Journal of Universal Knowledge Management, vol. 1, no. 1 (2006), 26-35
submitted: 14/5/06, accepted: 29/5/06, appeared: 28/8/06 © J.UKM

A Description Infrastructure for Audiovisual Media Processing Systems Based on MPEG-7

Peter Schallauer, Werner Bailer, Georg Thallinger
(JOANNEUM RESEARCH
Institute of Information Systems & Information Management
Graz, Austria
{firstname.lastname}@joanneum.at)

Abstract: We present a case study of establishing a description infrastructure for media processing systems. The description infrastructure consists of an internal metadata model and access tools for using it. Based on an analysis of requirements, we selected, out of a set of candidates, MPEG-7 as the basis of our metadata model. The openness and generality of MPEG-7 allow using it in a broad range of applications, but increase complexity and hinder interoperability. Profiling has been proposed as a solution, with the focus on selecting and constraining description tools. We have defined the Detailed Audiovisual Profile as the profile to be used in our metadata model which includes semantic constraints in order to ensure interoperability between MPEG-7 based systems. For practical work with the metadata model, we have implemented a MPEG-7 library and a client/server document access infrastructure.

Keywords: metadata, MPEG-7, content analysis, content description, profile, semantics
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

1 Introduction

A growing amount of audiovisual data are produced, processed and stored digitally. Many applications, for example those dealing with multimedia archive and media monitoring, are required to handle large amounts of digital audiovisual data. The main challenge is to index this data in order to make them searchable and thus (re-)usable. This requires the audiovisual content to be annotated, which can either be done manually, in an extremely work- and thus cost-intensive process, or by applying content-analysis algorithms that automatically extract descriptions of the audiovisual data. In both cases, the aim is to create metadata, which contains a concise and compact description of the features of the audiovisual content. Metadata descriptions may vary considerably in terms of profundity, comprehensiveness, granularity, abstraction level, etc. depending on the application area, the tools used and the effort made for creating the description.

In this paper, we present a case study of establishing a metadata description infrastructure for a audiovisual media processing system including content-analysis, documentation, search and retrieval functionalities. The description infrastructure consists of the internal metadata model and the tools for accessing, modifying and storing metadata descriptions. The system, which handles video, audio and still images, consists of the following components:

- An ingesting component, which is used to import audiovisual data into the system and to perform and control automatic content-analysis tools, which extract a number of low- and mid-level metadata.
- A manual documentation component, which is used for textual descriptions and description of high-level semantic information, which cannot be extracted automatically.
- A search component for query formulation and result presentation, which provides search options for both textual and content-based queries.
- A backend infrastructure providing storage and search functionalities.

The system is designed to enable an optimized annotation workflow and parallel operation of all the components.

The internal data model of comparable systems is usually proprietary. Until a few years ago, there were no standards available, that could be used as a basis for such a data model. A number of standards for metadata of audiovisual content are designed as exchange formats only and thus cannot be used for an internal data model (cf. Section 3) With the standardization of MPEG-7, the internal data model of some systems has been based on MPEG-7 or derived some concepts from it, especially in research systems (e.g. [Döller und Kosch 2003] [Gagnon et al. 2004]). We intended to follow a similar approach of basing the metadata model on an existing standard. However, as will be described in Section 4, some application specific adaptations are necessary.

The paper is organized as follows: We start by describing the requirements of such a system with respect to the description infrastructure. Based on these requirements we define an internal metadata model. We have chosen to base our metadata model on MPEG-7 and we describe the rationale for this choice. One of the consequences resulting from the generality and flexibility of MPEG-7 is the necessity to define a subset of the standard to be used, which is formalized as a profile. We briefly summarize the rationale and the concepts behind the Detailed Audiovisual Profile [Bailer and Schallauer 2006], which we have defined for our internal metadata model. We then discuss the access tools developed and present two applications in which the description infrastructure has been used.

2 Requirements on a Description Infrastructure

This section describes the requirements imposed on the metadata infrastructure by a system as described above. All of them are technology independent in terms of storage technology, implementation language and tools. We have organized them into those concerning the metadata model and those concerning architectural aspects of the access tools that will be part of the description infrastructure. These access tools will be based on the metadata model and enable applications to work with it.

2.1 Requirements on the AV description metadata model

The metadata model is the core element for the more metadata-centric types of media processing systems, such as for example media monitoring or retrieval systems. In the following, we describe the main properties of the metadata model of such a system.

Comprehensiveness. The metadata model must be capable of modelling a broad range of multimedia descriptions (e.g. descriptions of different kinds of modalities and descriptions created with different analysis and annotation tools).
Fine grained representation. The data model must allow describing 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.
Structured representation. The metadata model must be able to hierarchically structure descriptions with different scopes and descriptions assigned to fragments of different granularity.
Modularity. The metadata model should avoid interdependencies within the description, such as between content analysis results from different modalities (e.g. audio and visual). The metadata model shall also separate descriptions which are on different levels of abstraction (e.g. low-level feature descriptions and semantic descriptions). This is important, as descriptions on higher abstraction levels are usually based on multiple modalities and often use domain specific prior knowledge.
Extensibility. It must be possible to easily extend the metadata model to support types of descriptions not foreseen at design time or which are domain or application specific.
Interoperability. It shall be easily possible to import metadata descriptions from other systems or to export to other systems.

2.2 Requirements on the access tools

Access components are software tools that enable the components of the systems to use the metadata model. They shall abstract the technical details of the metadata description (e.g. storage format, database scheme) and provide an object-oriented access to the metadata model. The key properties of access components are *fine grained access*, *independence of the storage technology* and support of a *distributed architecture.* A detailed discussion of the requirements can be found in [Bailer et al. 2005].

3 Definition of a Metadata Model

When defining the internal metadata model, we wanted to build as much as possible on existing standards for the description of audiovisual data. One reason is to facilitate interoperability with other systems and the other is that it simply does not make sense to reinvent the wheel.

There is a number of standards that are candidates for being used as the basis of our metadata model. Those candidates are the Dublin Core Metadata Initiative (DCMI) [ISO 2003a], the EBU P/Meta [EBU 2003] standard, BBC's SMEF data model [BBC 2000], the SMPTE MXF Descriptive Metadata Scheme (DMS-1) [SMPTE 2004] and MPEG-7 [ISO 2001a]. After reviewing their strengths and weaknesses, we have selected MPEG-7. The rationale for this decision is described in the following.

MPEG-7, formally named Multimedia Content Description Interface, is a standard for describing multimedia content, independent of the encoding of the content, and allows different levels of granularity of the description. MPEG-7

descriptions can be represented either as XML (textual format, TeM) or in a binary format (BiM). A good overview can be found in [Martinez 2002].

MPEG-7 has been designed as a metadata model, while some of the other candidates are mainly metadata dictionaries or have been designed as metadata exchange formats (as discussed for P/Meta in [Carter 2003]). Other standards thus lack of comprehensiveness, as typically only certain subsets are needed for exchange. Many of the standards designed as exchange formats also lack sufficient structuring capabilities, as they are rather modelled as flat lists of attributes than as hierarchical structures.

MPEG-7 has been designed for a broad range of applications. Thus most of the concepts are very general and widely applicable. Some of the other standards, for example those from the broadcast domain, are tailored towards this application area and thus lack of some generality.

MPEG-7 supports fine grained description of fragments of the content, and is the most flexible standard for describing different levels of abstraction. It allows defining arbitrary fragments of the content and does not limit structuring of these fragments.

The data model of MPEG-7 is very flexible in terms of structuring capabilities. This is especially true for the spatial, temporal and spatiotemporal structuring tools defined in part 5 [ISO 2001c] of the standard. This flexibility allows modularizing the description, which is a prerequisite for fine grained access to parts of the document. The structuring tools also allow hierarchical organization of description fragments with different scope.

The fact that MPEG-7 has been defined using XML Schema simplifies mapping MPEG-7 descriptions to object structures in order to build APIs. The use of XML Schema for the definition of the data model also facilitates extensibility. This is an important advantage, as no standard can fulfil all requirements of the internal metadata model of a system and some application specific extensions will be required. According to the MPEG-7 conformance guidelines [ISO 2001d], a data model based on MPEG-7 with some extensions still represents a MPEG-7 compliant content description.

The XML Schema based definition of the standard also supports a document oriented approach, which allows modelling a relation between one multimedia document to one associated metadata document. The fact that MPEG-7 descriptions can also be serialized as XML documents also increases practical usability because of the number of available XML processing tools and the fact that it is human readable.

4 Using MPEG-7 as Metadata Model for Audiovisual Content Description

MPEG-7 provides a high amount of 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. Some of the reasons described above for choosing MPEG-7 are directly related to this flexibility and openness. However, when practically using MPEG-7, two main problems arise from these features: complexity and hampered interoperability. The MPEG-7 requirements group has early recognized these issues [ISO 2001e].

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 may cause hesitance in using the standard in real-world applications.

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 description schemes and descriptors used 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 has the consequence that an additional layer of definitions is necessary to enable full interoperability between systems using MPEG-7.

4.1 MPEG-7 Profiles

4.1.1 Concept of Profiles

Profiling has been proposed to partially solve these problems [ISO 2003c]. 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. The definition of profiles will also facilitate interoperability between different applications working with MPEG-7 descriptions.

There are three steps to define a profile [ISO 2003c]:

- Selection of tools supported in the profile, i.e. the subset of descriptors and description schemes that are used in descriptions conforming to the profile.
- Definition of constraints on these tools, such as restrictions on the cardinality of elements and on the use of attributes.
- Definition of constraints on the semantics of the tools, which describe their use in the profile more precisely.

The results of tool selection and the definition of tool constraints are formalized using the MPEG-7 DDL [ISO 2001b] and result in an XML schema like the full standard. The semantic constraints only play are marginal role in the document mentioned above, although they are in our view the most interesting ones to tackle the complexity and interoperability problems by the use of profiling.

4.1.2 Adopted Profiles

Several profiles have been under consideration for standardization and three profiles have been adopted as part 9 of the standard [ISO 2003b], the XML schemas of the profile are defined in part 11 [ISO 2005]. 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 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 of the standard.

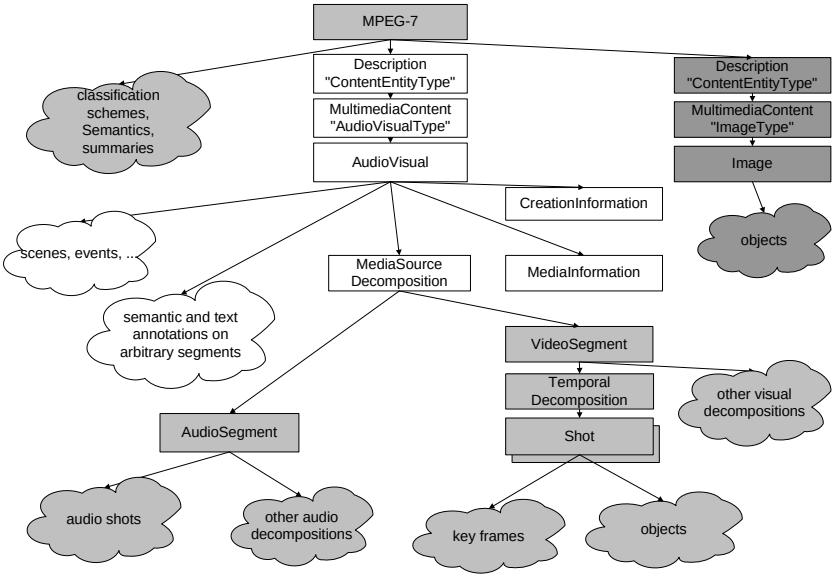

Figure 1: Overview of the structure of descriptions conforming to DAVP.

It has to be noted, that all of the adopted profiles just define the subset of description tools to be included and some tool constraints. None of the profile definitions includes constraints on the semantics of the tools that clarify how they are used in the profile.

4.2 Detailed Audiovisual Profile

In [Bailer and Schallauer 2006] a profile for the detailed description of single audiovisual content entities called Detailed Audiovisual Profile (DAVP) has been proposed. The profile includes many of the MDS tools, such as a wide range of structuring tools, as well as tools for the description of media, creation and production information and textual and semantic annotation, and for summarization. The structure of a MPEG-7 description using DAVP is shown in [Fig. 1]. In contrast to the adopted profiles, DAVP includes the tools for audio and visual feature description, which was one motivation for the definition of the profile. The other motivation was to define a profile that supports interoperability between systems using MPEG 7 by avoiding possible ambiguities and clarifying the use of the description tools in the profile. Defining a subset of functionalities, as it has been done in the definitions of the adopted profiles, reduces complexity, but it does not solve the interoperability

problem. DAVP thus includes a set of semantic constraints, which play a crucial role in the profile definition. They define the use of the MPEG-7 description tools in the context of the profile and allow describing the relations between the description tools included in the profile and constraining their use depending on the context. Only the semantic constraints can facilitate interoperability across systems by ensuring exchangeable MPEG-7 descriptions. Due to the lack of formal semantics, these constraints are only described in textual form in the profile definition [DAVP 2005].

5 Access Tools and Applications

In this section we describe the software tools and components we designed and implemented for establishing the MPEG-7 based description infrastructure and two applications in which the profile and the tools are used.

For using the MPEG-7 based metadata model in the components of a system we have implemented a C++ API for parts 3, 4 and 5 (visual, audio, MDS) of MPEG-7 [MPEG-7 Library 2006]. This library enables application developers to create multimedia content descriptions, manipulate them, serialize them to XML and de-serialize them – with validation – from XML. One major design goal was to simplify extending single classes to allow the developer to enrich interface functionality for certain descriptors. The library is freely available.

We have also designed a client/server infrastructure for access to metadata documents, with the focus of functionality on the server side. The document server provides read/write access to MPEG-7 documents for a number of clients and allows the exchange of whole documents or fragments thereof. Access to parts of documents is crucial for the efficiency of the system, as MPEG-7 XML documents of larger media items tend to have considerable size. A SOAP web service interface is provided to the clients.

5.1 Multimedia Mining Toolbox

The Multimedia Mining Toolbox provides users and application developers with tools for powerful combined text and content based search on multimedia data (video, audio, still images). The digital content is automatically analyzed and annotated. Manual annotation and usage of legacy metadata is included for text based search. An overview of the system is shown in [Fig. 2].

The *media-analyze* component is responsible for media import and automatic metadata extraction. During import fully automatic content analysis is performed (shot boundary detection, camera motion estimation, key frame extraction, moving object segmentation, extraction of low-level visual features).

The *media-find* component provides very fast access to the digital archive by supporting the formulation of combined text and content-based (similarity based) queries for visual content. The tool enables to efficiently search for all features automatically extracted by the *media-analyze* tool.

The innovative *media-summary* viewer visualizes an entire video on one screen in terms of a temporal summary/overview and by providing efficient navigation functionality by shot structure and key frames. All content can be played back at

various speeds and trimmed for further editing, the actual position is always synchronized with the temporal summary.

5.2 PrestoSpace

The objective of PrestoSpace (http://www.prestospace.org) is to develop an integrated approach to the preservation of and access to audiovisual archives, the so-called "PrestoSpace Factory". It covers the complete workflow from digitisation, preservation, documentation, restoration and storage to access and delivery. The description infrastructure discussed above is used in the documentation process to collect information produced by automatic content analysis tools, annotated manually and imported from legacy metadata. The metadata related to audiovisual content description are represented using MPEG-7 DAVP.

Automated defect and quality analysis is performed on the material to be preserved in order to assess its condition and plan restoration steps. For this purpose, extensions of MPEG-7 for the description of visual defects and quality measures have been proposed.

6 Conclusion

We have implemented a description infrastructure for a distributed audiovisual content-analysis and retrieval system. The metadata model is based on MPEG-7, which has turned out to be an appropriate choice because of its generality and flexibility, especially of the concepts defined in part 5 of the standard. The hierarchical structure of the descriptions makes them modular and allows fine-grained access to parts of descriptions.

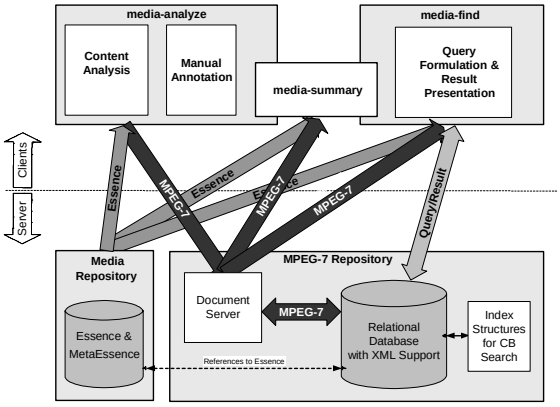

Figure 2: System overview of description infrastructure components in the Multimedia Mining Toolbox.

Our experience shows, that it is necessary to restrict the generality (and thus the complexity) of MPEG-7 by defining a subset of the MPEG-7 tools to be used in a metadata model. The way to formalize this subset is the definition of a profile, and we have thus defined the Detailed Audiovisual Profile (DAVP). The novelty of this work is that DAVP is the first MPEG-7 profile, where semantic constraints play a crucial role in the profile definition. They define the use of the MPEG-7 description tools in the context of the profile and allow describing the relations between the description tools included in the profile and constraining their use depending on the context. This means that only semantic constraints can facilitate system interoperability by ensuring exchangeable MPEG-7 descriptions.

The Detailed Audiovisual Profile (DAVP) has been successfully used for audiovisual content description in a wide range of applications. The semantic constraints, which especially ensure the modularity of descriptions, have proven to be the key enabler for interoperability.

To ensure conformance to profiles on a semantic level, and thus use profiles as a means for interoperability between systems using MPEG-7, the definition of future profiles must include a definition of their semantic constraints. If the validation of conformance in terms of semantic constraints shall be done automatically, an appropriate formalization of these semantic constraints has to be found.

Acknowledgements

The work described in this paper has been supported by several colleagues within JOANNEUM RESEARCH whom the authors would like to thank here. This work has been funded partially under the 5th Framework Programme of the European Union within the IST project “MECiTV” (IST-2001-37330, http://www.meci.tv) and partially under the 6th Framework Programme of the European Union within the IST project “PrestoSpace” (IST-FP6-507366, http://www.prestospace.org).

References

[Bailer and Schallauer 2006] W. Bailer and P. Schallauer, The Detailed Audiovisual Profile: Enabling Interoperability between MPEG-7 Based Systems, Proceedings of 12th International Multi-Media Modeling Conference, Beijing, CN, Jan.2006.

[Bailer et al. 2005] W. Bailer, P. Schallauer, M. Hausenblas, G. Thallinger, MPEG-7 Based Description Infrastructure for an Audiovisual Content Analysis and Retrieval System, Proc. Conference on Storage and Retrieval Methods and Applications for Multimedia, San Jose, CA, USA, Jan. 2005.

[BBC 2000] BBC, SMEF Data Model 1.5, 2000.

[Carter 2003] A. Carter, Data-modelling terminology and P/Meta, EBU Technical Review, Nr. 294, 2003.

[DAVP 2005] MPEG-7 Detailed Audiovisual Profile (DAVP). URL: http://mpeg-7.joanneum.at

[Döller und Kosch 2003] M. Döller and H. Kosch, An MPEG-7 Multimedia Data Cartridge, Proc. SPIE Conference on Multimedia Computing and Networking 2003 (MMCN03), Santa Clara, CA, Jan. 2003.

[EBU 2003] EBU, The EBU Metadata Exchange Scheme, EBU Tech 3295, Mar. 2003.

[Gagnon et al. 2004] L. Gagnon et al., MPEG-7 Audio-Visual Indexing Test-Bed for Video Retrieval, Proc. of Internet Imaging V Conference. San Jose, CA, USA, Jan. 2004.

[ISO 2001a] Information Technology—Multimedia Content Description Interface. ISO/IEC 15938, 2001.

[ISO 2001b] Information Technology—Multimedia Content Description Interface, Part 2: Description Definition Language. ISO/IEC 15938-2, 2001.

[ISO 2001c] Information Technology—Multimedia Content Description Interface, Part 5: Multimedia Description Schemes, ISO/IEC 15938-5, 2001.

[ISO 2001d] Information Technology—Multimedia Content Description Interface, Part 7: Conformance, ISO/IEC 15938-7, 2001.

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

[ISO 2003a] Information and documentation—The Dublin Core metadata element set. ISO 15836, 2003.

[ISO 2003b] Study of MPEG-7 Profiles Part 9 Committee Draft. ISO/IEC JTC1/SC29/WG11 N6263, Dec. 2003.

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

[ISO 2005] Information Technology—Multimedia Content Description Interface, Part 11: MPEG-7 profile schemas. ISO/IEC 15938-11, 2005.

[Martinez 2002] J. Martinez (ed.), “MPEG-7 Overview”, ISO/IEC JTC1/SC29/WG11 N4674, March 2002.

[MPEG-7 Library 2006] Joanneum Research MPEG-7 Library v. 2.0. URL: http://mpeg-7.joanneum.at

[SMPTE 2004] SMPTE, Material Exchange Format (MXF) Descriptive Metadata Scheme - 1, SMPTE 380M-2004.