints in AVDP in order to automatically validate documents w.r.t. the profile definition, beyond schema validation.

VAMP is a web based application with a graphical user interface deployed at http://vamp.joanneum.at. A command line client application can be downloaded to (batch) process local files.

## NHK Metadata Production Framework (MPF)

MPF is a specification proposed by NHK from 2006 that provides a common environment for the effective generation of content-based metadata for video. The MPF provides the mechanism of combining various multimedia analyses as modules for generating the desired metadata. The MPF adopts AVDP as its metadata model from version 3 on, and specifies two interface types for module control and metadata operation. Metadata Editor is part of the reference software by which user can test the basic functionality of MPF. Therefore if you follow the specified interface and develop the module with your own information extraction algorithm, you can test it on the Metadata Editor easily. The Metadata Editor and related materials are downloadable from the following site (<http://www.nhk.or.jp/strl/mpf/>). The Metadata Editor user interface is shown on the right. The metadata generated by the editor can be export/import as a MPEG-7 data which is compliant to AVDP specification



## Further examples and tools

A collection of examples and tools is available at EBU Metadata Developer Network Knowledge Base

* <http://workspace.ebu.ch/display/ecmmdn/Knowledge+base+-+contributions>
* to access the knowledge base, please register at http://tech.ebu.ch/groups/mdn

# References

[1] ISO/IEC 15938-9:2005/Amd 1:2012, “Extensions to profiles and levels”, 2012.

[2] ISO/IEC TR 15938-11:2005/Amd 1:2012, “Audiovisual description profile (AVDP) schema”, 2012.

[3] ISO/IEC 15938-10:2005, “Information technology -- Multimedia content description interface -- Part 10: Schema definition”, 2005.

[4] ISO/IEC 15938-3:2002, “Information technology -- Multimedia content description interface -- Part 3: Visual”, 2002

[5] ISO/IEC 15938-3:2002/Amd 1:2004, “Visual extensions”, 2004

[6] ISO/IEC 15938-4:2002, “Information technology -- Multimedia content description interface -- Part 4: Audio”, 2002

[7] ISO/IEC 15938-4:2002/Amd 1:2004, “Audio extensions”, 2004

[8] ISO/IEC 15938-5:2003, “Information technology -- Multimedia content description interface -- Part 5: Multimedia description schemes”, 2003

[9] ISO/IEC 15938-5:2003/Amd 1:2004, “Multimedia description schemes extensions”, 2004