Automatic Metadata Editing using Edit Decisions

Werner Bailer, Harald Stiegler, Georg Thallinger

JOANNEUM RESEARCH, Institute of Information Systems and Information Management
Steyrergasse 17, 8010 Graz, Austria
{firstname.lastname}@joanneum.at

Keywords: metadata editing, MPEG-7, edit decision, EDL

Abstract

Throughout the audiovisual media production workflow, content from various sources is combined and edited. In many stages of the workflow metadata is involved, but the final product is usually just the edited essence, while the metadata is lost. If the edit decisions taken for the essence are applied to the metadata descriptions of the sources, a metadata description for the target content can be generated automatically. We analyse the problem of automatic metadata editing and present a software framework for this task. Using MPEG-7 descriptions and an EDL, we demonstrate the feasibility of our approach with an example.

1 Introduction

Audiovisual media production workflows are complex processes involving a number of stages. In many of these stages, metadata are either produced or used. For example, during shooting settings of the camera and the position of the equipment can be recorded, which can be used for inserting effects later in the workflow. There are a number of different kinds of metadata involved, ranging from very low-level information such as color look-up tables to high-level information such as text annotations describing the content of the scene. The metadata formats used in the different stages of the workflow are quite diverse, and so are the sources of the material being used. The fact that material is shot with different equipment, computer-generated or re-used from archives contributes to the heterogeneity of the available metadata.

While there are interfaces to exchange metadata between some stages of the workflow (often depending on the compatibility of the tools being used), there is no continuous metadata chain throughout the entire production workflow. The outcome of the production process is of course the edited essence. A big part of the metadata that has been available throughout the production process is not available any more at the end of the chain. If needed later, e.g. re-using material from previous productions or for archiving purposes, it must be reconstructed. Automatic content analysis of the output essence can only help to reconstruct part of this information, but especially metadata of high semantic value is lost.

1.1 Motivation

The IP-RACINE project [7] aims at improving the digital cinema production workflow, among others by establishing a continuous metadata workflow throughout the production chain. One prerequisite are tools for conversion between different metadata formats, the other is to preserve metadata across the editing step. For the latter we propose to perform *metadata editing*. The approach is not only applicable to digital cinema production, but to any digital audiovisual media production process.

Editing is the process of condensing a large amount of raw material into the movie that is the final result of the production process [1]. In the editing process clips of the source material are put together to create a new material using cuts, transitions, keying, matting, etc.

With the term *metadata editing* we refer to the creation of metadata for the new content (the result of the editing process) using the metadata that may be available for the different source materials. In literature, the term metadata editing sometimes denotes manually modifying and completing metadata information. In the context of this work, metadata editing means applying the editing steps that have been performed on the essence also automatically to the associated metadata. The edit decisions taken for the essence (represented for example as EDL or using AAF) are used as input and applied to the metadata descriptions of the source essence.

In this paper the term *metadata* denotes the description of the audiovisual content, but not the description of the production process, such as editing decisions, as well as other information such as production schedules, etc.

The motivation to perform metadata editing is to preserve metadata descriptions that have been available for the source content and that would otherwise be lost and would have to be created again for the newly produced content. For types of metadata that can be extracted fully automatically in a reliable way (typically low- and mid-level metadata), an alternative approach is to extract the metadata again from the edited essence. As the automatic approaches cannot provide sufficient reliability, the analysis results of the source content have to be validated and corrected manually. This information is lost when re-running automatic metadata extraction on the output, which is not the case when using automated metadata editing.

Of course it is most appealing to apply metadata editing to those types of metadata, which are difficult and costly to

create, i.e. high-level metadata that have to be annotated manually.

1.2 Related Work

To the knowledge of the authors, there is not much literature available on the specific problem we are trying to solve. In [8], a framework for editing metadata together with the essence by merging descriptions of the same content and to project portions of one metadata document into another is proposed. The approach uses a proprietary XML-based metadata format.
There are a number of problems, which are related to editing audiovisual metadata: Merging of metadata in general, which for example occurs in library or museum cataloguing, when metadata from different collections are used in order to complement the descriptions. Merging of heterogeneous metadata from different sources is also one of the central problems of the Semantic Web. Another somewhat related problem is the use of metadata to facilitate video editing (e.g. [3]).

The rest of this paper is organized as follows: Section 2 presents an overview of the metadata editing process and the representations for metadata and edit decisions involved. We then analyse the problem of metadata editing in more detail by considering specific properties of metadata elements and edit operations. In Section 4 we propose a framework for automatic metadata editing. Section 5 discusses the implementation of some components of the framework and presents an example for metadata editing. We conclude with a discussion of the achieved results and the issues to be further investigated.

2 The Metadata Editing Process

2.1 Workflow

In order to perform metadata editing, we need the metadata descriptions of the source materials and a representation of the edit decisions. Metadata editing thus takes place in the postproduction after editing has been performed, i.e. at least the edit decisions on low-resolution material must have been defined. Then metadata editing can be applied automatically in order to generate a metadata description of the material, which is the result of the postproduction process. The basic formulation of the problem of metadata editing is straight-forward and simple:

- Take the source metadata descriptions and the edit decisions as input,
- extract parts from the source metadata descriptions according to the source information in the edit decisions,
- combine these parts into the target metadata description using the information in the edit decision.

If the editing step is not possible due to contradictions or ambiguities in the source metadata descriptions (cf. Section 3), the user can manually validate and correct the created descriptions. Figure 1 shows an overview of the metadata editing process.
Two types of data formats are involved in the metadata editing process: metadata formats (or metadata standards) and formats for describing edit decisions.

2.2 Metadata Representation

Throughout the audiovisual production workflow, a number of heterogeneous types of metadata exist, and many of them are represented in specific formats and standards. Apart from a number of proprietary formats, the following metadata standards for audiovisual metadata are relevant in the production workflow: MPEG-7, formally named Multimedia Content Description Interface [9], the Dublin Core metadata set [4], the SMPTE Metadata Dictionary [13], MXF DMS-1 [11], EBU P/Meta [5], as well as the technical metadata in the header of file formats such as DPX or JPEG2000. In order to deal with these diverse types of metadata, we need to map all the different metadata inputs to a unified metadata representation, which can serve as the internal metadata model of the metadata editing system. As discussed in [12], the MPEG-7 Detailed Audiovisual Profile [2] is a good basis for such a metadata model, as it is sufficiently comprehensive to cover the requirements of the different formats. It is thus used as the central metadata representation in the proposed metadata editing framework.

2.3 Formats for Edit Decisions

There are two main formats for describing edit decisions, which are relevant: An *Edit Decision List (EDL)* is an ASCII file format that lists edit decisions in terms of the source to be recorded, the target to record it to and the references to physical tapes and time codes. EDL is used to interchange editing metadata like cuts, transitions and simple effects between editing stations or offline editing and online finishing. EDL generally cannot describe file-based material or complex editing operations. There is a standardized version of EDL, SMPTE 258M, but it is not widely used. There exist several EDL flavours, which are similar, but differ in many small details.
The *Advanced Authoring Format (AAF)* is often called a super-EDL, but in fact this metadata format can contain even more information than that description implies. An AAF file can contain the essence itself along with the metadata or only references to the essence. Timelines can consist of several tracks, each with transitions and operators between clips inside a track and between tracks. Several timelines with varying priorities can be included.

3 Analysis of the Automatic Metadata Editing Problem

As we have seen in Section 2.1, the formulation of the automatic metadata editing problem is simple. However, the complexity of the problem depends on the types of metadata and the edit operations involved. We analyse in the following

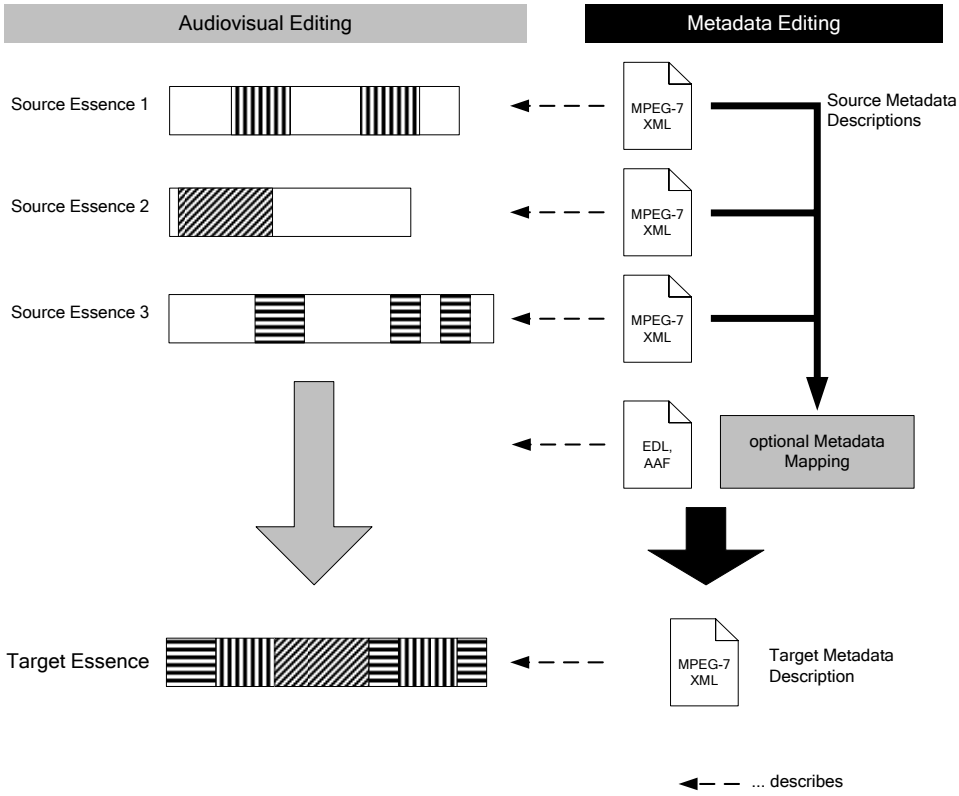

Figure 1: Overview of the metadata editing process.

the properties of different types of metadata and the relevant editing operations and discuss the implications for automatic metadata editing.

3.1 Metadata Properties

In the audiovisual production workflow, we encounter a number of quite diverse types of metadata. In the following, we review the properties of metadata elements, which are relevant for automatic metadata editing.

Scope. Metadata elements may refer to the whole content (global metadata) or to just a segment of it. If the scope of a metadata element is smaller than the complete content, edit decisions will directly influence the metadata element. The influence of editing on the global metadata can be compared to merging metadata descriptions of the same content. This is in many cases simpler, but there may be conflicting parts in the global description that need to be resolved.

Time dependency. This is the most important property with respect to metadata editing, as any metadata element that refers to a segment in time and not just a single time point must be split and/or merged in the editing process.

Splitting must be done, if the segment being described is split by an edit operation. The complexity of splitting the source segment depends on whether the value of the metadata element is constant over the whole segment or changes over time. In SMPTE EG 42, these types of metadata are called *dynamic* and *static* metadata respectively. For splitting dynamic metadata, the sampling density of the values over time determines the precision of the split that can be achieved.

Merging in the target description is only necessary if the metadata element is not naturally delimited by the edit due to its type. For example, a visual property such as the camera motion cannot continue across the edit, while a semantic property such as the topic of a segment can do so.

Spatial dependency. Metadata elements may refer to all of the content or just to a spatial/spatiotemporal segment (e.g. a static or moving object). The spatial scope of metadata elements is relevant, if editing operations combining parts from different sources into one result frame are used, such as keying or some gradual transition effects. In order to perform the metadata editing operation precisely, knowledge about the spatial properties of the editing operation (e.g. masks) is required. The problems are similar to those related to the temporal scope of the metadata elements. However, while the temporal scope of a metadata description is specified explicitly in most cases (in most metadata formats the description is attached to the segment it applies to, or the time range for which it is valid is specified), the spatial scope of the metadata may not always be specified. Ideally, metadata referring to one object in the scene would be attached to a spatial/spatiotemporal segment of the description representing this object. This is often not the case, partly because of the fact that many metadata standards/formats do not support this fine-grained kind of description. There might be a description of the objects appearing in the scene without specifying the spatial position, which prevents generating a correct metadata description for the target content.

Modality/channel dependency. A metadata element may refer to all modalities of a content (e.g. genre is sports), to a specific modality (e.g. color distribution of a shot) or even just one channel (e.g. spoken text in the center audio channel). This property of the metadata elements is relevant in cases where split operations are performed and edits only apply to some channels of the content. In the case that metadata elements describe several channels, it has to be decided if it is still relevant for the subset of channels used in target content. Whether metadata refer to a specific modality or all of them strongly depends on the type of metadata. From their nature, low-level metadata are usually specific to one modality, while high-level metadata (textual and semantic metadata) are often valid for several modalities of the audiovisual content.

High-level metadata. If high-level metadata, such as textual or semantic annotations are included in the description of the source material, the semantics of the metadata have to be taken into account when merging them into the target description. In order to process them automatically, a formalized description of the semantics of the elements of the metadata description is required (cf. [2]).

Concepts and named entities. Many annotations may contain references to concepts and named entities (such as persons, organizations, places, events, …), either as a formal semantic description or just as a textual annotation. When combining metadata descriptions from different sources, it is important to ensure consistency between these annotations and the concepts being referenced. For example, the same person or place may be referred to with different names, names in different languages or just different spelling. This is one of the core problems of the Semantic Web and thus technologies from this area can be used for solving it.

Dependencies between metadata elements. Some elements of the metadata description may be related to other elements or based on them. For example, high-level metadata, such as classification of segments in terms of topics or genres, are – if they have been extracted using automatic content analysis – often based on low-level features of different modalities. Thus re-using only some of the channels of the source content may invalidate the high-level metadata.

Provenance and confidence. For the editing step it is in some cases relevant to know whether a metadata element from the source description has been extracted automatically or annotated manually. In order to know how to combine possibly conflicting descriptions, their degree of confidence must be known.

Context dependency. Context dependency of metadata descriptions is probably the hardest problem related to automatic metadata editing. The problem again mainly concerns high-level metadata and has a number of different facets.

One is related to metadata describing context and editorial aspects of the source material (of course this does not apply to unedited source material). There may be metadata describing the editorial structure of the material (scenes, stories, …). When parts of the material are re-used, their metadata are no longer relevant. The semantics of the source metadata descriptions have to be understood in order to discard this kind of metadata. In general, it will not be possible to automatically create this kind of metadata for the new content, unless sufficient metadata from the sources and the edit decisions are available to create it.

Similarly, metadata are often created based on domain knowledge, especially when using automatic content analysis tools. If material is used outside of the context of the original domain, the metadata may no longer be useful.

This leads to a very fundamental problem: Editing often determines or changes the semantics of content, independent of the context in which the source material has been taken. This holds for both combining content of one modality (such as shown for images by Kuleshov in the 1920s [1]) and for combining different modalities (e.g. the use of music to influence the viewer's perception of the visual impressions). Similarly, metadata elements describing the affective content of the material (cf. [6]) or interpreting it, are potentially invalidated by the editing operation.

3.2 Relevant Editing Operations

The following types of editing operations are considered relevant as input for metadata editing. The complexity of metadata editing depends crucially on the type of editing operation.

Cut. If a segment is delimited by a cut, the spatial extent of metadata does not need to be considered. This also refers to splits, which can be seen as cuts that are applied independently to different channels.

Gradual transitions. Gradual transition effects (such as dissolves, wipes, etc.) may create complex spatial scopes for the different source contents. However, as their duration is rather short, it is in practice sufficient to annotate the type of transition and not necessary to create detailed metadata for the transition segment.

Keying, matting, alpha blending. All these effects combine content in a way that parts of more than one source content are visible at a time. Thus the spatial scope of the metadata has to be taken into account in order to decide if a certain metadata element has to be put into the target metadata description or not.

Audio mixing. The problem seems to be similar to the visual effects above, but due to the lack of the spatial dimension, all of the source content is perceivable to a certain degree. Thus the target metadata description can be created in general by collating the source metadata descriptions.

3.3 Conclusions for the Design of the Metadata Editing Framework

The issues discussed above can be summarized as follows: Problems are caused by metadata elements that have a larger spatial/temporal/spatiotemporal scope and by edits that combine more than one source for some time range. The approach needed for editing thus depends crucially on the properties of the metadata elements. Thus the framework for metadata editing should be modular, in order to handle different types of metadata separately and use specific

approaches for creating the corresponding elements in the target metadata description.
We have seen from the discussion above, that for processing high-level metadata descriptions and merging descriptions of concepts and named entities the use of Semantic Web technologies is required. In order to automatically apply tools for reasoning and rule processing, a formalization of the semantics of the metadata model is required. Thus the metadata model presented in [12] will be used as unified metadata representation in the editing framework. The formalization of the semantics of this metadata model is currently in progress.

4 A Software Framework for Metadata Editing

4.1 Architecture

The software framework for metadata editing is designed in a modular way in order to keep it flexible and extensible. The modularity concerns the support of input formats for edit decision information, the metadata input formats and the editing tools for the different types of metadata. As has been discussed before, the editing operation may be quite different for different kinds of metadata, thus each kind of metadata is handled by specialized components.
The metadata editing framework consists of the following components:

- The metadata editing engine, which is the core and control component of the framework.
- A set of metadata editing plugins, which perform editing operations for certain types of metadata.
- A data structure for edit decisions and plugins for parsers for different data formats for edit decision information (e.g. EDL, AAF).

The data structure represents each channel of the target timeline as a list of segments. Each segment refers to one source or multiple sources (in case of gradual transitions and overlay effects). The data structure also handles associating source content with the corresponding essence files and metadata document.
The internal metadata representation is based on the MPEG-7 Detailed Audiovisual Profile (DAVP) [2]. The metadata documents are internally represented using the JRS MPEG-7 library [10]. The metadata editing engine and the plugins are discussed in more detail in the following.

4.2 Metadata Editing Engine

The metadata editing engine is the core component of the framework. It holds the representation of the edit decisions, the metadata documents for the source contents and the metadata document for the target content. The metadata editing engine handles some types of metadata by itself and controls the execution of the plugins.
Global technical metadata. Most of the technical metadata (e.g. resolution, sampling rates, encoding) cannot be taken from the sources but have to be newly generated for the target essence. If the rendered output of the editing process is already available, the technical metadata can be extracted from the target essence. Otherwise, if the edit decision information contains these parameters of the target essence, it can be taken from there.
Global descriptive metadata. Global metadata refer to all of the content (i.e. one of the source contents or the target content). There is no need for segment-wise merging, and the details of the edit decisions need not be taken into account for generating the global descriptive metadata for the output. Performing editing of global descriptive metadata is comparable to traditional metadata merging, i.e. harmonizing different metadata descriptions for the same content. This requires eliminating redundancies and checking for consistency and possible conflicts.
Due to the diverse types of global descriptive metadata they will be handled by specific plugins. For some types of metadata, merging global descriptive metadata is the same task as merging the corresponding type of descriptive metadata of a single temporal/spatial/spatiotemporal segment.
Time-/space-dependent descriptive metadata. The different types of time/space-dependent descriptive metadata will always be handled by the specific plugins. For some plugins, the execution sequence is relevant (for example, a description of the visual properties of a content may require the shot structure of the target description). The metadata editing engine will ensure that the editing plugins are executed in the right order.

4.3 Metadata Editing Plugins

A metadata editing plugin handles a specific type of metadata with well defined spatial and temporal scope and semantics. It takes splitting and merging decisions for the involved metadata segments. It shall be able to automatically resolve ambiguities and conflicts, and, if this is not possible, the plugin should report the possible problems so that the user can correct them in a later stage. There might be interdependencies between plugins, if there are interdependencies between metadata elements. They will be modelled by the order in which the plugins are executed by the metadata editing engine.
The plugins listed in the following are planned to be implemented in the metadata editing framework.
Shot structure and key frames. This plugin creates the shot structure and the associated key frame representation for the target timeline. The sources for the target shot structure are both the edit decisions (which create new shot boundaries) and the source metadata descriptions, as a source segment may already contain cuts and gradual transitions.
Low-level image features. Feature descriptions (e.g. colour, texture) of key frames or other single frames can be treated similarly as the (key) frames representing the shots, to which they are attached.
Camera motion and motion activity. Depending on the granularity of the description, the source segments can be split more or less precisely. Merging of target segments is not necessary, as both descriptions are naturally delimited by shot boundaries.

(Moving) objects and their trajectories. If static or moving objects are described in a more detailed way, such as describing their bounding rectangle or polygon and their trajectories, this information has to be taken into account during editing. Like for other visual feature descriptions, no target segments need to be merged.

Automatic speech recognition (ASR) transcripts. When editing is applied to one or more of the audio channels containing speech, ASR transcripts have to be edited. The problem is to detect the correct points to split the source segments, which might be hard, if no sufficient information about the temporal alignment is available.

Audio segment classification. Classification of audio segments usually describes the type of content of the segment, such as silence, speech, music, etc. As these types of content may be overlapping, multiple classifications may apply to one segment. The same is true for audio mixing, which unites the description of the audio channels of the source content.

Occurrence of named entities. This applies to descriptions of objects and concepts which are either visible or audible in the scene (appearing on screen, being the source of a sound or being explicitly named) or which are related to the scene without being perceivable. Editing requires splitting the description in the source content, which may only be related to a specific channel or modality.

It is necessary to map the named entity to a uniform identifier. Not only can there be different names referring to the same named entity, but the named entity may be referred to by a role within the context of the scene.

Global production metadata. Global production metadata (such as production place and time, the name of the director, etc.) of one content or clip applies to all of the segments stemming from this clip. The production metadata of the different source contents may have the same meaning, even if they are expressed differently. In the case that the descriptions from the different sources differ, the result of the editing process will be local production metadata for the segments, otherwise they can be collated into a new global metadata description for the target content.

Segment classification. Classification of segments may refer to various facets of the content, such as the genre, location (indoor/outdoor, cityscape/landscape), time of day or year, etc. Depending on the level of abstraction of the classification, automatic editing may or may not be possible.

The most important prerequisite is to model the set of classes and their relations, in order to be able to decide, which classifications refer to the same facet and may thus be contradicting, and which are complementary. Many types of classifications are not affected by the majority of editing operations. Editing operations such as keying may affect some classifications (e.g. when combining part of a shot classified as “indoor” with part of a shot classified as “outdoor”), while editing operations related to the temporal dimension will only have an effect on classifications with a high abstraction level. For example, the genre of a segment may be modified, when used in a different context.

5 Results

This section describes the implemented components of the framework presented in Section 4 and presents an example of metadata editing performed using the prototype implementation.

5.1 Implementation

Currently the supported formats are the MPEG-7 metadata standard and edit decision lists (EDLs). The implementation of the EDL parser reads the supported edit decisions from different flavors of EDLs (CMX, GVG and SMPTE 258M). It parses the associations between reel names and the actual essence file names, if available in the EDL file.

Besides the metadata editing engine, the plugins for merging shot and key frames and low-level visual features have been implemented.

5.2 Example

As input we have two short clips of rushes. For each of them, an MPEG-7 description has been created. The description contains key frames, which are sampled based on the visual activity in the content. One of the descriptions is shown in Figure 2, the other one looks likewise. Figure 3 shows the EDL representing the editing operation, which creates a new content made of parts from both of the source clips. The shot structure of the target material is generated automatically, as well as the description of representative key frames of the shots. Figure 4 shows the resulting MPEG-7 description for the edited content.

6 Conclusion and Future Work

Automatically applying edit decisions to the metadata of the source essence for generating a complete metadata description for the output is an important tool to establish a continuous metadata chain throughout the audiovisual media production process. We have analysed the problem of automatic metadata editing, taking into account specific properties of the different types of metadata elements. Metadata editing is a hard problem in some cases, but feasible if some conditions are met:

- The input metadata are described in a standard/format, for which a formal semantic definition is available.
- Provenance and confidence of the input metadata should be given.
- Unique identification of concepts and named entities significantly facilitate merging of description fragments.
- Editing operations that produce outputs combining more than one source content at the same time are not performed for all types of metadata, in particular those with a spatial dependency.

```
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
   <Description xsi:type="ContentEntityType">
      <MultimediaContent xsi:type="AudioVisualType">
         <AudioVisual>
            <MediaInformation>< !-- ... --></MediaInformation>
            <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:av.programme"/>
            <CreationInformation>< !-- ... --></CreationInformation>
            <MediaTime>
               <MediaTimePoint>T00:00:00:0F25</MediaTimePoint>
               <MediaDuration>PT00H00M25S14N25F</MediaDuration>
            </MediaTime>
            <MediaSourceDecomposition criteria="modalities">
               <VideoSegment id="clip1">
                  <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.programme"/>
                  <MediaTime>
                     <MediaTimePoint>T00:00:00:0F25</MediaTimePoint>
                     <MediaDuration>PT00H00M25S14N25F</MediaDuration>
                  </MediaTime>
                  <TemporalDecomposition gap="false" overlap="false" criteria="visual shots">
                     <VideoSegment xsi:type="ShotType" id="take1">
                        <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.shot"/>
                        <MediaTime>
                           <MediaTimePoint>T00:00:00:0F25</MediaTimePoint>
                           <MediaDuration>PT00H00M25S14N25F</MediaDuration>
                        </MediaTime>
                        <TemporalDecomposition criteria="key frames">
                           <VideoSegment id="take1_1">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:00:00F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="take1_2">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:05:20F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="take1_3">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:06:05F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="take1_4">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:08:15F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="take1_5>
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:22:03F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="take1_6>
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:23:01F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                        </TemporalDecomposition>
                     </VideoSegment>
                  </TemporalDecomposition>
               </VideoSegment>
            </MediaSourceDecomposition>
         </AudioVisual>
      </MultimediaContent>
   </Description>
</Mpeg7>
```

Figure 2: MPEG-7 description of a source clip.

```
TITLE: MetadataEditTest

001  CLIP1    V     C         00:00:06:00 00:00:08:23 00:00:00:00 00:00:02:23
* FROM CLIP NAME: clip1.avi
002  CLIP2    V     C         00:00:01:21 00:00:05:05 00:00:02:23 00:00:06:07
* FROM CLIP NAME: clip2.avi
```

Figure 3: The edit decision list used for metadata editing.

```
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" >
   <Description xsi:type="ContentEntityType">
      <MultimediaContent xsi:type="AudioVisualType">
         <AudioVisual>
            <MediaInformation>< !-- ... --></MediaInformation>
            <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:av.programme"/>
            <CreationInformation>< !-- ... --></CreationInformation>
            <MediaTime>
               <MediaTimePoint>T00:00:00:0F25</MediaTimePoint>
               <MediaDuration>PT00H00M6S7N25F</MediaDuration>
            </MediaTime>
            <MediaSourceDecomposition criteria="modalities">
               <VideoSegment id="videosegment-title">
                  <TemporalDecomposition criteria="visual shots" gap="0" overlap="0">
                     <VideoSegment id="CLIP1 take1" xsi:type="ShotType">
                        <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.shot"/>
                        <MediaTime>
                           <MediaTimePoint>T00:00:00:0F25</MediaTimePoint>
                           <MediaDuration>PT0H0M2S23N25F</MediaDuration>
                        </MediaTime>
                        <TemporalDecomposition criteria="key frames">
                           <VideoSegment id="CLIP1 take1 3">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:00:05F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="CLIP1_take1_4">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:02:15F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                        </TemporalDecomposition>
                     </VideoSegment>
                     <VideoSegment id="CLIP2 take2" xsi:type="ShotType">
                        <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.shot"/>
                        <MediaTime>
                           <MediaTimePoint>T00:00:02:23F25</MediaTimePoint>
                           <MediaDuration>PT0H0M3S9N25F</MediaDuration>
                        </MediaTime>
                        <TemporalDecomposition criteria="key frames">
                           <VideoSegment id="CLIP2 take2 2">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:03:15F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                           <VideoSegment id="CLIP2 take2 3">
                              <StructuralUnit href="urn:x-mpeg-7-davp:cs:StrUnitCS:2005:vis.keyframe"/>
                              <MediaTime>
                                 <MediaTimePoint>T00:00:05:24F25</MediaTimePoint>
                              </MediaTime>
                           </VideoSegment>
                        </TemporalDecomposition>
                     </VideoSegment>
                  </TemporalDecomposition>
               </VideoSegment>
            </MediaSourceDecomposition>
         </AudioVisual>
      </MultimediaContent>
   </Description>
</Mpeg7>
```

Figure 4: Generated MPEG-7 description of the edited output.

- Metadata describing the context or the affective content are not taken into account.

We have presented a software framework for metadata editing that is flexible enough to fulfil the diverse requirements for editing the different types of metadata. A central building block of the framework is the unified metadata representation which we have based on the MPEG-7 Detailed Audiovisual Profile (DAVP).

During the analysis of the metadata involved we have seen, that using Semantic Web technologies is necessary, as soon as we are dealing with semantic metadata such as textual annotations, annotations of named entities, etc. In these cases it is not possible to follow simple rules in order to combine the metadata elements, but the semantics of the metadata elements have to be understood in order to perform correct editing. The plugins that will be developed for these types of metadata will thus make use of these technologies. This also requires the formal description of the semantics of the unified metadata representation, which is currently in progress.

Acknowledgements

The authors would like to thank Andreas Hofmann for the implementation of the EDL parser as well as Roland Mörzinger, Herwig Zeiner and Werner Haas for their feedback and support. This work has been funded partially under the 6th Framework Programme of the European Union within the IST project "IP-RACINE" (IST-2-511316, http://www.ipracine.org).

References

[1] S. Ascher and E. Pincus, *The Filmmaker's Handbook*, revised edition, Plume 1999.

[2] W. Bailer and P. Schallauer, "The Detailed Audiovisual Profile: Enabling Interoperability between MPEG-7 Based Systems". *Proceedings of 12th International Multi-Media Modeling Conference*, IEEE Press, pp. 217-224, Beijing, CN, 2006.

[3] J. Casares, A. Chris Long, B. A. Myers, R. Bhatnagar, S. M. Stevens, L. Dabbish, D. Yocum and A. Corbett, "Simplifying video editing using metadata", *Proceedings of the Conference on Designing Interactive Systems*, pp. 157—166, London, UK, 2002.

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

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

[6] A. Hanjalic, *Content-based analysis of digital video*, chapter 5, Kluwer, 2004.

[7] IP-RACINE project website. URL: http://www.ipracine.org (Jul. 4, 2006).

[8] C. L. Madhwacharyula, M. S. Kankanhalli and P. Mulhem, "Content based editing of semantic video metadata," *IEEE International Conference on Multimedia and Expo 2004*, vol. 1, pp. 33-36, June 2004.

[9] Information Technology—Multimedia Content Description Interface. ISO/IEC 15938:2001.

[10] JOANNEUM RESEARCH MPEG-7 Library. URL: http://mpeg-7.joanneum.at (Jul. 4, 2006).

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

[12] P. Schallauer, W. Bailer and G. Thallinger: "A Description Infrastructure for Audiovisual Media Processing Systems Based on MPEG-7", *Journal of Universal Knowledge Management*, Special Issue on Multimedia Metadata, 2006 (to appear).

[13] SMPTE Metadata Dictionary, SMPTE RP210.8-2004.