THE MPEG-7 AUDIOVISUAL DESCRIPTION PROFILE STANDARD FOR DESCRIBING RESULTS OF AUTOMATIC ANNOTATION SERVICES

Mike MATTON\textsuperscript{a}, Werner BAILER\textsuperscript{b}, Masanori SANO\textsuperscript{c}, Alberto MESSINA\textsuperscript{d}, Jean-Pierre ÉVAIN\textsuperscript{e} and Peter SCHALLAUER\textsuperscript{b}

\textsuperscript{a}Vlaams Radio & Televisieomroeporganisatie; \textsuperscript{b}JOANNEUM RESEARCH; \textsuperscript{c}NHK STRL; \textsuperscript{d}RAI - Radiotelevisione Italiana; \textsuperscript{e}European Broadcasting Union

Broadcast archives are valuable, and adding metadata is a crucial part of both the archival process and the re-activation of archive content. Metadata standards are crucial to enable interoperable content exchange. As manual annotation is very time consuming and expensive, many archives lack the resources to perform this task and rely on automatic extraction of time-coded metadata. These technologies have in common that they produce metadata on a fine level of granularity, e.g., per segment, frame or even regions in a frame. The EBU MIM-SCALE working group has developed the MPEG-7 Audiovisual Description Profile (AVDP), which reduces the complexity of the overall MPEG-7 MDS schema by selecting only the subset needed for broadcasting. This paper describes this profile and some selected applications that are able to work with AVDP.

Keywords: Metadata standard | MPEG-7 AVDP | automatic information extraction

INTRODUCTION

BACKGROUND

Broadcast archives are valuable, and all stakeholders agree that adding metadata is a crucial part of the archival process and a fundamental requirement for the exploitation of archived assets. Of course, together with the frequent exchange of content, archives and broadcasters also like to exchange the associated metadata, seamlessly. Thus, in order to achieve interoperability also at this level, metadata standards are essential.

Metadata can originate from a variety of sources: production systems, manual annotations and automatic annotations. As manual annotation is very time consuming and expensive, many archives lack the resources to perform this systematically, and use tools for the automatic extraction of time-coded metadata. There are plenty of tools that can derive information from an audiovisual source, some of which are becoming mature and ready to be deployed on a large scale. Most of these tools generate very detailed metadata providing time reference for every annotation.

However, due to interoperability and exchangeability requirements, there is a need for a common metadata standard in which these annotations can be represented. The European Broadcasting Union (EBU) is very active in devising these standards. One of these standards is the EBU Core metadata set, an extension of Dublin Core designed specifically for media (including but not limited to broadcast archives). It provides a set of descriptors for technical, administrative and descriptive metadata. However, this standard...
is insufficient for representing results of automatic analysis tools. MPEG-7 on the other hand – a metadata standard created by MPEG† – has this capability but, at the same time, raises some interoperability issues if used in its entirety. Therefore, the MPEG-7 Audiovisual Description Profile (AVDP) profile was defined to create an interoperable standard for describing results of automatic content analysis tools.

**TASK ORIENTED ANNOTATION AND SEARCH WORKFLOWS**

In the future, content annotations and search workflows will increasingly become automated. Already, there are workflow engines on the market that allow automation of business processes up to a certain level.

One problem is the integration of different components from different vendors in a single workflow. This is a challenge as every component might use its own internal data model. However, to be integrated in a single workflow, a representation needs to be agreed upon to exchange data and metadata. A second challenge is that resources are increasingly distributed. Media researchers want to be able to retrieve content, irrespective of the physical location of the content.

The European project TOSCA-MP tackles these challenges by building such an integrated workflow for media annotation and search workflows. Figure 1 shows an overview of the logical system components of this TOSCA-MP system. Four main components can be identified:

- The Distributed Repository Framework (DRF) is responsible for storing and managing data and metadata. It consists of many distributed repositories including external repositories to which the framework can connect.
- The Services provide many atomic services in a media annotation and search workflow. Examples are automatic annotation services such as speech recognition, indexing services, semantic enrichment, etc.
- The Metadata Production and Management Framework binds the system together. It allows to define workflows which connect individual services together.
- Finally, the User Interfaces are the components that are visible to the end user (e.g. documentalist, media researcher, etc.). Examples are the search user interface, a user interface for monitoring the workflow processes and so on.

![Figure 1 - Logical system design of the TOSCA-MP system](image-url)

† ISO/IEC JTC 1 SC29 WG11
However, for binding these components together, standards are needed to allow the
different services in a workflow to communicate with each other. As MPEG-7 AVDP has
been specifically created for representing results of automatic annotation services, this
standard suits this purpose very well and it has been chosen as the main standard for
exchanging metadata in the TOSCA-MP project.

Presenting the entire system in this paper would lead us too far. Instead, we refer the
interested reader to different available publications such as Bailer, Altendorf, Messina,
Negro and Thallinger (2012). Even more information is available on the TOSCA-MP web
site‡, including the public deliverables of the project.

The remainder of this paper is organised as follows: First, different metadata standards
are described, then MPEG-7 AVDP is described in detail. Then some applications
working with AVDP are described. Finally, we conclude and summarize the contributions.

EXISTING METADATA STANDARDS

In the following, some important metadata standards for multimedia are described. We
then discuss the MPEG-7 AVDP profile, the main topic of this paper.

DUBLIN CORE

Dublin Core§ is a metadata element set that is intended to be a common set of elements
that can be used across many different resource types. It has been approved as a U.S.
15836).

The standard contains 15 basic descriptive elements, which can be tagged with a
qualifier and which can occur many different times. The elements are: contributor,
coverage, creator, date, description, format, identifier, language, publisher, relation,
rights, source, subject, title and type. Note that this list contains descriptive, technical a
nd administrative metadata fields. While Dublin Core is widely agreed upon as a common
standard, the expressiveness of it for broadcast media is usually too limited. Moreover,
Dublin Core cannot cope with segmented and regional descriptive metadata. The
specification of Dublin Core can be found in the Dublin Core Metadata Element Set,
version 1.1**.

EBU CORE

The existence of the EBU Core†† metadata standard dates back to 2000. The standard
was originally devised as a refinement of Dublin Core for audio archives, but since then it
has been extended several times. The current scope of EBU Core is identified as the
minimum information needed to describe radio and television in broadcasting. It
addresses creation, management and preservation of audiovisual content. Next to a
formal XML Schema, it is also available as an RDF (Resource Description Framework,
Klyne and Carroll (2004) ‡‡) ontology, and it is consistent with the W3C Media Annotation
Working Group ontology‡‡. For more information we refer to the EBU Technical
document Tech 3293§§.

‡ http://tosca-mp.eu
§ http://www.dublincore.org
** http://dublincore.org/documents/2010/10/11/dces/
†† http://tech.ebu.ch/lang/en/MetadataEbuCore
‡‡ http://www.w3.org/TR/mediaont-10/
EBU Core is registered in Class 13 of the SMPTE RP210 dictionary to allow mapping to MXF (among other applications) and will be used as the basis of the announced SMPTE Core.

MPEG-7
MPEG-7, ISO/IEC 15938 (2001), is devised as a standard in which multimedia content can be annotated. It has been designed for a broad range of applications. It allows for a very fine grained annotation of the multimedia content and it is able to describe time-segmented metadata, which usually is the result of automatic analysis tools.

The standard definition consists of different levels. The basic element of the standard is a language for specifying the elements of the standard, called Description Definition Language (DDL). The next one is a set of description schemes (DS) and descriptors (D) that can be used to describe technical, administrative of descriptive metadata. There are specific descriptors for audio and video features. Another element in the standard is a scheme for coding the description in order to provide a standard to store it, or to multiplex it with the content.

The MPEG-7 standard is a very open standard, leaving a lot of choice to the application designer on how to structure the metadata. This can be a benefit, but it is also clearly a disadvantage for interoperability, to overcome which, an additional layer of definitions is necessary. MPEG-7 allows this extra layer to be provided by means of MPEG-7 profiles. Therefore, the MPEG-7 Audiovisual Description Profile (AVDP) has been defined, specifically for describing the results of automatic multimedia analysis systems. This profile is described in some more detail in the next paragraph.

THE MPEG-7 AUDIOVISUAL DESCRIPTION PROFILE (AVDP)

In order to address the issue related to the representation and exchange of automatic content analysis tools, The EBU MIM-SCAIE working group***, which brings together experts in the field to identify and assess for methods to include automatic content analysis techniques in media production and archival workflows, has started the definition of a new MPEG-7 profile, the Audiovisual Description Profile. Like in the MPEG coding standards, a profile is a subset of the complete standard covered the requirements of a specific application area. AVDP is tailored to the requirements of representing metadata created using semi-automatic analysis tools in media production and archiving processes. In 2012, AVDP has become an ISO standard [3].

PURPOSE AND FUNCTIONALITIES

The profile includes functionalities for representing results of several – (semi)-automatic or manual – information extraction processes. Based on the requirements from the broadcast domain, functionalities for describing the results of temporal segmentation, text recognition, person identification, format detection, genre detection, keyword extraction, speech recognition, subject classification and copy/near duplicate detection have been taken into account. As there is ample literature on these automatic information extraction techniques, we will not describe them in detail here, but focus on the representation of results. Many of these tools generate time-based metadata, and AVDP supports representing the results as a set of parallel time lines. The different results can be described in different temporal granularity. Some tools also produce results that are linked to a specific region in the video, such as face detection. AVDP also supports the representation of static and dynamic regions. For all results, the confidence level of the automatic detection as well as for describing information extraction tools, version, contributors, and the date/time the tool was applied can be described.

*** http://tech.ebu.ch/groups/pscale
STRUCTURE AND SEMANTICS

In contrast to other MPEG-7 profiles, AVDP defines the structure of the metadata document and the semantics of the different parts in this structure very clearly. It goes beyond defining just a subset of elements to be included, but limits the degrees of freedom of many elements and the variability in representation in order to avoid ambiguities and ensure interoperability also on a semantic level. A typical pattern in an MPEG-7 document is a hierarchical structure of segments and decompositions of these segments. Here AVDP enforces the use of type identifiers from a controlled vocabulary to describe the semantics of these constructs.

Modularity has been an important design goal of AVDP. It helps to facilitate accessing specific information in the document, to reduce the integration effort when describing results of different tools in one document and to avoid interdependencies between different parts of the document. Modularity of AVDP is provided by the following design rules:

- Separating metadata generated with different tools,
- Separating metadata with different abstraction level (e.g., directly derived from audiovisual data vs. generated with help of external knowledge),
- Separating metadata derived from different modalities, and
- Separating segmentation of content and representative elements such as 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. The second is used to describe several contents and their relations (e.g., similar segments or copies). The third describes the result of summarization and the related audiovisual contents.

AVDP provides a set of description tools to describe the results of various kinds of media analysis with visual and audio low-level features. In contrast to other MPEG-7 profiles, AVDP is based on version 2 of MPEG-7, and includes the low-level visual and audio descriptors defined in parts 3 and 4 of the standard. The constraints on description tools in AVDP concern those defined in part 5 (Multimedia Description Schemes, MDS) of the standard, while the descriptors from parts 3 and 4 are hardly constrained. The main constraints in the schema are the restriction of AVDP documents to support only complete content descriptions and summaries. As the focus is clearly on detailed content description, most tools for describing collections, user profiles, etc. have been excluded. The Semantic DS has been excluded, as there are much simpler tools at hand for referencing terms in ontologies, which can be represented in OWL, leveraging a rich environment of existing tools.

A number of constraints aim at improving interoperability, such as the limitation to relative and incremental time specification (including rules for placing time base and units), requiring IDs on all description scheme type elements, restricting string qualifications to URIs whenever possible, and requiring particular subsets of description metadata at specific elements. The profile definition also recommends the use of specific classification schemes, in particular for structural components such as decomposition criteria and segment types.

Figure 2 visualizes the high level structure for content description supported by this profile. The root audiovisual segment (AudioVisual) expresses a whole content to which automatic media analysis is applied, and it is decomposed by temporal decompositions (TD) into AudioVisualSegments (AVS). Each temporal decomposition corresponds to the result of an analysis/annotation process. If specific audio and/or video descriptors are needed, the AVS are 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 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.
CONTROLLED VOCABULARIES
Whenever possible, AVDP makes use of controlled vocabularies for identifying types of elements or describing content analysis results. MPEG-7 offers a representation for controlled vocabularies called classification schemes, but AVDP can support any type of vocabulary, as long as URIs can identify its elements. This includes ontologies, for which a standardized representation (OWL) and a wide range of tools exist. Thus, the MPEG-7 Semantic DS, which supports similar constructs, has not been included in AVDP.

In addition to already existing classification schemes, there is a set of defined by the EBU and published at the EBU website. This includes some classification schemes specifically defined to identify constructs in AVDP.

APPLICATIONS
We will present and demonstrate four selected applications for handling MPEG-7 AVDP metadata. The first one is a semantic validation service for MPEG-7 AVDP documents, and the others include automatic tools for metadata generation and user interfaces for verification and annotation.

VAMP VALIDATION SERVICE
Validation of metadata documents is an important issue whenever documents are produced, exported or imported. Standard tools, most notably XML Schema validators, can perform validation on a syntactic level. However, in order to validate metadata documents on a semantic level, additional tools are needed to capture the formal
semantics of metadata standards in scope. This applies to all aspects of the profile specification that cannot be represented using the constructs of XML Schema. Therefore we have implemented an approach for validating metadata documents on a semantic level based on a description of formal semantics of the metadata format used. This service is called VAMP, described in Troncy, Bailer, Höffernig and Hausenblas (2010, January), and serves as the reference validator for MPEG-7 AVDP.

First, the MPEG-7 input document is checked for syntactic validity against the MPEG-7 XML schema and the selected profile XML Schema. A syntactically valid MPEG-7 input document is a necessary precondition to start the semantic validation. Second, the MPEG-7 description is converted into RDF with respect to an ontology capturing the semantics of the selected profile. In this step a XSL transformation and additional classification rules are applied. Finally, these RDF triples are the input data for the semantic consistency check. In this step, validation rules based on the profile and temporal ontology are used, as described in Höffernig, Hausenblas and Bailer (2007). After the semantic consistency check, possible validation violations can be retrieved using a SPARQL query.

VAMP can be accessed through a web based user interface or a RESTful web service. The web application, shown in Figure 3, is available at through a website‡‡‡. When validating a document, all detected semantic errors are listed in the result window. For each semantic error, all XML elements, which caused this error, are listed. These XML elements are identified by XPath expressions, which enable the direct observation of the error locations in the input MPEG-7 document. In addition, a Java client can be downloaded, which allows batch processing local files by sending validation requests for the documents to the web service.

QUALITY ASSESSMENT AND VERIFICATION

The second application VidiCert allows for efficient video essence quality assessment§§§. Ensuring quality is an essential part of moving image and video production, post-production, delivery and archive operations as well as migration. Quality assessment of audiovisual content is an extremely time- and therefore cost intensive part of the work. The aim of VidiCert is to significantly automate the video quality assessment process in a two-step approach.

In the first step, video or movie content is fully automatically analyzed by VidiCert Analyser in regard to certain visual and audio impairments, e.g. image noise, strong video distortions, black frames, digital tape dropouts, blurriness or silent sections. In the second step the gathered information is presented to the user in the VidiCert Summary application for efficient verification. The result is a human verified quality report. Both applications are using MPEG-7 AVDP for storing all metadata related to the video and its quality including those visualized in Figure 4. Application areas for essence quality assessment of video and movies are:

- Quality assurance of video and movie production
- Incoming/outgoing quality assessment in video and movie post-production
- Estimation of the restoration effort for specific content
- Quality assurance before broadcast
- Efficient quality assurance for archive digitization and migration to file based environments
- Selection of and search for content with specific quality properties
- Real-time quality assessment of security and safety video
- Quality assurance for online video streaming services

‡‡‡ http://vamp.joanneum.at
§§§ http://www.vidicert.com
Efficient visualization and verification of impairment analysis results supports an operator to get a quick overview of the condition of the material and to allow for manual corrections and final quality judgment by the operator.

In the following we describe the user interface shown in Figure 4, which is composed of these four main parts: Global timeline views (1) show the occurrence of defect events for the full temporal range of the video. A global timeline view also shows the shot structure and the temporal zoom period for the timeline views in (3). For efficient verification, a defect list component (2) shows defect events and their properties. Timeline views showing a zoomed temporal resolution providing a level of temporal detail that can be freely adjusted are shown in (3) of the user interface. A video player (4) with frame accurate positioning support and audio playback is also provided.

The video player is the central component of the user interface. All other components synchronize with it. The video player can be positioned on an extra monitor for full resolution playback. The other components like the event list component and the timeline views can also be displayed on a second monitor.
Figure 4 – User Interface for efficient interactive verification of automatic detections.

All components provide additional navigation functionalities. The key frame and stripe image timeline views shown in the bottom of (3) provide a quick visual overview of the video content. Key frames and stripe images are aligned on the timeline according to their respective time points. Navigation is possible by clicking on the timeline, or by moving the scroll wheel for frame accurate positioning.

Timeline views showing impairment detection results either visualize continuous or segment-based quality measures. Continuous quality measures are visualized in form of line or bar charts like the visual activity and the noise/grain level within specific time ranges. Detections having an event-like character are also visualized on timeline views by indicating the temporal segment of the detection. These are for example video breakups, uniform color, digital tape dropout and test pattern segments. The different views appear both over the full video range in (1) and for the selected zoom period in (3). For uniform color detections the respective segments are additionally filled with the color detected.

The time an operator can devote to verify automatic analysis results is typically limited and it may be the case that not all defect detections can be manually verified. So the time the operator has available should be optimally used. For this it is very useful to be able to handle the most relevant detections first. To support this, the detections listed in the defect list view can be sorted by all columns. When sorting by severity an operator can efficiently verify the most relevant detections first. A detection can either be approved, discarded, or postponed for later verification by the operator. After such a manual verification the next detection in the list not yet verified will be selected. This verification process is supported by a special mode where the video will play in a loop around the currently selected detection including a configurable pre roll and post roll time.

The user interface allows also for manually adding defect events on the timeline, where defects are defined by a classification scheme. Additionally an operator can rate the overall quality of the video by selecting predefined rating classes.

All the information visualized in Figure 4 and all the decisions made by an operator during his work are stored in an MPEG-7 AVDP compliant XML document.
WEB-BASED VISUALISATION OF ANALYSIS RESULTS
A third selected application consists of a web application for visualizing different kinds of annotations from an AVDP document. The web application is entirely implemented using HTML5 and JavaScript, i.e. it can run on a state of the art web browser without the need to install any plugins. The application can be customized to include views for different segment-based annotation (e.g., concept detection results), graph views (e.g., some of the quality annotations described above) as well as textual information (e.g., speech to text and translation results). In addition, key frames and stripe images are visualized. All views are synchronized with the video player, and both the player controls and all views can be used for navigation in the video. The temporal level of detail of the views can be adapted. In order to support dual display environments, the video player can be detached into a separate window that can be moved to a second screen.

![Figure 5 – web based visualisation of automatic content analysis results](image)

METADATA PRODUCTION FRAMEWORK
The fourth selected application is a metadata editor provided as a reference software of MPF (Metadata Production Framework) by NHK STRL. The MPF is being developed and published from 2006. Its aim is to provide a common platform where a user can combine various multimedia analysis tools easily and flexibly in order to obtain the corresponding metadata directly or to support the metadata production activity. The basic policy is modularizing all processes related to metadata generation and its handling in the same way. To realize the environment, MPF specifies the common metadata representation and two kinds of interfaces, i.e. a module-control interface and a metadata-operation interface. The MPF specification has also been updated several times and the AVDP is adopted as its metadata model from version 3.0 onwards. More details of MPF can be found on the MPF web site****.

Figure 6 shows a screenshot of the metadata editor. As mentioned before, the metadata editor is a reference software of MPF, through which we can test a sequence of metadata generation with several automatic analysis tools. (1) Load the target video; (2) select the process modules with necessary parameters and execute them; (3) check the results and manually correct them if necessary; (4) output the data, i.e. save it to a file or send it to the database. After loading the target video, to run a process module, a track should be created, as depicted in the lower part of Figure 6, and then the module is assigned to the

**** http://www.nhk.or.jp/strl/mpf/
track and executed. The module control API and the part of AVDP which shall be handled by the modules is well defined such that the results of other tracks can be reused in another module. I.e., a cascade process can be achieved in this way. Generated metadata is expressed as a rectangle on the track. When the rectangle is clicked, its corresponding description is displayed in the upper-left part of the metadata editor in the MPEG-7 format. The metadata can be modified in the metadata editor and can be saved to an external file and/or in the database. The metadata editor can handle only one video content at a time with manual operation. Therefore, programs for an automatic batch process are also provided along with a real use case scenario.

MPF can be applied to many situations and NHK has also tested it in several actual applications. The latest trial is to use the MPF-based system to disclose a huge amount of footage taken in the Great East Japan Earthquake for shot-based pre-classification. This is still an ongoing project.

![Figure 6 - Metadata Editor based on “MPF” (NHK)](image)

**CONCLUSION**

Nowadays the ability to exchange information (metadata) along with content is a primary requirement to support efficient production workflows, and to exploit the value of archives. To achieve these goals, metadata standards are indispensable tools.

In this paper we have presented AVDP, a metadata standard able to cope with annotations generated by automatic content analysis techniques. We have motivated the necessity of this standard and have explained the rationale behind and the reason why it has been created. The standard has been described on a technical level. Finally, we presented four different applications among which a semantic validation service for MPEG-7 AVDP documents, automatic tools for metadata generation and user interfaces for verification and annotation.
The AVDP standard is becoming very valuable now that automatic metadata extraction techniques are becoming mainstream in broadcast production and archiving.

ACKNOWLEDGEMENT
This work has been funded partially under the 7th Framework Programme of the European Union within the projects “TOSCA-MP” (ICT FP7 287532), and “DAVID” (ICT FP7 600827).

REFERENCES