Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

INS

Information Systems

INformation Systems

VAMP: semantic validation for MPEG-7 profile descriptions

R. Troncy, W. Bailer, M. Hausenblas, M. Höffernig

REPORT INS-E0705 APRIL 2007

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3681

VAMP: semantic validation for MPEG-7 profile descriptions

ABSTRACT
MPEG-7 can be used to create complex and comprehensive metadata descriptions of multimedia content. Since MPEG-7 is defined in terms of an XML schema, the semantics of its elements has no formal grounding. In addition, certain features can be described in multiple ways. MPEG-7 profiles are subsets of the standard that apply to specific application areas and that aim to reduce this syntactic variability, but they still lack formal semantics. We propose an approach for expressing the semantics explicitly by formalizing the constraints of various profiles using ontologies and logical rules, thus enabling interoperability and automatic use for MPEG-7 based applications. We have implemented VAMP, a full semantic validation service that detects any inconsistencies of the semantic constraints formalized. Another contribution of this paper is an analysis of how MPEG-7 is practically used. We report on experiments about the semantic validity of MPEG-7 descriptions produced by numerous tools and projects and we categorize the most common errors found.

2000 Mathematics Subject Classification: H.5.1 Multimedia Information Systems
1998 ACM Computing Classification System: H.5.1 Multimedia Information Systems
Keywords and Phrases: VAMP, MPEG-7 Validation Service, MPEG-7 Semantic Constraints

VAMP: Semantic Validation for MPEG-7 Profile Descriptions

Raphaël Troncy[1], Werner Bailer[2],
Michael Hausenblas[2] and Martin Höffernig[2]

[1] CWI Amsterdam, P.O. Box 94079,
1090 GB Amsterdam, The Netherlands
raphael.troncy@cwi.nl
[2] JOANNEUM RESEARCH Forschungsgesellschaft mbH,
Institute of Information Systems and Information Management,
Steyrergasse 17, 8010 Graz, Austria
firstName.lastName@joanneum.at

Abstract

MPEG-7 can be used to create complex and comprehensive metadata descriptions of multimedia content. Since MPEG-7 is defined in terms of an XML schema, the semantics of its elements has no formal grounding. In addition, certain features can be described in multiple ways. MPEG-7 profiles are subsets of the standard that apply to specific application areas and that aim to reduce this syntactic variability, but they still lack formal semantics. We propose an approach for expressing the semantics explicitly by formalizing the constraints of various profiles using ontologies and logical rules, thus enabling interoperability and automatic use for MPEG-7 based applications. We have implemented *VAMP*, a full semantic validation service that detects any inconsistencies of the semantic constraints formalized. Another contribution of this paper is an analysis of how MPEG-7 is practically used. We report on experiments about the semantic validity of MPEG-7 descriptions produced by numerous tools and projects and we categorize the most common errors found.

1 Introduction

The amount of multimedia data being produced, processed and consumed is growing, as is the number of applications dealing with multimedia content. In many of these applications, metadata descriptions of the content are important. MPEG-7 [15], formally named Multimedia Content Description Interface, is designed as a standard for representing these descriptions in a broad range of applications. In order to cover diverse requirement scenarios [22], many *descriptors* and *descriptions schemes*, as well as the relationships between them, have been defined. The descriptors and description schemes are together referred to as *description tools*, and a *description* is a particular instantiation of these. There are description tools for diverse types of annotations on different semantic levels, ranging from very low-level features, such as visual (e.g. texture, camera motion) or audio (e.g. spectrum, harmonicity), to more abstract descriptions.

The flexibility of MPEG-7 is based on allowing descriptions to be associated with arbitrary multimedia segments or regions, at any level of granularity, using different levels of abstraction. The downside of the breadth targeted by MPEG-7 is its complexity and its fuzziness [4, 20, 27]. For example, very different syntactic variations may be used in multimedia descriptions with the same intended semantics, while remaining valid MPEG-7 descriptions. Given that the standard does not provide a formal semantics for these descriptions, this syntax variability causes serious interoperability issues for multimedia processing and exchange, for example on the Web.

To reduce this syntax variability, MPEG-7 has introduced the notion of *profiles* that constrain the way multimedia descriptions should be represented for particular applications. Profiles are therefore a way of reducing the complexity of MPEG-7 (i.e. only a subset of the whole standard can be used) and of solving some interoperability issues (i.e. English guidelines are provided on how the descriptors should be used and combined). However, these additional constraints are only represented with XML Schema [29], and, for most of them, cannot be

automatically checked for consistency by XML processing tools. In other words, profiles provide only very limited control over the semantics of the MPEG-7 descriptions [11, 19, 25]. Because of this lack of formal semantics, the resulting interoperability problems prevent an effective use of MPEG-7 as a language for describing multimedia.

In this paper, we present VAMP[1], a semantic VAlidation service for MPEG-7 Profiles. VAMP generalizes the method we proposed for the single DAVP profile [26] by formalizing how MPEG-7 descriptors should be used in commonly-used profiles. In contrast to other work [1, 8, 11, 28], we do not intend to completely map the MPEG-7 description tools onto an OWL ontology [7, 14], but rather use Semantic Web technologies to represent those MPEG-7 semantic constraints defined in natural language that cannot be expressed using XML Schema. We have also gathered and analyzed numerous MPEG-7 descriptions generated by various tools. We report in this paper on how semantically valid these descriptions are and we provide a categorization of the most common interoperability problems we found.

The paper is organized as follows. In the next section, we briefly introduce the notion of MPEG-7 profiles and we analyze several MPEG-7 descriptions generated by various tools. In section 3, we provide a categorization of the most common interoperability problems encountered. In section 4, we present the VAMP service and we detail how the MPEG-7 profiles can be formalized, building first an OWL ontology and rules capturing the semantic constraints, and developing tools converting the XML-based MPEG-7 descriptions to RDF triples. In section 5, we compare our approach with other attempts to formalize the MPEG-7 knowledge and we discuss the scope of our methodology before concluding the paper (section 6).

[1] VAMP is available as a web application at `http://vamp.joanneum.at` and as a web service.

2 MPEG-7 Usage Analysis

The MPEG-7 XML Schema defines numerous elements and types, as well as rules for their valid combinations. The standard, however, allows the specification of different descriptions with equivalent semantics. This raises interoperability problems when exchanging MPEG-7 descriptions since applications may use the standard differently. For example, the same decomposition of a video into shots and key frames can be represented by multiple MPEG-7 descriptions [26].

This problem has been recognized by both the MPEG working group and the various tools that partially support the standard. Profiles have thus been proposed as a possible solution. In the following, we first introduce the notion of *profiles* (section 2.1) and we then show how several multimedia annotation tools (section 2.2) address this interoperability problem by reducing and further constraining the MPEG-7 description tools.

2.1 Profiling MPEG-7

The specification of a profile consists of three parts, namely [16]: i) *description tool selection*, i.e. the definition of the subset of description tools to be included in the profile, ii) *description tool constraints*, i.e. definition of constraints on the description tools such as restrictions on the cardinality of elements or on the use of attributes, and iii) *semantic constraints* that further describe the use of the description tools in the context of the profile.

The first two parts of a profile specification are used to address the *complexity* problem, that is, the complexity of a description that can be measured by its size or the number of descriptors used. Limiting the number of descriptors and description schemes (either by excluding elements or constraining their cardinality) reduces this complexity. Both the selection and the usage constraints of the description tools are specified using the MPEG-7 DDL. They result in a specific and more constrained XML Schema. The third part of a profile specifi-

cation tackles the *interoperability* problem. Semantic constraints are expressed in natural language to clarify the ambiguities associated with the use of the remaining description tools selected in the first two parts. This informal specification of the constraints, however, prevents an automated process from checking the correct use of MPEG-7 profiles for describing multimedia content.

Five MPEG-7 profiles are currently in widespread use: three of them have been defined in Part 9 of the standard[2] [17], and we consider two other "de-facto" profiles, not (yet) standardized, but also widely used by the multimedia community:

Simple Metadata Profile (SMP) describes single instances or collections of multimedia content as entire entities or clips with textual metadata only and no spatial decomposition. The motivation of this profile is to support simple metadata tagging similar to ID3[3] for music and EXIF[4] for images, and to support mobile applications such as 3GPP[5]. A partial mapping from these vocabularies to SMP has been specified.

User Description Profile (UDP) consists of tools for describing the personal preferences and usage patterns of users of multimedia content in order to enable automatic discovery, selection, personalization and recommendation of multimedia content. This profile contains all MPEG-7 description tools that were adopted by the TV-Anytime Forum, and are referenced by the TV-Anytime Metadata specification [23].

Core Description Profile (CDP) consists of tools for describing general multimedia content such as images, videos, audio and collections using the top-level types defined in Part 5. A typical use of this profile is the description of the structural and semantic aspects of video content of a TV program and its corresponding materials. This includes managing the media materials, distributing them and archiving them. Just as the two

[2]Five other profiles are discussed in [17] but have been later merged or withdrawn.
[3]`http://www.id3.org/`
[4]`http://www.exif.org/`
[5]`http://www.3gpp.org/`

previous profiles, it does not include the visual and audio descriptors defined in Parts 3 and 4 of MPEG-7.

Detailed Audio-Visual Profile (DAVP) describes single multimedia content entities, based on a comprehensive structural description of the content and including all audio and visual low-level feature descriptors.

TRECVID Profile gathers the descriptors used for representing the results of the TREC Video Retrieval Evaluation[6] yearly competition. It describes the shot structure of a video and the key frames representing each shot.

Profile	**Descriptors**	**Semantic Constraints**
Simple Metadata Profile (SMP)	45	6 + 0
User Description Profile (UDP)	102	8 + 0
Core Description Profile (CDP)	153	27 + 2
Detailed Audio-Visual Profile (DAVP)	274	35 + 50
TRECVID Profile	30[a]	4 + 2

Table 1: The number of MPEG-7 descriptors and semantic constraints specified in each profile

[a]This number is an approximation based on the instance descriptions, since the TRECVID schema has never been specified.

These five profiles put different emphasis on the *complexity* and *interoperability* problems mentioned above. For each profile, we have counted the number of descriptors and we have evaluated the number of semantic constraints it contains (Table 1). More precisely, for each descriptor included in a profile, we looked at its informal semantics written in English in the standard, and we examine the constraints that cannot be represented with XML Schema. Therefore, our evaluation considers both the original MPEG-7 constraints and those specified additionally in the profiles. We observe that the standardized profiles aim at complexity reduction and hence significantly reduce the included set of allowed descriptors (with respect to the 1200 MPEG-7 elements) while defining few semantic constraints. In contrast, DAVP excludes some descriptors such

[6]`http://www-nlpir.nist.gov/projects/trecvid/`

as the user preferences or the collection description schemes, but keeps most of the others [4]. The focus is on the definition of the semantic constraints for the remaining descriptors included in the profile. Similarly, the TRECVID profile has reduced the set of descriptors to those applicable to its specific application area and agreed upon the use of these descriptors.

2.2 Gathering MPEG-7 Descriptions

The W3C Multimedia Semantics Incubator Group maintains a comprehensive list[7] of tools that can generate MPEG-7 descriptions. These tools do not necessarily comply with a profile, but they also try to address the interoperability problem by further constraining the subset of descriptors they support. This complexity reduction, however, comes often with the price of having hard-coded constraints instead of explicit semantics. We present a selection of these tools, categorized according to their predominant media type (image, audio and video), although some of them can handle multiple media.

2.2.1 Image Related Tools

Caliph & Emir[8] is a semi-automatic annotation tool for images that supports free text and graph-based semantic annotations as well as a number of visual feature extractors. Furthermore, pre-existing metadata, such as EXIF or IPTC tags inside images, is converted into MPEG-7 following the mapping rules given in the SMP profile.

The M-OntoMat-Annotizer[9] supports the manual annotate regions of still images, linking RDF(S) domain specific ontologies to low-level MPEG-7 visual descriptors. The semantics of these visual descriptors is formalized in a Visual Descriptor Ontology (VDO) represented in RDFS [2].

[7] http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources
[8] http://www.semanticmetadata.net/features/
[9] http://www.acemedia.org/aceMedia/results/software/m-ontomat-annotizer.html

2.2.2 Audio Related Tools

The MPEG-7 Audio Analyzer[10] implements all 17 low-level audio descriptors defined in Part 4, while the MPEG-7 Spoken Content Demonstrator[11] generates the output of an Automatic Speech Recognition (ASR) system using the `SpokenContent DS`, which is composed of around 20 descriptors.

The MPEG-7 Audio Encoder[12] allows also to extract all the audio descriptors, but it further constrains their use in two new XML Schema.

2.2.3 Video Related Tools

IBM VideoAnnex[13] is a semi-automatic annotation tool for videos that generates temporal shot segmentation and supports spatial decomposition of key frames. The annotations make use of controlled vocabularies defined using the `ClassificationScheme DS` (see Part 5 of [15]).

Frameline 47[14] uses an advanced content schema based on MPEG-7 so as to be able to annotate entire video files, or segments and groups of segments from within that video file.

Muvino[15] is a very simple tool for manually annotating videos (free text annotation and keyword based). It supports some general metadata about the video, the temporal decomposition into segments and some semantic descriptors such as place and time.

The Metadata Production Framework (MPF) [18] proposed by NHK is an industrial application of the Core Description Profile (CDP). The authors address the complexity and ambiguity problems of MPEG-7 by proposing a metadata model that further restricts CDP by excluding some elements and reducing the cardinality of others. Although they recognize the problem of the lack of semantics in the standard, the descriptions of "meaning of each descriptor in MPF"

[10] `http://mpeg7lld.nue.tu-berlin.de/`
[11] `http://mpeg7spkc.nue.tu-berlin.de/`
[12] `http://mpeg7audioenc.sourceforge.net/`
[13] `http://www.research.ibm.com/VideoAnnEx`
[14] `http://frameline.tv/`
[15] `http://vitooki.sourceforge.net/components/muvino/code/index.html`

only marginally goes beyond the textual description of the semantics one can find in MPEG-7. They describe, however, the "model definition policy" used for the specification of MPF, which contains some basic design criteria and definitions of the semantics of the main parts in the structure of the MPF data model. As the data format specification is not the only element of MPF, but there is also a Metadata Editor application, the semantic constraints are directly hard-coded into the application.

2.3 Summary

We have collected a large set of sample descriptions in order to analyse how MPEG-7 is used in practice. These examples cover a broad range of applications and use different subsets of MPEG-7 descriptors. Profiles are sometimes used (and even further constrained) or could have been specified from the scope of the application. The interoperability problems, however, cannot be solved by just extending the XML schema and the semantics is often directly hard-coded in the tools. We argue that true interoperability can be obtained if the semantics is made explicit and can be formally checked for consistency.

Some tools generate errors. For example, the IBM VideoAnnex tool automatically produces shot lists of videos. For some video clips the tool produces shot segments with a negative duration, or overlapping segments, even though the `overlap` attribute of the `TemporalDecomposition` has the value `false`. The resulting description will validate according to the XML Schema (of MPEG-7 or one of the profiles) but will not be semantically valid.

We have analyzed from these MPEG-7 descriptions the possible errors and identified the semantic constraints that need to be formalized. We detail these errors in the next section and present how the interoperability problem is solved in VAMP.

3 Interoperability Problems

In this section, we summarize the errors that we found, which we discuss in three categories: the inconsistencies related to the temporal information (section 3.1), the media information (section 3.2), and the semantic information (section 3.3). All the violations discussed here yield perfectly valid documents with respect to the MPEG-7 XML schema but raise inconsistencies with the semantic constraints that express the intended semantics of the standard.

3.1 Temporal-related Violations

The representation of time is an essential component for media having a temporal dimension. MPEG-7, however, defines only a simple syntactic pattern for representing the time points and the time durations. We present common inconsistencies underlying this representation as well as the possible misuse of the temporal decomposition descriptors. We advocate then an alternative time representation.

3.1.1 Common violations

The ISO 8601 standard is generally considered as the reference "specification of the representation of dates in the proleptic Gregorian calendar[16] and times and representations of periods of time" [13]. The corresponding datatypes in XML Schema use lexical formats inspired by the ISO standard and include some deviations such as an optional minus sign in the lexical representation, the possibility of having more than 9999 years or the inclusion of a time zone [29]. Unfortunately, these datatypes are not used in MPEG-7, which instead, redefines a simple pattern format for the media time point:

```
<simpleType name="mediaTimePointType">
  <restriction base="mpeg7:basicTimePointType">
    <pattern value="(\-?\d+(\-\d{2}(\-\d{2})?)?)?(T\d{2}(:\d{2}(:\d{2}
                    (:\d+)?)?)?)?(F\d+)?"/>
```

[16]The proleptic Gregorian calendar includes dates prior to 1582 (the year it came into use as an ecclesiastical calendar).

```
  </restriction>
</simpleType>
```

and for the media duration:

```
<simpleType name="mediaDurationType">
  <restriction base="mpeg7:basicDurationType">
    <pattern value="\-?P(\d+D)?(T(\d+H)?(\d+M)?(\d+S)?(\d+N)?)?(\d+F)?"/>
  </restriction>
</simpleType>
```

Based on this decision, the following inconsistencies can be observed:

Invalid time specification. MPEG-7 introduces different new lexical patterns to represent media times and real-world dates and times. The patterns definition allows the specification of invalid dates and times. For example, 31^{st} of February would be a valid date according to the time point pattern shown above. Another shortcoming deals with the frame precision in the media time pattern: for example `T00:01:23:27F25` would be a valid time point whereas it points to the fraction 27 of 25 that is impossible to compute. Similarly, a fraction rate of 0 cannot be computed but could still be represented with this pattern.

Negative segment duration. MPEG-7 segments are described by a start time point and a duration. The optional minus sign of the patterns allows negative duration for segments in a temporal decomposition while this would make no sense.

Inconsistent temporal decomposition A temporal decomposition of a segment into subsegments is only meaningful if the the time range filled by each of the subsegments is at most the time range of the segment being decomposed, i.e. a part of a temporal segment cannot start before or end after its parent segment.

Gap and overlap A temporal decomposition can be qualified whether the subsegments in the decomposition overlap or have gaps between them. These properties are specified with the `gap` and `overlap` attributes of the decomposition that have a `true/false` value. There is, however, no mechanism to check whether the actual time description of the segments conforms to the value of the attribute or not.

Formalizing the representation of dates and times, for example using OWL-Time [9] solves some of these problems. The 8-ary predicate `duration` is converted into eight binary relations, which are more convenient for description logic-based markup languages such as OWL, so that the consistency of the time specification can be checked.

3.1.2 Time Representation

Representing durations smaller than one second becomes a problem when different sampling rates are involved and precise time points need to be computed. Conversion between time points and durations specified with respect to different sampling rates requires converting them to a common representation. This representation can be the second, with the drawback of using floating point numbers and possible precision loss, or some defined sampling rate. In some frameworks, the millisecond or nanosecond unit (e.g. in Microsoft DirectShow) is used. Choosing an arbitrary integer sampling rate f as a common representation leads to rounding errors for rates that are coprime[17] with f.

Using the least common multiple of the sampling rates involved is a solution to this problem [27]. For example, considering video frame rates of 24, 25 and 30 and audio sampling rates of 44,100 and 96,000 this leads to a common sampling rate of 14,112,000. This approach is used in the DETECT content analysis framework [24] and in the FERIA framework [6]. When all sampling rates involved are not known in advance or when there are many different rates

[17] In mathematics, two integers a and b are said to be coprime or relatively prime if their greatest common divisor is 1.

to consider, this approach is however impractical. For example, considering the exact NTSC frame rate of 30000/1001 will increase the factor calculated above to 4.2336×10^{11}, which is already about 100 times beyond the value range of a 32bit integer. If only comparisons and simple calculations are needed, then using floating point numbers or a reasonably large common sampling rate will lead to acceptably small errors.

3.1.3 Analogy with space representation

Similar to the temporal decomposition, the spatial and the spatio-temporal decompositions suffer from the same limitations in MPEG-7. For example, if a region of an image is decomposed into subregions, the subregions must lie inside the parent region. The violations related to the values of the `gap` and `overlap` attributes can thus also be raised. Consistency checking is, however, much more difficult to implement than for the time ranges due to the two-dimensional nature of the regions.

3.2 Media Information-related Violations

The description of information about properties of the media can be specified at multiple places in MPEG-7. While the presence and cardinality of the elements can be controlled using XML Schema, the semantics between the global media information and the actual description can mismatch. The following inconsistencies can thus be observed:

Inconsistent media content types. The `Content` element in `MediaFormat` is used to describe the content type of the medium being described (e.g. image, video), using a reference to a classification scheme. The same information is contained in the type of the `MultimediaContent` element of the description but these two values can mismatch. For example, the `xsi:type="ImageType"` specifies the multimedia content being described, but the `MediaFormat` could be stated as audio.

Inconsistent modality information. The `MediaProfile` describes the visual and audio encoding (e.g. a master quality and a low resolution preview), or each stream if several streams in different encoding are available. This information must also match the content type, but again, there is no way to check that the values are consistent. For example, different modalities can be present in the structural description (e.g. one video and two audio channels) even though the media information contains contradicting information about the modalities (e.g. states that the content is mono-audio).

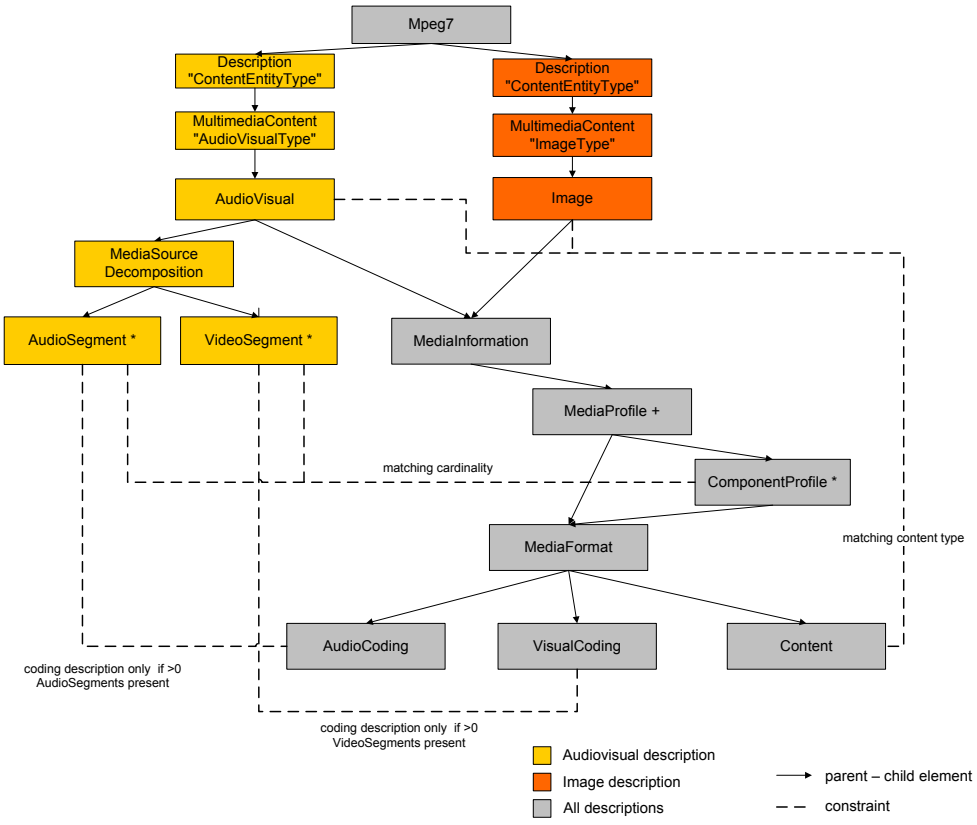

Figure 1: Example of a media information violation

Figure 1 shows the semantic constraints between the elements of the media information description and the top-level elements segments in the audiovisual description. The dashed lines indicate potential violations.

3.3 Classification Scheme-related Violations

An MPEG-7 `ClassificationScheme` is a generic mechanism for defining multilingual and controlled vocabularies. The set of terms and definitions belonging to a scheme is organized in a taxonomy, and is identified by a URI to be further referenced as values for descriptors. Part 5 of the standard already defines some basic classification schemes, e.g. for enumerating the media types, the different encoding, or some TV genres.

The appropriateness of a classification scheme in a certain context is a source of possible violations of the semantic constraints. More precisely, the `ClassificationSchemeBaseType` has two attributes: `uri` which identifies the classification scheme and `domain` which gives a list of XPath expressions containing the MPEG-7 description schemes that can reference the terms of the scheme. A description, however, can contain unforeseen descriptors using terms from this scheme, i.e. the classification scheme does not contain appropriate terms for the context in which it is used.

Once a classification scheme is dereferenced, the terms identified might not be retrieved, i.e. there are broken links. A classification scheme can also import other classification schemes which makes the task of resolving the referenced terms more difficult.

The errors detailed in this section cannot be checked with XML Schema validators. Semantic constraints are defined informally in the standard and cannot be processed by automated tools. We therefore propose a method for formalizing these constraints, implemented in the VAMP service.

4 VAMP: A Semantic Validation Service for MPEG-7 Descriptions

The violations of the semantic constraints trigger interoperability problems even though the result is perfectly valid MPEG-7 descriptions. We had already analyzed the semantic constraints of the Detailed Audiovisual Profile (DAVP) and formalized a subset of them [26]. Here, we generalize this approach to all profiles and present VAMP, a validation service for MPEG-7 semantic constraints (section 4.1). We show that the formalization of the semantic constraints amounts to explicitly capturing the semantics of a given profile as well as some additional logical rules (section 4.2). Finally, we describe the implementation of the VAMP service, available as a web interface for humans, and as a REST-style Web service for agents (section 4.3).

4.1 General Methodology

We propose the following layered approach to validate *semantically* the conformance of MPEG-7 descriptions to a given profile:

XML/syntactic: well-formedness. The well-formedness[18] of the input description is verified;

XML/syntactic: validity. The XML validity of the input description against the MPEG-7 schema and possibly a profile schema is checked;

RDF/semantics: constraints. The consistency of the input description with the ontology and logical rules formalizing the semantic constraints of a profile is computed.

Figure 2 depicts these various steps in the VAMP service. We propose the use of Semantic Web languages to formalize the semantic constraints, and later inference tools to check the semantic consistency of the descriptions. This is carried out with an appropriate combination of the following languages [12]:

[18] `http://www.w3.org/TR/REC-xml/#sec-well-formed`

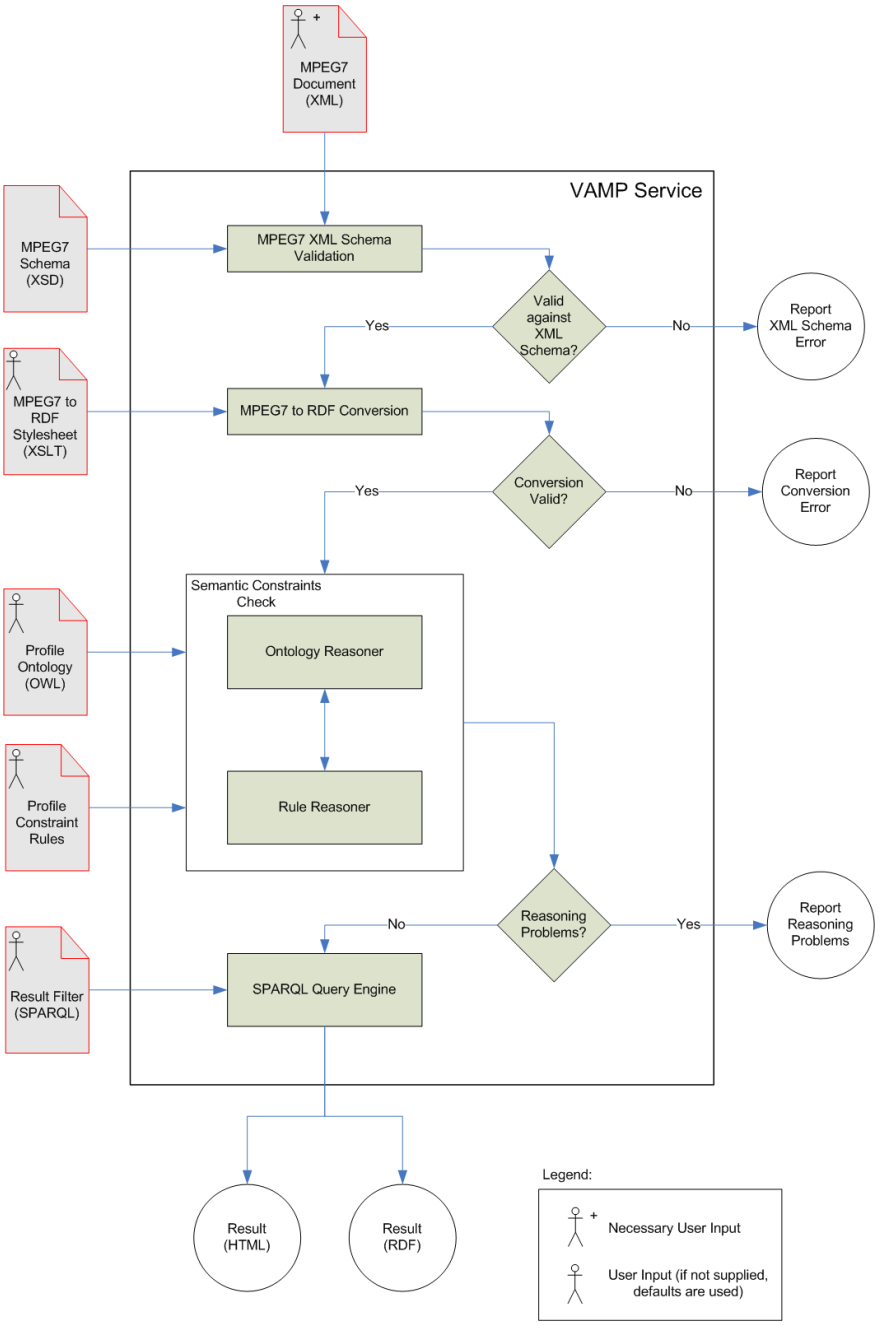

Figure 2: General architecture of the VAMP service

- XML Schema [29] to define the structural constraints, that is, which types are allowed and how they can be combined;
- OWL-DL [7] to formally capture the intended semantics of the descriptors contained in a profile which have semantic constraints;
- Horn clauses [5] to express relationships between syntactically different but semantically equivalent descriptors;
- XSLT to convert MPEG-7 descriptions into RDF depending to the profile. The RDF data asserts the class-membership of particular descriptors given their properties.

Achieving interoperability for MPEG-7 descriptions thus requires formally describing the profile the MPEG-7 description purports to adhere to, and converting automatically from the descriptions, the instances of the concepts modeled. We now show an example of formalization of temporal semantic constraints.

4.2 Formalizing the MPEG-7 Semantic Constraints

Table 2 reproduces the informal semantics of the MPEG-7 `SegmentDecompositionType` descriptor. This example illustrates typical constraints that cannot be checked with XML processing tools and need to be formalized.

4.2.1 Modeling Semantic Constraints with an Ontology Language

Figure 3 gives a partial formalization of the `SegmentDecompositionType` descriptor in the OWL Abstract Syntax (OWL-AS) [21]. It starts with the namespaces declaration, followed by the definition of the concepts used. The (optional) `criteria` attribute is modeled through the object property `davp:hasCriteria`, taking another concept as its value (`davp:Criteria`), depending on the criteria type (e.g. camera motion). The `gap` and `overlap` attributes are represented through two datatype properties (`davp:hasGap` and `davp:hasOverlap` respectively).

The XML representation of the description can then be converted into RDF using the ontology capturing the semantics of the profile. The OWL-DL expressivity is, however, insufficient for capturing all the semantic constraints. For example, the boolean values of the `gap` and `overlap` attributes can mismatch their actual truth values based on the actual time points delimiting the segments. Logic programming [5] and specifically Horn clauses are able to check the consistency of such information.

Name	***Definition***
`SegmentDecompositionType`	Describes decompositions of segments (abstract). The specialized segment decomposition tools extend the `SegmentDecomposition DS` `SegmentDecompositionType` extends `DSType`. A segment decomposition requires that the union of the extents defined by the sub-segments does not extend beyond the extents defined by the parent segment.
`criteria`	Indicates the criteria used in the segment decomposition (optional). Examples of criteria include "color homogeneity" and "camera motion type."
`overlap`	Indicates whether or not the segments resulting from the segment decomposition overlap in space and/or time. This attribute value is "false" by default.
`gap`	Indicates whether or not the segments resulting from the segment decomposition leave gaps with respect to the parent segment. A segment decomposition has gaps if the union of the child sub-segments does not correspond exactly to the parent segment. This attribute value is "false" by default.
`TemporalSegment-DecompositionType`	Abstract type from which the specialized temporal segment decomposition DSs are derived. The `TemporalSegmentDecomposition DS` describes a temporal decomposition of a segment. `TemporalSegmentDecompositionType` extends `SegmentDecompositionType`.

Table 2: Intended semantics of the `SegmentDecompositionType` from [15], Part 5, page 257

4.2.2 Modeling the Additional Knowledge using Rules

The logical rules depicted in Figure 4 are used to detect the temporal segments which start earlier than their parent segments, which would violate a temporal semantic constraint.

```
Namespace(rdf  = <http://www.w3.org/1999/02/22-rdf-syntax-ns#>)
Namespace(xsd  = <http://www.w3.org/2001/XMLSchema#>)
Namespace(rdfs = <http://www.w3.org/2000/01/rdf-schema#>)
Namespace(owl  = <http://www.w3.org/2002/07/owl#>)
Namespace(davp = <http://iis.joanneum.at/mpeg-7/davp#>)

 Class(davp:DSType partial)
 Class(davp:SegmentDecompositionType partial
       davp:DSType)
 Class(davp:TemporalSegmentDecompositionType partial
       davp:SegmentDecompositionType)
 Class(davp:Criteria partial)

 ObjectProperty(davp:hasCriteria
                range(davp:Criteria))
 DatatypeProperty(davp:hasGap
                  domain(davp:TemporalSegmentDecompositionType)
                  range(xsd:boolean))
 DatatypeProperty(davp:hasOverlap
                  domain(davp:TemporalSegmentDecompositionType)
                  range(xsd:boolean))
```

Figure 3: Formalization of `SegmentDecompositionType` in OWL

`add_media_time_point`. The media time point of a segment is described with the `davp:hasMediaTimePoint` property from the profile ontology. The property value is `xsd:string` representing a time point in the ISO 8601 format (see section 3.1.2). For comparing time points, we convert the representation in seconds with `calculateMediaTimePointInSeconds` function. The rule then produces a new RDF triple with the property `davp:hasMediaTimePointInSeconds` for any subject which has a `davp:hasMediaTimePoint` property, along with a typed literal object (the media time point in seconds, `xsd:double`).

`check_media_time_points`. The second rule compares then two media time points. If the media time point from a child segment (`?child`) is less than the media time point from its parent segment (`?parent`), then an error is flagged and typed (`davp:MediaTimePointError`) to be further processed in order to give a meaningful explanation of the violation to the end user.

```
@prefix davp: <http://iis.joanneum.at/mpeg-7/davp#>.
@prefix ex: <http://iis.joanneum.at/mpeg-7/davp/example#>.
 ...

 [add_media_time_point:
  (?segment davp:hasMediaTimePoint ?MTPString),
   noValue(?segment davp:hasMediaTimePointSec),
   calculateMediaTimePointInSeconds(?MTPString, ?MTPNumInSec)
   ->
  (?segment omp:hasMediaTimePointSec ?MTPNumInSec)
 ]

 [check_media_time_points:
  (?child davp:hasParent ?parent),
  (?child davp:hasMediaTimePointSec ?MTPChild),
  (?parent davp:hasMediaTimePointSec ?MTPParent),
  lessThan(?MTPChild, ?MTPParent)
  ->
  (?child davp:hasError davp:MediaTimePointError)
 ]
```

Figure 4: Formalization of `SegmentDecompositionType` with additional Horn clauses

4.2.3 Semantic Constraints and Reasoning

Once the semantic constraints have been formalized, they need to be checked for consistency. In contrast to the Semantic Web, VAMP is a closed system. Actually, we assume that all information needed to validate an MPEG-7 description is available: in the MPEG-7 input document itself, in the profile-dependent transformation, in the semantic constraints profile ontology and in the semantic constraints profile rule base.

As a Description Logic [3] language, OWL is based on an *open world assumption* with "negation as unsatisfiability", that is, something is false if and only if it contradicts existing information. In contrast, rule-based systems tend to be based on a *close world assumption*. It is possible to use both OWL and rules by explicitly closing the world [10]. Such a need exists in VAMP, for example for checking the minimal cardinality of numerous descriptors in MPEG-7 descriptions[19].

[19]See ongoing discussion about this topic, at `http://lists.owldl.com/pipermail/pellet-users/2007-March/001355.html`.

4.3 Implementation

This methodology has been implemented in the VAMP service, available as a web interface for humans and as a REST-style Web service for agents. For the RDF processing, Jena 2.4[20] is used. The validation of the semantic constraints is done via a DIG-connected Pellet[21] reasoner. Jena rules[22] provide for a sound, and integrated reasoning system that allows for both forward and backward reasoning.

4.3.1 VAMP as a Semantic Web Application

The interface for a human user is the VAMP Web interface, depicted in Figure 5. The web application uses Ajax and Java servlet technologies.

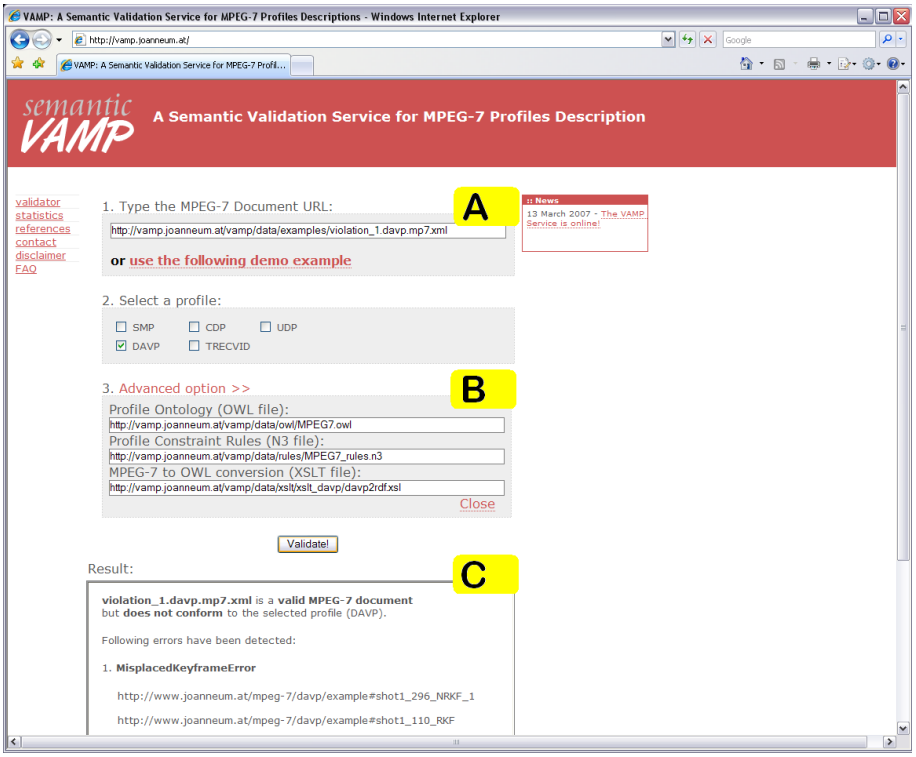

Figure 5: The VAMP Web interface

[20] http://jena.sourceforge.net/
[21] http://pellet.owldl.com/
[22] http://jena.sourceforge.net/inference/

The user enters the URI of the description to be validated (A). In an advanced mode, optional parameters corresponding to an alternative formalization of the semantic constraints can be entered (B). The `Validate` button provides a meaningful explanation of the errors detected in the description (C).

4.3.2 VAMP as a Web Service for the Semantic Web

VAMP is also available as a web service so that the validation functionality can be embedded into any application. We provide a REST-style web service interface for the validation service. Similar to the graphical user interface, the client of the Web Service provides an input MPEG-7 description to be validated and can specify other predefined resources (XSLT stylesheet, ontology, rules, query) identified by URIs.

The service can then generate the results of the SPARQL query in two different formats: i) an XML format, which can be easily further processed by XSLT depending on the application's needs; ii) the RDF graph that is built up in the service containing all the instances contained from the document.

5 Related Work and Discussion

Several attempts have been made to map the MPEG-7 description tools onto an OWL ontology[23], which we present in section 5.1). We then argue why MPEG-7 and its formal representation should co-exist (section 5.2). We finally discuss the scope of our approach which goes beyond the validation of MPEG-7 descriptions (section 5.3).

5.1 Existing MPEG-7 Ontologies

Automatic mappings from the MPEG-7 XML Schema to OWL covering the whole standard have been proposed [8, 28]. The resulting ontology, however, is unable to capture the intended semantics not represented in the XML schema

[23] `http://www.w3.org/2005/Incubator/mmsem/wiki/Vocabularies`

without re-engineering work. Other attempts have manually modeled an MPEG-7 ontology. The result is, however, either restricted to the upper level elements and types of MPEG-7 [11], or adapted to a very specific use of the standard in a particular application [25]. These ontologies could be used in the VAMP service as an alternative modeling of the semantic constraints as soon as a transformation into RDF is provided. The validity will then not be checked against a particular profile.

5.2 Using MPEG-7 and its Formalization

Considering the various shortcomings of the MPEG-7 schema-based representation with respect to a formal representation of its semantics, and the existing work for obtaining a formal model, one can wonder if it is worth keeping the MPEG-7 XML-based format. We argue that both representations are useful and are suitable for different purposes.

Describing the structure of audiovisual content, such as the sequence of shots contained in a video, is fundamental for many applications. Representing a structure with the current semantic web languages is often too complex. Due to the directed graph model with unordered edges used by OWL/RDF, it is not possible to determine the order of segments in the ontology-based representation without explicitly representing it [25]. Furthermore, numerous MPEG-7 low-level descriptors are characterized for having numerical values such as vectors and matrices while encapsulating few semantics. Hence, there is little or no advantage in having a formal representation for these concepts since: i) it is inefficient for typical operations such as similarity matching, ii) it will generate too many triples that might go beyond the current scale of RDF stores (consider for example the description of visual descriptors of the key frames of several hours of video).

5.3 Generalizing the VAMP Approach

The approach presented in this paper is not limited to validating MPEG-7 documents. The basic idea of formalizing some semantic constraints of specific XML-based languages can be useful in a range of other applications. For example, VAMP could be used to validate semantically SMIL documents. In the advanced options, one would need to specify the URI of a SMIL ontology along with some associated logical rules capturing the intended semantics of this standard, and then provide the XSLT transformation. The SMIL document could then be checked with VAMP, even though the human-readable explanation of the various error types would need to be adapted.

6 Conclusion and Future Work

In this paper, we proposed a general approach to overcome the interoperability problems that result from the lack of formal semantics of the MPEG-7 description tools by formalizing their semantic constraints. The approach is based on the definition of profiles, which are not just subsets of the MPEG-7 standard, but that also define a set of semantic constraints that specify the use of the descriptors in a particular context. Our methodology advocates the specification of an ontology that includes the concepts being described in a profile, plus additional logical rules to fully capture the semantic constraints. We have demonstrated the feasibility of this approach by implementing VAMP, which is available both as a web application and as a web service. We have collected and analyzed numerous MPEG-7 descriptions from various tools from the multimedia community, and we have successfully applied VAMP for checking the constraints related to time ranges in temporal decompositions and to media information, highlighting the errors produced sometimes by these tools. The validation service is also now available for checking the semantics conformance of the MPEG-7 format used for representing shot boundary references, which would be really useful for the TRECVID community when exchanging results.

When formalizing semantic constraints, the question of strictness consistency arises. There is, of course, always a tradeoff between flexibility and strictness with respect to description tool semantics. If we require the semantic constraints to be very strict, this might prevent the use of any structures in the description not foreseen in the profile definition, even if they are used as an extension and do not interfere with the structures defined in the profile. Thus it could be an option to introduce different levels of conformance to profile semantics. We are working on this concept that we name "semantic levels", by analogy to the levels of profiles in MPEG standards allowing different complexity. The idea is to define several levels of strictness in terms of semantic constraints for each profile which can then be used depending on application requirements. The definition starts with the most "liberal" semantic level: an ontology and a set of rules modeling the most basic semantic constraints of the profile. These constraints should only solve interoperability problems by avoiding ambiguities, but not unnecessarily restrict the use of optional elements or extensions. Based on this simple definition, stricter levels can be derived by adding further constraints to the ontology and defining additional rules.

Representing formally the semantic constraints of the MPEG-7 description tools is not only useful for semantically validating the descriptions, but also for establishing mappings between profiles and heterogeneous MPEG-7 descriptions. Actually, the greatest potential with semantic definitions of MPEG-7 profiles is in the ability to use these descriptions to relate the content to other audiovisual segments described using alternative MPEG-7 profiles or other domain ontologies such as EXIF or the ID3 tags. Current multimedia applications on the web need to index multimedia metadata from heterogeneous sources. Formalizing the semantics of the profiles used for representing this metadata allows to express mappings between heterogeneous descriptions based on their semantics. In the future, we plan to investigate further how the approach presented in this paper can be used in this particular use case.

Acknowledgments

The authors would like to thank Alia Amin (CWI) for the design of the VAMP interface, Philip Hofmair and Rudolf Schlatte (JRS) for their help in the implementation of VAMP, and Lynda Hardman (CWI) for her feedback on earlier versions of this paper. The research leading to this paper was partially supported by the European Commission under the contracts FP6-027026, "Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space", IST-2-511316, "IP-RACINE: Integrated Project - Research Area CINE" and FP6-027122, "SALERO: Semantic AudiovisuaL Entertainment Reusable Objects".

References

[1] Richard Arndt, Raphaël Troncy, Steffen Staab, and Lynda Hardman. Adding Formal Semantics to MPEG7: Designing a Well-Founded Multimedia Ontology for the Web. Technical Report KU-N0407, University of Koblenz-Landau, 2007.

[2] Thanos Athanasiadis, Vassilis Tzouvaras, Kosmas Petridis, Frederic Precioso, Yannis Avrithis, and Yiannis Kompatsiaris. Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. In *5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot'05)*, Galway, Ireland, 2005.

[3] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[4] Werner Bailer and Peter Schallauer. The Detailed Audiovisual Profile: Enabling Interoperability between MPEG-7 based Systems. In *12th In-*

ternational MultiMedia Modelling Conference (MMM'06), pages 217–224, Beijing, China, 2006.

[5] Chitta Baral and Michael Gelfond. Logic Programming and Knowledge Representation. *Journal of Logic-Programming*, 19–20:73–148, 1994.

[6] Marc Caillet, Jean Carrive, Cécile Roisin, and François Yvon. Engineering Multimedia Applications on the basis of Multi-Structured Descriptions of Audiovisual Documents. In *International Workshop On Semantically Aware Document Processing And Indexing (SADPI'07)*, Montpellier, France, 2007.

[7] Mike Dean and Guus Schreiber. OWL Web Ontology Language: Reference. W3C Recommendation, 10 February 2004. `http://www.w3.org/TR/owl-ref/`.

[8] Roberto Garcia and Oscar Celma. Semantic Integration and Retrieval of Multimedia Metadata. In *5^{th} International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot'05)*, Galway, Ireland, 2005.

[9] Jerry R. Hobbs and Feng Pan. Time Ontology in OWL. W3C Working Draft, 27 September 2006. `http://www.w3.org/TR/owl-time/`.

[10] Ian Horrocks, Peter F. Patel-Schneider, Sean Bechhofer, and Dmitry Tsarkov. OWL rules: A proposal and prototype implementation. *Journal of Web Semantics*, 3(1):23–40, 2005.

[11] Jane Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *First International Semantic Web Working Symposium (SWWS'01)*, Stanford, California, USA, 2001.

[12] Jane Hunter and Carl Lagoze. Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles. In *10^{th} International World Wide Web Conference (WWW'01)*, pages 457–466, Hong Kong, 2001.

[13] International Organization for Standardization. Representations of dates and times, second edition. ISO 8601, 15 December 2000.

[14] Frank Manola and Eric Miller. RDF (Ressource Description Framework) Primer. W3C Recommendation, 10 February 2004. `http://www.w3.org/TR/rdf-primer/`.

[15] MPEG-7. Multimedia Content Description Interface. ISO/IEC 15938, 2001.

[16] MPEG Requirements Group. MPEG-7 Interoperability, Conformance Testing and Profiling, v.2. ISO/IEC JTC1/SC29/WG11 N4039. Singapore, March 2001.

[17] MPEG Requirements Group. MPEG-7 Profiles and Levels under Consideration. ISO/IEC JTC1/SC29/WG11 N6039. Gold Coast, October 2003.

[18] MPF. Metadata production framework specifications (v. 1.0.1E). Technical report, NHK Science and Technical Research Laboratories, October 2006. `http://www.nhk.or.jp/strl/mpf/english/index.htm`.

[19] Frank Nack, Jacco van Ossenbruggen, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). *IEEE Multimedia*, 12(1), 2005.

[20] Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4), 2004.

[21] Peter F. Patel-Schneider, Patrick Hayes, and Ian Horrocks. OWL Web Ontology Language: Semantics and Abstract Syntax. W3C Recommendation, 10 February 2004. `http://www.w3.org/TR/owl-semantics/`.

[22] Fernando Pereira. MPEG-7 Requirements Document v.16. ISO/IEC JTC1/SC29/WG11/N4510. Pattaya, Thailand, December 2001.

[23] Silvia Pfeiffer and Uma Srinivasan. TV Anytime as an application scenario for MPEG-7. In *Workshop on Standards, Interoperability and Practice*, Los Angeles, California, USA, 2000.

[24] Harald Stiegler. Analysis framework module developer's guide (v.1.1). Technical report, JOANNEUM RESEARCH, Institute of Information Systems and Information Management, March 2007. `http://www.joanneum.at/uploads/tx_publicationlibrary/AnalysisFramewor%k_ModuleDeveloperGuide.pdf`.

[25] Raphaël Troncy. Integrating Structure and Semantics into Audio-visual Documents. In *2nd International Semantic Web Conference (ISWC'03)*, pages 566–581, Sanibel Island, Florida, USA, 2003.

[26] Raphaël Troncy, Werner Bailer, Michael Hausenblas, Philip Hofmair, and Rudolf Schlatte. Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. In *1st International Conference on Semantics And digital Media Technology (SAMT'06)*, pages 41–55, Athens, Greece, 2006.

[27] Raphaël Troncy, Jean Carrive, Steffen Lalande, and Jean-Philippe Poli. A Motivating Scenario for Designing an Extensible Audio-Visual Description Language. In *The International Workshop on Multidisciplinary Image, Video, and Audio Retrieval and Mining (CoRIMedia)*, Sherbrooke, Canada, 2004.

[28] Chrisa Tsinaraki, Panagiotis Polydoros, and Stavros Christodoulakis. Interoperability support for Ontology-based Video Retrieval Applications. In *3rd International Conference on Image and Video Retrieval (CIVR'04)*, Dublin, Ireland, 2004.

[29] XML Schema. W3C Recommendation, 2 May 2001. `http://www.w3.org/XML/Schema`.