Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

INS

Information Systems

INformation Systems

Towards a syntax for multimedia semantics

F.-M. Nack, H.L. Hardman

REPORT INS-R0204 APRIL 30, 2002

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization 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

Towards a Syntax for Multimedia Semantics

Frank Nack and Lynda Hardman
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

ABSTRACT
This article describes the current state of the art on representing the four essential conceptual facets of a multimedia unit, namely the form and substance of content and the form and substance of its expression, and points to the still unsolved problems regarding the syntax for media semantics. We first provide a brief overview of the general features of the MPEG-7 standard and its different parts. This serves as a description of the state of the art in content description for audio-visual media. We then analyse the ability of one of these parts for its capability to define structures for describing media semantics. We describe the problems of two currently conflicting MPEG-7 representations of expression-based media semantic, which should be equivalent. We then discuss high-level aspects of media semantics, namely the general problems of an ontology for media semantics. Finally, we talk about the problems of applying the theoretical concepts to real applications.

1998 ACM Computing Classification System: H.5.2, I.3.4, I.3.8
Keywords and Phrases: Multimedia semantics, MPEG-7, DDL, MDS, Semantic Web, XML-Schema, information spaces.
Note: The work was carried out under the Ontoweb, a thematic network of the European Commission.

1. INTRODUCTION

The growing amount of digital audiovisual information and the need to transform it into applicable knowledge requires a machine-processable representation of its associated semantics.
The semantics of a media unit depends on the context in which it is used, where the meaning of context here is twofold. On the one hand it represents the use of the material in the current application and on the other hand it represents the overall placement of the information in the domain the application is applied to.
The semantics of a media unit are organised around surface structures (*expression*) and deep structures (*content*). Expression is the means used to represent the media itself, whereas content is the representation of the conceptual items, which are expressed through the media. Both expression and content, however, depend on *substance* and *form*, where substance represents the natural material for content and expression, but form represents the abstract structure of relationships, which a particular media demands.
Examples of the four aspects in a sports video are: the substance of expression is the mpeg format of the video, the form of expression are the cuts, the substance of content is represented by the ball, and the form of content is given by the structure of the event sequences, in sports usually the sequence of highlights. Figure 1 shows the relationships between the differing structures found in multimedia unit in more detail.

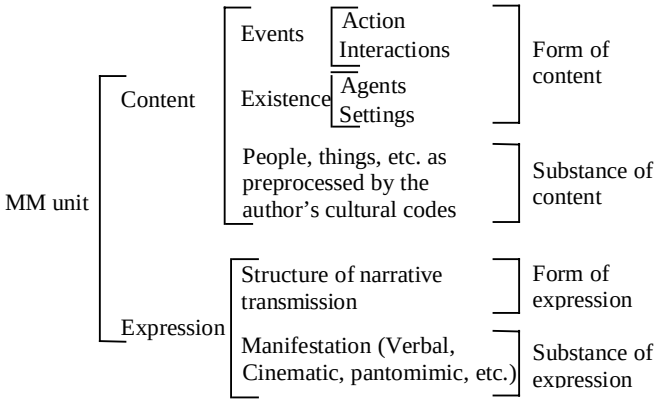

Figure 1: Relationship between multimedia elements (adopted from Chatman (1978, p.26))
In this article we are mainly interested in the advances on and remaining problems of representing these four essential conceptual aspects of a multimedia unit. Though the modelling of these four key aspects provides the means for the

communication of information, we additionally look at the language for expressing the representational structures (schemata) and their organisation in the form of related documents.
The discussion is predominantly motivated by MPEG-7, and where appropriate, we compare MPEG-7 to other approaches, in particular - the most widely known alternative towards machine-processable content description – the Semantic Web activity of the W3C [W3C Semantic Web 2002].
The paper is structured as follows. We first provide a brief overview of the general features of the MPEG-7 standard and its different parts serving as a description of the state of the art in content description for audio-visual media. We then analyse the ability of the description definition language (DDL) on its ability to define structures for describing media semantics. In section 4 we describe the problems of conflicting representations of expression-based media semantic, which should be equivalent but are not. Section 5 is devoted to high-level aspects of media semantics, namely the problems of an ontology for media semantics. Section 6 finally discusses the problems of addressing the theoretical concepts to real world problems. The paper concludes with an overview of research questions.

2. DESCRIPTION OF AUDIOVISUAL MEDIA

The goal of the Multimedia Content Description Interface [ISO MPEG-7 2001a-f] is to provide a standardised means of describing audiovisual data content in multimedia environments. Its scope is to facilitate the description of content of multimedia data, so that this data can be searched for, browsed, filtered or interpreted either by search engines, filter agents, or any other program.
MPEG-7 offers a set of audiovisual description tools in the form of descriptors (Ds) and description schemata (DS) describing the structure of the metadata elements, their relationships and the constraints a valid MPEG-7 description should adhere to. These structures[1] form the basis for users to create application specific content descriptions, i.e. a set of instantiated description schemata and their corresponding descriptors. The standard is organised in 8 parts, each responsible for a particular aspect of the functionality:

Systems specifies the tools for preparing descriptions for efficient transport and storage, compressing descriptions, and allowing synchronization between content and description. It is important to mention that MPEG-7 descriptions may be delivered independently of, or together with the content they describe [ISO MPEG-7 2001a].

The Description Definition Language (DDL) specifies the language for defining the standard set of description tools (Description schemata (DS), descriptors (Ds), and datatypes) and for defining new description tools. The main parser requirements are defined here. [ISO MPEG-7 2001b]. Note that additional essential datatypes are defined in the parts Audio, Video and, in particular, the MDS.

Visual consists of structures and descriptors that cover basic visual features, such as colour, texture, shape, motion, localization, and face recognition. The syntax of the descriptors and description schemata is provided in normative DDL specifications and the corresponding binary representations. Moreover, normative definitions of the semantics of all the components of the corresponding descriptors and description schemata are provided [ISO MPEG-7 2001c].

Audio specifies a set of low-level descriptors for audio features (e.g., spectral, parametric, and temporal features of a signal), and high-level description tools that are more specific to a set of applications. Those high-level tools include general sound recognition and indexing schemata, such as for instrumental timbre, spoken content, audio signature and melody. Moreover, normative definitions of the semantics of all the components of the corresponding descriptors and description schemata are provided [ISO MPEG-7 2001d].

The part *Multimedia Description Schemes (MDS)* specifies the generic description tools pertaining to multimedia including audio and visual content. The MDS covers

- the basic elements for building a description (this section also defines additional datatypes used in the visual and audio part, which are not covered by the DDL datatype definitions),
- the tools to describe content and relate the description to the data and
- the tools to describe content on an organisation, navigation and interaction level [ISO MPEG-7 2001e].

Reference Software provides reference software to the standard [ISO MPEG-7 2001f].

Conformance specifies the guidelines and procedures for testing conformance of implementations of the standard [ISO MPEG-7 2001f].

Extraction and use specifies provides guidelines and examples for the extraction and use of descriptions. [ISO MPEG-7 2001f].

[1] DS and Ds are usually described by the standard as descriptive tools, whereas the parser, for example, represents technology. We use the terms in this article accordingly.

Clearly, the standard addresses a broad spectrum of representational problems, from high-level conceptual descriptions of the content itself and its production down to the smallest detail on a low-level feature level. The rest of the article is devoted to the question of which problems regarding the representation of multimedia semantics are still unsolved.

3. THE LANGUAGE PROBLEM

In this section we discuss the requirements for a language that facilitates the syntactic means for establishing semantic descriptions of multimedia. These requirements form the basis of the examination of the MPEG-7 approach, where the fusion of language syntax and schemata semantics establishes a fundamental problem for the applicability of the provided structures and technology.

There are a number of requirements for a language that facilitates the description of multimedia content. The language must be able to

- support the definition of syntactic rules to express and combine description structures at various levels of detail, which results in the provision of a rich set of syntactic, structural, cardinality and datatyping constraints.
- state spatial, temporal and conceptual relationships between the components of a description and between descriptions, so that a meaningful discourse about, or with, descriptions, through algebraic, logic, or functional means, is possible.
- facilitate a diverse set of linking mechanisms between descriptions and the data that is described, which includes, in particular, means of segmentation for temporal media.
- be platform and application independent and human- and machine-readable.

Within MPEG-7, the DDL is intended to address the above requirements. The language provides basically the same structure-oriented language elements as XML-Schema [W3C XML Schema 2002][2]. The only extensions to XML-Schema cover the ability to define arrays and matrices and to provide two additional datatypes that is basicTimePoint and basicDuration, which allow specific temporal descriptions [see ISO MPEG-7 2001b, pp. 9 – 14]. Any available MPEG-7 parser addresses consequently only these extensions in addition to the other language constructs.

The XML orientation of the suggested language facilitates platform and application independence and human- and machine-readability. However, as it merely adopts the syntactic elements to represent structures in the form of schemata, the language

- lacks particular media-based datatypes,
- it does not facilitate mechanisms to support semantic descriptions, such as the relation-oriented schemata language RDF(S) [W3C RDF 2002] or ontology-based technology such as DAML+OIL [DAML+OIL 2002],
- it does not provide particular linking mechanisms as for example provided by XLink [W3C Xlink 2002] and Xpath [W3C Xpath 2002].

In fact, the DDL offers fewer media specific tools than other multimedia-based description approaches in the form of presentational languages such as SMIL (WWW-based media animation and integration of style) [W3C SMIL 2002], SVG (with CSS for graphics) [W3C SVG 2002] and XHTML (with CSS for formatted text) [W3C XHTML 2002], or transformation methods such as XSLT (document transformation) [W3C XSLT 2002] and CSS (control of style appearance) [W3C CSS 2002]. Note that these W3C approaches are unable to recognize visual media's dynamic nature or its variety of data representations and their mixes.

The strength of the DDL, namely supporting the definition and alteration of schemata, is used in MPEG-7 to define normative schemata that address the language requirements outlined earlier. Mainly in the MDS we find a plethora of structures for

- specific datatypes required for the description of form and substance of media expression [ISO MPEG-7 2001e, pp. 49-103]. Extensions are provided in the parts Visual [ISO MPEG-7 2001c] and Audio [ISO MPEG-7 2001d];
- linking, identification and localisation tools, mainly based on XML-Path but extended with particular temporal constructs, that provide a basic means of establishing references within a description and linking to the associated multimedia-data [ISO MPEG-7 2001e, pp. 74-103];
- graphs of relations, where the basic unit of a relation is built, similar to RDF, on a conceptual triple that allows the establishment of named relations between the parts in a description. The organisation of relations is restricted to a defined set of 11 topological and set-theoretic graph-relation types [ISO MPEG-7 2001e, pp. 179 – 191].;

[2] Earlier work describes the discussion about the orientation of the language (procedural or descriptive) and the final decision process in more detail [Nack & Lindsay 1999, Hunter & Nack 2000].

- forms of spatio, temporal and spatio-temporal segmentations for video, audio, audio-visual, multimedia, and ink content, including a set of temporal and spatial relations [ISO MPEG-7 2001e, pp. 251 – 400 and 458 - 540];
- a set of 45 semantic relations that allow the description of narrative structures [ISO MPEG-7 2001e, pp. 401 - 457].

All of these constructs and mechanisms are valuable for describing the semantics of a single multimedia object or collections in the form of a multimedia unit – but they are not part of the description language[3].
The consequences of merging the description language (syntax) with the schemata for providing description structures (semantics) are far reaching. As the essential semantic aspects for the description of multimedia are defined in standardised schemata they have to be used in the provided way and any alteration, including the combination of schemata, will be outside the scope of the standard. More crucially, any alteration on one of the "language related" schemata will not only alter the semantics of the description but also the description language itself. Such alterations are, however, unavoidable since as a great number of schemata describe solutions for particular problems for a fraction of multimedia applications, as demonstrated by the tools for video segmentation in Figure 2 on the previous page, alterations are unavoidable.

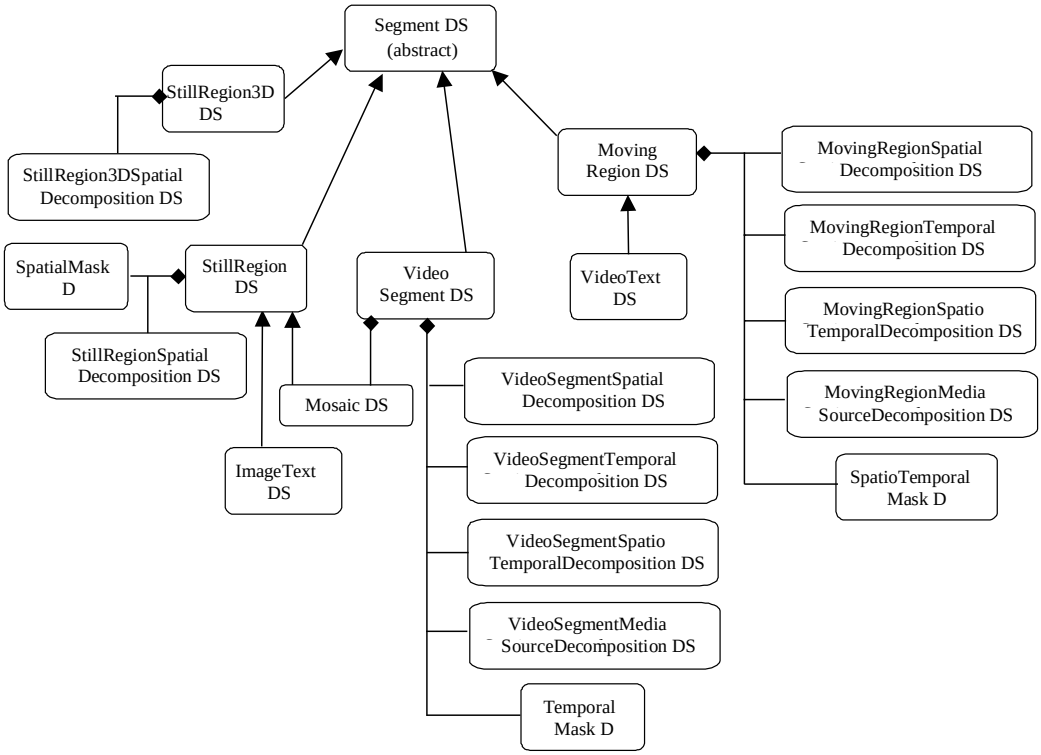

Figure 2: Overview of the tools for describing segments of visual content in MPEG-7 [ISO MPEG-7 2001e, p. 285].

Moreover, dispersing language elements into description schemata asks for an evaluation complexity close to a validation level no parser can cope with. In fact, at the time of writing there is no MPEG-7 parser that could possibly handle the existing structures.
The problem of defining a language that provides the syntax for describing multimedia semantics is still not solved but we are a great deal closer to it. The next goal is to identify semantically relevant syntax elements in the semantic-related schemata and abstract them in a way that they can extend the DDL. In other words, transform the DDL from a mere structure-oriented language into a rigorous media-based description language.
The syntactic properties of this language would facilitate re-use and inferencing about material for specific purposes, potentially leading to interoperability with other description formats, such as RDF(S)[4] or DAML+OIL. This would increase the potential for interoperability and integration of MPEG-7 semantics with metadata descriptions from other domains, such

[3] In fact the philosophy behind XML-Schema is more practical because it simply provides a means for establishing schemata without caring about their potential semantics. The DDL + Semantic-Schemata are rigorous about use and implied semantics. In fact, MPEG-7 reads in large parts as an ontology, a topic to be discussed later.

[4] There are attempts to combine RDF(S) and MPEG-7 [Hunter & Lagoze 2001, Hunter 2001].

as the Open Media Framework [OMF 2001], the Multimedia Home Platform (MHP) (as part of the Digital Video Broadcasting (DVB) Project [DVB 2001]), the TV Anytime Forum [TV Anytime 2001], the Dublin Core Metadata Initiative [Dublin Core 2001], MPEG-21 [ISO-MPEG-21 2001], NewsML [NewsML 2001], the Gateway to Educational Materials project [GEM 2001] or the Art & Architecture Thesaurus browser [AAT 2001].

4. MEDIA EXPRESSION – THE REPRESENTATION PROBLEM

In this section we consider the fundamental problems of representing the semantics of media expression.. Though we pay attention to the representation of expression form, we mainly focus on the representation of the dynamic aspects of expression substance (low-level features), where the MPEG-7 approach suggests the provision of two formats, i.e. a binary (procedural) and textual (conceptual) representation. Though the need for both formats is undeniable, it requires structural modelling that is at the current stage not fully achieved.

As stated in the introduction, it is the expressional form and substance of multimedia objects that require special attention when it comes to semantics. The process of visual and audio signification, though based on common human knowledge and thematic structures (expression form), provides its own temporal-spatial realities based on patterns of juxtaposition of the media intrinsic parts (expression substance). The information provided on a perceptual level using objective measurements, such as those based on image or audio processing or pattern recognition, play an important role regarding the aesthetics of a multimedia unit and consequently its subjective interpretation.

Support for the form of expression requires a rich set of presentation models. In MPEG-7 a set of schemata is suggested that provides structures for multimedia summaries, points of view, partitions and variations [ISO MPEG-7 2001e, pp. 458 - 540] and various forms of collections on a probabilistic, analytical or classification level [ISO MPEG-7 2001e, pp. 541 - 600]. These schemata are very detailed, but they impose particular semantics on the user. In fact, the approach taken by SMIL [W3C SMIL 2002][5] representing a textual serialisation of temporal and spatial aspects for multimedia presentations seems more promising because it is less rigid and thus more easily applicable.

The approach taken by MPEG-7 for representing substance of expression, i.e., the semantics low-level audio and visual features, addresses an important problem, but it is problematic in its solution.

The two parts devoted to the description of features are Visual and [ISO MPEG-7 2001c] and Audio [ISO MPEG-7 2001d]. The visual part is devoted to the description of features such as colour, texture, shape, motion, or localisation, whereas the audio part focuses on various series types (scalable, scalar, vector etc.), waveform, power, spectrum, harmonicity, silence, sound, spoken content, etc. Both parts address the problem of representing the dynamic nature of audiovisual semantics and try to solve it by providing a binary (algorithmic) and textural (schema) description structure. Both intend to provide the same information, since a requirement for the system specification of MPEG-7 is that “MPEG-7 data can be represented either in textual format, in binary format or a mixture of the two formats, depending on application usage. A bi-directional loss-less mapping between the textual and the binary representation is possible.” [ISO MPEG-7 2002a, p. 10]. This, however, turns out not to be the case.

Consider the following example of the descriptor “ColourStructure” specifying both colour content (similar to that of a colour histogram) and the structure of this content as described in Part 3: Visuals [ISO MPEG-7 2001c, pp. 50 – 56]. Figure 3 illustrates the textual representation, whereas Table 1 describes the essential part of the binary representation.

Despite the fact that the binary syntax is accompanied by long textual and graphical descriptions giving detailed semantic definitions covering the extraction algorithm, the re-quantization, the colour space and colour quantization (see table 2 as an example) and the raw "ColourStructure" histogram accumulation method, there is obviously a strong discrepancy between the supposedly equivalent textual and binary representation forms.

[5] SMIL is also among the formats the Extensible MPEG-4 Textual Format (XMT) is addressing. XMT is MPEG4’s textual representation of the Binary Format for Scenes (BIFS) [ISO MPEG-4 2001].

```
<complexType name="ColorStructureType" final="#all">
 <complexContent>
  <extension base="mpeg7:VisualDType">[6]
   <sequence>
    <element name="Values">
     <simpleType>
      <restriction>
       <simpleType>
        <list itemType="mpeg7:unsigned8"/>
       </simpleType>
       <minLength value="1"/>
       <maxLength value="256"/>
      </restriction>
     </simpleType>
    </element>
   </sequence>
  <attribute name="colorQuant" type="mpeg7:unsigned3"
   use="required"/>
  </extension>
 </complexContent>
</complexType>
```

Figure 3: Textual representation syntax for the ColourStructure

ColorStructure{	Number of bits	Mnemonic
colorQuant	3	uimsbf
NumOfValuesCode	8	uimsbf
for (m=0; m<M; m++) {		
Values[m]	8	uimsbf
}		
}		

Table 1: Binary representation syntax

colorQuant	operating point
0	Forbidden
1	32 (HMMD)
2	64 (HMMD)
3	128 (HMMD)
4	256 (HMMD)
5-7	Reserved

Table 2: Semantics of colorQuant.

For example, none of the semantic descriptions relevant for the interpretation of the schema and provided in the standard made it into the textual description. In fact, the schema merely provides the structure of the result space, that is the size of the matrix that contains the results of the extraction algorithm. Thus, the general assumption that textual and binary representations are interchangeable is false. At the moment, it is not clear how the two representational forms can be merged and perhaps it is not even possible. If that is the case, then it is at least necessary to improve the textual representation, which is essential because a parser is not able to evaluate the binary representation. While this problem may appear trivial, it has far reaching consequences because the use of low-level features for semantic-based descriptions is one of the few mechanisms available for the automatic annotation of media.

[6] The types VisualDType, unsigned8 and unsigned3 are defined in the MDS and not in the DDL.

5. SEMANTIC MAPPING – THE ONTOLOGY PROBLEM

The representational problem of media expression also throws light on another problematic issue, i.e. the mapping of semantics or the ontology problem. In knowledge representation it is well known that the description of a domain provides particular semantic concepts, which are hard, if not impossible, to be mapped onto ontologies that cover the same or a similar semantic space. This is because of the differences in the use of language or conceptual structure. In this section we first show that the ontological commitment in MPEG-7 is implicit and describe the resulting representational problems. Then we show that the same dilemma also applies to the semantics of low-level features. This unexpected problem is made in particular explicit in the Audio and Video parts.

A large number of schemata in the standard establish ontological structures, since semantics are always purpose driven. That means that there are no description structures that are domain or context independent. In MPEG-7 most schemata are inspired by the domain of broadcasting and audiovisual-based entertainment (see for example the VideoEditingSegment, the AgentDS, PlaceDS, or the user preference description schemata in the MDS). The large number of schemata, often describing similar aspects of the same semantic problem, and their interlocked nature, indicate the ontological role at least of the MDS. However, the attempt of abstraction to achieve domain independence makes it impossible to use those schemata as ontology items. The goal of abstraction is also mainly responsible for the syntax and semantic problem, as described in section 3,

However, there is one attempt in MPEG-7 to address the language problem within ontologies, i.e. the classification schema. The classification schema facilitates the organisational wrapper for a controlled vocabulary built out of terms and the relations between them. The relations organise the terms in the form of a hierarchy, indicating if one term is broader or narrower in its meaning than another, a synonym or, in the given set of relations, the one of highest relevance. Thus, a classification schema in some sense covers aspects of a thesaurus. The classification schema allows the incorporation of other classification schema, though no indication is given, if this feature only takes account of the inclusion of other MPEG-7 classification schemata or also the insertion of or connection to other ontologies. Unfortunately, there is also no information provided about how the mapping from previously unconnected terms should be achieved.

Thus, the problem of mapping high-level media semantics is not solved yet and it remains questionable if the MPEG-7 approach of profiling schemata provides suitable solutions.

The representation problem for the substance of media expression, as described in section 4, also points to the ontology problem. It is one achievement of MPEG-7, in particular the parts Audio and Visual, that it made the problem of semantic mapping for the level of expressional semantics explicit, though without acknowledging it.

The optical and audible patterns in audiovisual media communicate on the basis of precise perception. However, this low-level data needs interpretation. For example, an image shown in isolation is a form of utterance that provides an identifiable semantics. The same image presented in a sequence might appear with a modulated semantics because the order created new levels of meaning. The same effect appears in sequences of image sequences, i.e. video. Thus, even on the level of low-level feature description, we have to provide collections of objective measurements for media units representing prototypical style elements, and the context establishing relationships between them. In other words, we create descriptions of expressional detail based on their own ontology, as described in Nack et al. [2001], Schreiber et al. [2001], and Dorai & Venkatesh [2001]. Each of these descriptions varies on the level of structural depth and the use of feature descriptors (use of different features based on different extraction algorithms), though the approaches describe the same or similar semantics.

Thus, the ability to isolate various semantic representations of visual and audible artefacts leads towards combinatoric, where the resulting forms of non-equivalent structure require mapping mechanisms, to allow material to be searched for, browsed, filtered or interpreted either by search engines, filter agents, or any other program, which will not be able to know about all potential representational forms.

The only potential way of solving the problem of “semantic mapping” of expression is the provision of a suitable language for expressing the syntax of multimedia semantics, as outlined in section 3, which in turn will alleviate the problems of representing low-level semantics of media expression, as described in section 4.

6. THE APPLICABILITY PROBLEM

The ontology problem of media expression points to another important difficulty, namely the applicability of schemata to a domain. This covers not only aspects such as expressiveness of provided structures for current domain semantics and extensibility to emerging semantic needs, but also addresses the problem of generation and maintenance of media-based knowledge spaces. In this section we state the requirements for organisational structures that facilitate the handling of media-based knowledge spaces. These requirements form the basis of the examination of the MPEG-7 approach, where a hierarchical structure on a single document basis establishes a fundamental problem for the applicability of the provided structures and technology.

There are a number of requirements for dynamic knowledge spaces:

- The nature of annotations is necessarily imperfect, incomplete, and preliminary because they accompany and document the dynamic progress of understanding a concept. Thus, dynamic semantic structures that adapt to changing contexts to facilitate discourse are required.
- There is no such thing as a single and all-inclusive content description. Thus, there is a need for mechanisms to establish collective sets of descriptions growing over time (i.e. no annotation will be overwritten but extensions or new descriptions will appear in the form of new documents).
- There is evidence that a great deal of useful annotation can be provided by human effort [Nack and Putz, 2001]. Hence, it is necessary that the generation and maintenance of media semantics incorporate the annotation in the form that structures can be read and altered by humans.

Devoted to the goal of efficient access to and retrieval of audiovisual material, MPEG-7 suggests, as a potential solution to the first two of the above requirements, a hierarchical structure with a flexible root element that determines the depth and structure of the tree. This decision has far-reaching consequences.

The fundamental problem of this approach is that a description of a media item is basically forced into one document. The critical part is the root element, which is designed on the vision of a "universal" description schema for a domain. Its instantiations can then be attached to the relevant media items. Naturally, the resulting descriptions are consistent and interoperable, even if the descriptions vary in their instantiated depth. However, this approach lacks the required flexibility of changing semantic needs. Though the structure of the schema can be complex, once it is created and used in instantiations, its structure cannot be altered. Any modification would cause inconsistencies with existing documents. However, the new semantic needs could be expressed in a new schema and its instantiations could then refer to the existing documents[7]. Since MPEG-7 allows linking between descriptions, this is a potential solution, though only partially because the link is necessarily uni-directional. Links in MPEG-7, however, do not provide any information about the semantics of the relationship between documents. Relations, which supply the desired semantics, and are provided by MPEG-7 through the introduction of "relationship element", can only be applied within a document. This still results in encapsulating the required network structure in a single document.

A potentially improved solution would be to fully deconstruct the single document approach. The goal is to allow the proposed graph structures in the standard not to function as document intrinsic solutions but rather apply them to an open semantic-oriented network structure. Various levels of clustering allow structure without losing flexibility and the potential for temporal growth.

The complicated construct of distinguishing between a complete and fragmental description expresses one instance of an overall level of complexity, which makes it difficult to apply the provided schemata. Other instances of the complexity problem are:

- There are a great number of abstract elements, which are used to establish class structure[8]. However, abstract elements cannot appear in instantiations. When an element is declared to be abstract, a member of that element's substitutable class must appear in the instance document. To indicate that the derived type is not abstract, the XML namespace mechanism is used (xsi:type). Thus, a thorough understanding of schemata development is required which makes instant schemata development for distinct domains hard, especially if the required schemata should cover simple descriptions, where the theoretically founded overhead is actually not required.
- The interlocked nature of schemata, resulting on the approach of providing an ontology-like but yet general set of schemata to describe media semantics, makes it very difficult for a user to identify the appropriate schemata and to use them in isolation.
- Due to the lack of a fundamental data model (the language problem as described in section 3) the provided structures show inconsistencies and duplications, which makes manual schemata generation difficult.

The structural complexity of MPEG-7 is obstructing the take-up of the standard, especially since few support tools exist for the manual generation of new schemata.

The TV Anytime Forum [TV Anytime 2001] perhaps gives an indication of how semantic structures provided by MPEG-7 will be used in the future. The TV Anytime Forum develops specifications for services based on consumer digital storage devices. The semantic structures, all written in XML Schema, are self-developments and cover the essential aspects of media description, i.e. content description, content referencing and location, rights management and protection, systems and transport.. Though the TV Anytime schemata are similar to the equivalent structures in MPEG-7, they are less complex in

7 The "description unit" might be intended to play that role. The problem with this construct is that it is deficient in most of the conceptual overhead of the "complete description", among which the lack of linking mechanisms is the most serious. In fact, a "description unit" performs merely as a free-floating description unrelated to real data.

8 The fundamental problem of class and instance, where sometimes an instance should also be a class, is implicitly addressed in MPEG-7 and also forms part of the language problem described in section 3.

their organisational structure. TV Anytime includes the MPEG-7 schemata on user-modelling, though without incorporating the complete MPEG-7 organisational overhead. Rather, TV Anytime uses MPEG-7 as a namespace and thus be able to incorporate just the required schemata [TV Anytime 2001a].

7. CONCLUSION

In this article we discussed the advances on representing the four essential conceptual facets of a multimedia unit, namely the form and substance of content and the form and substance of its expression. The discussion was predominantly motivated by MPEG-7, serving as a description of the state of the art in content description for audio-visual media. Additionally we looked at the language for expressing the representational structures (schemata) and their organisation in the form of related documents.

Though the new standard for the description of audio-visual media on a semantic level provides a number of solutions, we still face a situation where essential problems are not yet solved.

We showed that the current approach of fusing language syntax and schemata semantics is problematic and must be seen as a first step towards a language that facilitates the syntactic means for establishing semantic descriptions of multimedia. We also described the next steps in language development, i.e. to identify semantically relevant syntax elements in the semantic-related schemata and to include them into the DDL. The syntactic properties of this language would facilitate re-use and inferencing about material for specific purposes, potentially leading to interoperability with other description formats and with metadata descriptions from other domains.

We considered the fundamental problems of representing the semantics of media expression, focussing on the representation of expression substance. We showed that there is need for two formats, i.e. a binary (procedural) and textual (conceptual) representation and pointed out that this requires structural modelling that is at the current stage not fully achieved.

We also showed that the problem of semantic mapping of schemata still needs further research. Though MPEG-7 tries to achieve to be a highly interoperable standard among well-known industry standards and other related standards of different domains, it does not make explicit how the mapping between unrelated schemata can be achieved.

Finally, we pointed out a number of problems regarding the applicability of MPEG-7 of which the complexity level of the internal organisation and the compulsion into a hierarchical single document structure for a description are the most critical. We pointed out solutions, in particular how graph structures can be applied on non-equivalent documents describing the same media unit from different points of view.

The problem of describing the semantics of multimedia is complex and the results of the MPEG-7 standard provide us with a first step towards a syntax for multimedia semantics. All we can do is to build on the lessons learned.

The various problems addressed in the article suggest that semantics are domain dependent. As a result we have to accept that we have to leave the development of ontologies for low and high-level semantic concepts to the relevant industry bodies. Over time we might be able to distil a set of “generally applicable schemata” out of these domain semantics. These universal sets of standardised schemata can then be used, as demonstrated by TV Anytime’s inclusion of MPEG-7 schemata.

However, to allow interoperability between various descriptive representations within one domain or crossover domains, we have to provide a language that provides the syntax for describing multimedia semantics. Hence the goal is not to provide a description language that mainly allows the alteration of schemata – which is highly problematic for the applicability of a standard. The aim is rather to develop a powerful description definition language, as outlined in section 3. This problem is not yet solved but the syntactical flexibility of this language will allow the creation of representations that facilitate machine-processable or manual investigation based on interpretive or associative methods.

8. ACKNOWLEDGMENTS

The author wishes to thank in particular Wolfgang Putz from FHG-IPSI in Darmstadt and Jane Hunter from DSTC in Brisbane for insightful discussion and helpful comments. We also wish to thank our colleagues Jacco van Ossenbruggen and Lloyd Rutledge for useful discussion during the development of this work This work was funded under Ontoweb, a thematic network of the European commission.

REFERENCES

1. AAT (2001). http://www.getty.edu/research/tools/vocabulary/aat/
2. Chatman, S. (1978). Story and Discourse: Narrative Structure in Fiction and Film. New York: Ithaca.
3. DAML+OIL (2001). http://www.ontoknowledge.org/oil/
4. Dorai, C. & Venkatesh, S. (2001). Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics. Proceedings of the First Conference on Computational Semiotics for Games and New Media - COSIGN 2001, pp. 94 – 99, Amsterdam, 10 – 12 September, 2001.
5. Dublin Core (2001). http://dublincore.org/
6. DVB (2001). http://www.dvb.org/dvb_technology/index.html
7. GEM (2001). The Gateway to Educational Materials, http://www.thegateway.org
8. Hunter, J. (2001). Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. International Semantic Web Working Symposium (SWWS), Stanford, July 30 - August 1, 2001. http://archive.dstc.edu.au/RDU/staff/jane-hunter/swws.pdf
9. Hunter, J. & Lagoze, C. (2001). Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles. In: The Tenth International World Wide Web Conference, Hong Kong pp. 457—466, May 1-5, 2001
10. Hunter, J. & Nack, F. (2000). An Overview of the MPEG-7 Description Definition Language (DDL) Proposals. Signal Processing: Image Communication Journal, Special Issue on MPEG-7, Vol. 16, pp 271-293, 2000.
11. ISO MPEG-4 (2001). "*MPEG-4 Overview - (V.18 - Singapore Version)*". ISO/IEC JTC1/SC29/WG11 N4030, March 2001
12. ISO MPEG-7 (2001). "MPEG-7 Requirements Document V.15". ISO/IEC JTC1/SC29/WG11/"N4317, July 2001, Sydney
13. ISO MPEG-7 (2001a). "Text of ISO/IEC CD 15938-1 Information Technology - Multimedia Content Description Interface – Part 1 Systems". ISO/IEC JTC 1/SC 29/WG 11/ M7041, March 2001, Singapore
14. ISO MPEG-7 (2001b). "Text of ISO/IEC FCD 15938-2 Information Technology - Multimedia Content Description Interface - Part 2: Description Definition Language". ISO/IEC JTC 1/SC 29/WG 11 N4288, 2001-09-19
15. ISO MPEG-7 (2001c). "Text of ISO/IEC 15938-3/FDIS Information Technology - Multimedia Content Description Interface – Part 3 Visual". ISO/IEC JTC 1/SC 29/WG 11/N4358/ July 2001, Sydney
16. ISO MPEG-7 (2001d). "Text of ISO/IEC FDIS 15938-4:2001(E) Information Technology - Multimedia Content Description Interface – Part 4: Audio". ISO/IEC JTC 1/SC 29, 2001-06-09
17. ISO MPEG-7 (2001e). "Text of ISO/IEC 15938-5/FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes". ISO/IEC JTC 1/SC 29/WG 11 N4242, 2001-10-23
18. ISO MPEG-7 (2001f). "Overview of the MPEG-7 Standard (version 6.0)". ISO/IEC JTC1/SC29/WG11 N4509, Pattaya, December 2001.
19. ISO MPEG-21 (20001). MPEG-21 Multimedia Framework, ISO/IEC JTC1/SC29/WG11/N4511 Pattaya, December 2001 http://www.cselt.it/mpeg/standards/mpeg-21/mpeg-21.htm
20. Koivunen, M.-R. & Miller, E. (2001). W3C Semantic Web Activity. http://www.w3.org/2001/12/semweb-fin/w3csw
21. Nack, F. & Lindsay, A. (1999). Everything you wanted to know about MPEG-7: Part I & Part II. IEEE MultiMedia, July - September 1999,pp., 65 - 77,October – December 1999,pp.64-73, IEEE Computer Society.
22. Nack, F. & Putz, W. (2001). Designing Annotation Before It's Needed In Proceedings of the 9th ACM International Conference on Multimedia, pp. 251 - 260, Ottawa, Canada, Sept. 30 - Oct. 5, 2001
23. Nack, F., Windhouwer, M., Hardman, L., Pauwels, E., & Huijberts, M. (2001). The Role of High-level and Low-level Features in Semi-automated Retrieval and Generation of Multimedia Presentations. CWI-technical report, INS-R0108, 2001
24. NewsML (2001). http://www.newsml.org/
25. OMFI (2001). http://www.avid.com/omfi/index.html
26. Schreiber, A. T. G., Dubbeldam, B., Wielemaker, J., & Wielinga, B (2001). Ontology-based Photo Annotation, IEEE Intelligent Systems, pp 66 – 74, May/June 2001 (Vol. 16, No. 3). http://www.computer.org/intelligent/ex2001/x3066abs.htm

27. TV Anytime (2001). http://www.tv-any-time.org/
28. TV Anytime (2001a). Specification Series: S-3 On: Metadata Corrigenda 1 to S-3 V1.1, Document: COR1_ SP003v1.1, December 21, 2001
29. W3C CSS (2002). http://www.w3.org/Style/CSS/
30. W3C RDF Schema (2002). http://www.w3.org/RDF/
31. W3C Semantic Web (2002). http://www.w3.org/2001/sw/
32. W3C XSLT (2002). http://www.w3.org/Style/XSL/
33. W3C SMIL (2002). http://www.w3.org/AudioVideo/
34. W3C SVG (2002). http://www.w3.org/Graphics/SVG/Overview.htm8
35. W3C XPath (1999). http://www.w3.org/TR/1999/REC-xpath-19991116
36. W3C XHTML (2002). http://www.w3c.org/MarkUp/
37. W3C XML Schema (2002).
    Part 0: Primer http://www.w3.org/TR/xmlschema-0/
    Part 1: Structures http://www.w3.org/TR/xmlschema-1/
    Part 2 : Datatypes http://www.w3.org/TR/xmlschema-2/