Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

INS

Information Systems

INformation Systems

That Obscure Object of Desire: Multimedia Metadata on the Web (Part II)

Jacco van Ossenbruggen, Frank Nack, Lynda Hardman

REPORT INS-E0309

That Obscure Object of Desire: Multimedia Metadata on the Web (part II)

Frank Nack, Jacco van Ossenbruggen, Lynda Hardman

Abstract

Part I of this article provided our vision of a media-aware Semantic Web in the form of a business presentation scenario and derived from it a number of problems regarding the semantic content description of media units. We discussed the multimedia production chain, in particular emphasizing the role of progressive metadata production. As a result we distilled a set of media-based metadata production requirements and showed how current media production environments fail to address these. We then introduced relevant W3C Recommendations and ISO standards. In part II we analyze the abilities of the W3C and ISO for for defining structures for describing media semantics. We discuss syntactic and semantic problems, ontological issues for media semantics and the problems of applying the theoretical concepts to real world applications. We conclude with implications of the findings for future action with respect to the activities the community should take.

Keywords: Semantic Web, metadata production, multimedia production process, XML, XML Schema, RDF, RDF Schema, MPEG-4, MPEG-7, MPEG-21

1 Introduction

Machine-processable content is the main prerequisite for the more intelligent Web services that constitute the "Semantic Web". To be able to build tools that are aware of the semantics of both the content and the context of multimedia, we need a language that makes the semantics of media units explicit. In part I we discussed, based on our analysis of the metadata production process, the requirements for a language that facilitates the description of multimedia content. The requirements for our desired multimedia metadata format can be summarized as follows. It should:

1. be platform and application independent and human- and machine-readable.
2. support a definition language for media content description structures at various levels of detail, including a rich set of syntactic, structural, cardinality and multimedia datatyping constraints.
3. support the definition of the various spatial, temporal and conceptual relationships between the media items in a commonly agreed upon format,
4. facilitate a diverse set of linking mechanisms between the annotations and the data that is described, including, in particular, means of segmentation for temporal media.

Ultimately, when describing multimedia content on the Web, one has to pick a language suitable for doing so. Despite the different representational goals in the ISO and W3C approaches, at least both use the same serialization language: XML. The two approaches differ, however, widely in the way XML is used to describe multimedia content.

2 Semantic Web versus MPEG-7: Interoperability

In the following analysis of both description approaches we discuss various problems related to syntactic interoperability between the main languages used, namely XML, MPEG-7 DDL, RDF, RDF(S) and OWL. We then examine solutions and problems regarding semantic interoperability, in particular related to the definition and mapping of semantic-based descriptions. We analyze the ability of W3C and ISO technologies to address the expressiveness of media units to facilitate the process of audio-visual signification of multimedia. Finally, we consider the practical applicability of the provided concepts, methods and technologies.

2.1 Syntactic Interoperability: MPEG's DDL vs XML Schema

Within MPEG-7, the Description Definition Language (DDL) [6] is intended to address the language requirements listed in bullet points 1 – 4. The DDL provides basically the same structure-oriented language elements as XML-Schema[1]. The

[1] `http://www.w3.org/TR/xmlschema-1/`

only extensions to XML-Schema cover the ability to define arrays and matrices and to provide two additional datatypes, `basicTimePoint` and `basicDuration`, which allow specific temporal descriptions (see [6], pp. 9 – 14). Any available MPEG-7 parser addresses consequently only these extensions in addition to the other XML Schema-based language constructs.

Figure 1 on the following page and Figure 2 on page 4 provide a small example of a piece of MPEG-7 metadata. The first half of the example shows how to address the target video fragment. Note that in addition to the URI, the `MediaTime` is used to identify the first eight minute segment of the video file this piece of metadata applies to. Moreover, this part of the example shows how the coding format of the audio-visual component can be described by using the `MediaFormat Descriptor`. Here the description of the video covers its aspect ratio, the frame size and the frame rate per second.

Figure 2 on page 4 illustrates how the segmentation of the video in scenes and subscenes (`TemporalDecomposition`) can be achieved, where the `MediaTimePoint` provides the temporal start point of the audio-visual segment based on a Gregorian time scheme and the `MediaDuration` specifies the duration of the segment. This also illustrates a simple way of providing extra semantic annotations for a particular sequence supported by the `semantic` element.

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

1. lacks particular media-based datatypes. The datatypes used in the example are either standard XML-Schema datatypes (such as integers etc.) or media-specific datatypes defined in part 5 of the MPEG-7 standard, the Multimedia Description Schemes (MDS) [9].

2. does not 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. Again, the locating and segmentation techniques used in the example are plain URIs combined with descriptors for time segments, also defined in the MDS.

3. does not support the definition of semantic relations, as does RDF Schema[2], or ontology-based modeling, such as DAML+OIL[3] or OWL[4]. Semantics of relations between the syntax constructs, such as being used in the example of Figure 1 on the next page, are often only defined by English prose in the text of the standard, and hence lack the formal semantics that the Semantic Web languages have.

The strength of the DDL, however, lies in supporting the definition and adaptation of schemata. This is used in MPEG-7 to define normative schemata that provide not only the necessary syntax but also facilitate the description of the semantics of a single multimedia object or collections in the form of a multimedia unit. These schemata, however, are not part of the description language, but of the MDS. Here we find a plethora of structures for:

- specific datatypes required for the description of form and substance of media expression ([9] pp. 49–103]. Extensions are provided in the parts Visual [7] and Audio [8];
- linking, identification and localization tools, mainly based on XPath but extended with particular temporal constructs, that provide a basic means of establishing references within a description and linking to the associated multimedia data ([9], 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 organization of relations is restricted to a defined set of 11 topological and set-theoretic graph-relation types ([9], pp. 179–191);
- 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 ([9], pp. 251–400 and 458–540);
- a set of 45 semantic relations that allow the description of narrative structures ([9], pp. 401–457).

The syntactic description of general multimedia datatypes is thus not part of the description language, but is an integral part of concrete schemata embedding their specific semantics. The consequences are far reaching. As the essential semantic aspects for the description of multimedia are defined in standardized schemata they have to be used in the provided way and any modification, including the combination of schemata, will be outside the scope of the standard. More crucially, any modification on one of the "language related" schemata will not only alter the semantics of the description but also the description language itself. Such modifications are, however, unavoidable as a great number of schemata describe solutions for particular problems for a fraction of multimedia applications[5].

[2] `http://www.w3.org/TR/rdf-schema/`
[3] `http://www.daml.org/2001/03/reference.html`
[4] `http://www.w3.org/TR/owl-ref/`

[5] Moreover, dispersing language elements into description schemata asks for an evaluation complexity close to a validation level no parser can

```
<?xml version="1.0" encoding="UTF-8"?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001">
  <Description xsi:type="ContentEntityType">
    <MultimediaContent xsi:type="AudioVisualType">
      <MediaFormat>
        <VisualCoding>
           <Format href="urn:mpeg:mpeg7:cs:VisualCodingFormatCS:2001:1"
              colorDomain="color">
              <Name xml:lang="en">MPEG-1 Video</Name>
           </Format>
           <Pixel aspectRatio="0.75" bitsPer="8"/>
           <Frame height="288" width="352" rate="25"/>
        </VisualCoding>
      </MediaFormat>
      <AudioVisual id="Sue-and-martin-home-1">
        <MediaLocator>
          <MediaUri>http://www.example.com/videos/yup_lifestyle.mpg</MediaUri>
        </MediaLocator>
        <MediaTime>
          <MediaTimePoint>T00:00:00</MediaTimePoint>
          <MediaDuration>PT0H08M00S</MediaDuration>
        </MediaTime>
        ...
```

Figure 1: Example of a MPEG-7 sequential description (I): linking to a video fragment.

Thus, the MPEG-7 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. An issue that needs addressing is the identification of semantically relevant syntax elements in the semantic-related schemata and to include them into the DDL. This would allow the semantic web to make use of the implicit semantics of low-level binary descriptors of MPEG-7.

On the other hand, the lack of explicit semantics in MPEG-7 is, to some extent, inherent to the direct use of XML. The XML level of "self description" is limited to the extent that XML is only able to define the *syntax* of the elements in a language. There is no understanding of anything other than the hierarchical, syntactical structure of the document. What is needed is some way of specifying the semantics that is supposed to be communicated by the syntactical XML document structures [1]. However, to make these semantics explicit, and to communicate them in a machine understandable way, XML in itself is insufficient. Other layers, built on top of XML, are required to accomplish this. Within the Semantic Web, however, the semantics of these upper layers are RDF-based, and envisioned to be themselves machine-readable as much as possible. As a result we have a problem of syntactical interoperability between the two main developments (XML Schema in MPEG-7 and RDF-Schema for the Semantic web). As this is an important issue we investigate it now in more detail.

cope with. In fact, at the time of writing there is no MPEG-7 validator that can handle all the existing structures.

2.2 Syntactic interoperability: RDF vs XML

Both the Semantic Web and MPEG-7 metadata build syntactically on top of XML. Unfortunately, this does not solve even the syntactic interoperability issues for applications that need to use both approaches simultaneously. In particular, the use of RDF in most Semantic Web applications causes interoperability problems. While the decision to build the Semantic Web on top of RDF is often taken for granted, it results in a potentially large number of low level, pure syntax-oriented interoperability problems (that is, the type of problems XML was intended to solve).

Suppose that the "lifestyle video" fragment from the example scenario in part I of this article is published on the Web, distributed under an "Open Publication License". That this Web resource is indeed open content could, by interpreting the surrounding text on the HTML page from where it is linked to, be obvious to human readers, but not to a machine. To make this explicit, one could state this explicitly in RDF, and attach this statement as metadata to the Web page. In RDF triple terminology: the URL of the page (say, the (relative) URL `yup_lifestyle.mpg`) would denote the resource, the "dc:rights" label the prop-

```
        ...
        <TemporalDecomposition>
          <AudioVisualSegment id="Sue-firstphone-unwrapping">
            <Semantic><Label><Name>surprise</Name></Label></Semantic>
            <PointOfView viewpoint="martin">
              <Importance><Value>0.7</Value></Importance>
            </PointOfView>
            <PointOfView viewpoint="sue"/>
            <MediaTime>
              <MediaTimePoint>T00:00:48</MediaTimePoint>
              <MediaDuration>PT0H16M42S</MediaDuration>
            </MediaTime>
          </AudioVisualSegment>

          <TemporalDecomposition>
            ...
          </TemporalDecomposition>
        </TemporalDecomposition>

        <TemporalDecomposition>
          <AudioVisualSegment id="Sue-riding-car">
            <Semantic><Label><Name>stormy</Name></Label></Semantic>
            <MediaTime>
              <MediaTimePoint>T00:06:21</MediaTimePoint>
              <MediaDuration>PT0H00M14S</MediaDuration>
            </MediaTime>
          </AudioVisualSegment>
          <AudioVisualSegment id="Martin-with-children">
            ...
          </AudioVisualSegment>
        </TemporalDecomposition>
      </AudioVisual>
      </MultimediaContent>
  </Description>
</Mpeg7>
```

Figure 2: Example of MPEG-7 sequential description (II) the actual annotations describing the content of the video

Figure 3: Simple graphical representation of an RDF triple

erty, and the string “OPL” the value. Figure 3 shows the common graph notation.

While RDF in itself is syntax neutral, it defines an XML serialization syntax for interchange. In addition, it defines an abbreviated form. As a result, even the simple, single triple defined above can be serialized to XML in two ways, as shown in Figure 4.

Applications are expected to implement both forms and annotators are thus free to mix the two. In practice, many RDF files use both forms, making it hard to process using generic XML tools. For example, it is almost impossible to write an XSLT stylesheet for any but the most trivial RDF documents. This problem is exacerbated because the order in which RDF triples are serialized is irrelevant for most

```
<!-- Serialization syntax: -->
<rdf:Description rdf:about="yup_lifestyle.mpg">
  <dc:rights>OPL</dc:rights>
</rdf:Description>

<!-- Abbreviated syntax: -->
<rdf:Description rdf:about="yup_lifestyle.mpg"
                 dc:rights="OPL" />
```

Figure 4: Example of two XML serializations of the same RDF statement.

RDF applications. This is, however, not the case for XML applications where the order is significant. Similarly, an RDF application might decide to serialize descriptions in a nested form which does not change the RDF semantics. Again, in XML, the nesting of elements is significant and can thus not be changed.

So while RDF technically uses XML, it makes it very hard to use generic XML tools for RDF processing. Unfortunately, the reverse also holds, so that it is difficult to make RDF tools process generic XML [10]. Suppose that, in addition to the RDF metadata of our video fragment, our application also has access to the MPEG-7 metadata shown in Figure 1 and Figure 2. Despite the fact that it is encoded in XML, most RDF-based Semantic Web applications will not even be able to parse this on a syntactic level, unless one uses a non-standardized translation from MPEG's XML-based syntax to RDF, as advocated by Hunter [3].

The syntactic problems between the two major approaches in multimedia content description are, however, not the only issues that make it difficult to merge multiple sets of metadata. There are also different ways of defining semantics, as discussed below.

2.3 Semantic interoperability: defining semantics

The Semantic Web itself does not define any multimedia specific semantics. For defining application-specific semantics the Semantic Web relies on third-party specifications. The meaning of the "dc:rights" property in Figure 2, for example, is defined by the Dublin Core Metadata Initiative [2]. Attaching RDF metadata to a particular segment of a video (as is done in the MPEG-7 example in Figure 2) requires a way of addressing that specific fragment. Specification of such an addressing scheme is not considered to be within the scope of RDF or the other Semantic Web languages. Instead, it is left to a third party to develop such a scheme.

The approach of defining semantics on the Semantic Web is to provide relatively thin, but generic, layers that define increasingly complex semantic structures, and to defer the definition of domain and application specific ontologies to third parties. This approach can be contrasted to the MPEG-7 approach, which defines metadata syntax and semantics within the MPEG-7 standard. It also defines both the framework (including the DDL) and the actual ontologies. A large number of schemata in MPEG-7 establish ontological structures, as 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 [9]). 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 the schemata as ontology items. Nevertheless, the advantage of the approach taken by MPEG-7 is that it provides a large vocabulary of description terms, developed specifically for describing audiovisual material. A disadvantage is that the result is monolithic, with structures that are hard to be reused outside the MPEG-7 context. This problem cannot be underestimated, as the definition of semantics lays the groundwork for the next problem - that of mapping semantics.

Thus, with respect to the aim of the Semantic Web to make use of third-party specifications, the schemata developed in MPEG-7 are most relevant. As outlined earlier, there are particular language barriers that have to be removed before a full integration is possible. Moreover, it seems to us that the encapsulated nature of MPEG-7 needs to be opened up, namely through further modularization, to allow easier accessibility of the available schemata.

Even if the discussed issues can be resolved, there is more to be considered. In part I of this article we pointed out that the nature of annotations is necessarily imperfect, incomplete, and preliminary because they accompany and document the dynamic progress of understanding a concept, which usually open up questions of subjective interpretation. Thus, there is a need for mechanisms to establish collective sets of descriptions growing over time. The problem we then face is that of mapping semantics to make use of such structures.

2.4 Semantic interoperability: mapping semantics

The question to be asked, with respect to the Semantic Web, is: should an ontology layer use RDF(S) as its serialization syntax, or is it better to develop a (more concise) syntax directly in XML? In the RDF-based approach, one runs the risk of making integration with current and future XML-based approaches harder. Needless to say, the majority of current Web applications is XML-based, and even the MPEG-7 metadata framework is based on XML, not RDF. In addition, by using RDF syntax and building incremental syntax layers on top of that, one also needs to make sure that the underlying semantics can be layered in a similar fashion (for example, consider the potential problems when a pure RDF application interprets the semantics of an OWL document using the RDF serialization syntax. Ideally, the conclusions of the RDF application should be a subset of the conclusions an OWL application would make, but the two should not contradict one another).

On the other hand, by building the ontology layer directly on XML, one runs the risk of the development of two incompatible Semantic Webs: an XML/ontology-based "knowledge" Web versus an RDF/RDF Schema-based "metadata" Web. Clearly, the Web-Ontology Working Group chose the RDF-based approach. But the XML vs RDF question (see section 2.2) is closely related to one of the big controversies surrounding the Semantic Web in general: the question of whether the advantages of developing a common Semantic Web language stack, such as proposed by Tim

Berners-Lee[6], really outweigh the more pragmatic approach of defining knowledge interchange formats directly in XML on a per application domain and per user community basis. The latter is the approach many E-business initiatives are currently taking. In theory, a Semantic Web-based approach would require less *a priori* commitment between the different user groups, and would promote the use of generic (free and commercial) tools. The Semantic Web would standardize more levels of the information stack, agreement about the semantics defined by these levels and the possibility of using off-the-shelf tools would thus come "for free". In the XML-based alternative, users from a specific community would need to agree on these levels first, and then develop their own tools. Second, when new users add their own set of knowledge bases and tools, the Semantic Web promises a better infrastructure for interoperability between the two worlds. It still remains to be seen to what extent these promises prove to be realistic in practice.

MPEG-7's approach towards semantic interoperability is also critical because the standard acts as an ontology definition language as well as an ontology (see comments in section 2.1). This ambiguous conceptual state is a result of the decision to model the DDL on XML Schema rather than on RDF Schema. This choice was mainly political, as RDF Schema was at that time, and still at the time of writing, not a W3C recommendation and was thus not referable (for more insights w.r.t. the relationship between XML Schema and RDF Schema, see [1, 4]). The choice for XML-Schema as the serialization syntax had far reaching consequences. As a syntax oriented language, the DDL is inadequate as a basis on which to provide reasoning services, in particular subsumption based reasoning on the class and properties hierarchies. This required formal semantics that had to be established elsewhere, resulting in the `Semantic description tools` section in the MDS ([9], pp. 401-457). That MPEG-7 defined its own ontology environment is an asset with respect to interoperability within MPEG but it turned out as being a hurdle for the interoperability with other ontologies, as necessary mechanisms to connect into a source, such as an ontology, were not developed and the available linking mechanisms in MPEG-7 into external sources cover only other MPEG-7 documents or media items.

A potentially solution to overcome this problem of ontology interoperability is the `Classification schema`. This facilitates the organizational wrapper for a controlled vocabulary built out of terms and the relations between them. The relations organize 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 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.

A final issue to be addressed is the strong focus of both the current Document Web and the future Semantic Web on textual XML and page-based layout. Many of today's Web metadata and linking technology does not address the special needs of multimedia. The MPEG metadata framework was specifically developed to address those multimedia-specific issues.

3 Semantics for media expressiveness

In part I of the article we outlined that it is the process of audio-visual signification of multimedia objects that requires special attention when it comes to semantics. 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. Take the incorporated video sequences from our business authoring scenario in part I. These videos were not only added to the presentation to strengthen the logical flow within the presentation by conveying the lifestyle of the new product's target audience but the material also has to express the expectations of the audience (in the scenario the board of directors) and has to fit into the overall style of the presentation.

Support for the form of expression requires a rich set of presentation models. The following discussion is predominantly focused on MPEG-7 as the standard is devoted to representing the form and substance of media expression, whereas these issues are out of scope of the current W3C recommendations.

Despite the semantic relations, already introduced in section 2.1, MPEG-7 additionally suggests schemata that provide structures for multimedia summaries, points of view, partitions and variations ([9], pp. 458 – 540) and various forms of collections on a probabilistic, analytical or classification level ([9], pp. 541 – 600). These schemata are very detailed, but they impose particular semantics on the user. In fact, the approach taken by W3C, as illustrated in SMIL, representing a textual serialization of temporal and spatial aspects for multimedia presentations seems more promising because it is less rigid and thus more easily applicable.

[6] `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`

Similarly problematic is the the approach taken by MPEG-7 for representing substance of expression, i.e., the semantics of low-level audio and visual features as specified in the parts Visual [7][7] and Audio [8][8]. It must be clearly stated that it is not so much the conceptual ideas described in standard that are problematic. The dilemma is rather caused by the attempt to solve the challenge of representing the dynamic nature of audiovisual semantics by providing both a binary (algorithmic) and textural (schema) description structure. The intention is that both representational forms 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." ([5], p. 10).

This, however, turns out not to be the case. Both parts provide many semantic descriptions relevant for the interpretation of the individual binary format of a schema. Take the `ColorStructureType` ([7], pp. 50-56) as an example. The descriptor specifies both color content (similar to that of a color histogram) and the structure of this content. The binary format is accompanied by long textual and graphical descriptions giving detailed information about the extraction algorithm, re-quantization, color space and color quantization, and the `raw ColorStructure histogram accumulation`. All of this information is required to understand the meaning of every single element (bin) specified in the ColorStructure descriptor array of 8-bit integer values, $h(m)form \in \{0, 1, \ldots, M-1\}$.

None of this, however, made it into the textual description. In fact, the schema provides only the structure of the result space, that is the size of the matrix that contains the results of the extraction algorithm (see the DDL representation syntax on [7], p. 51). The assumption during the development of the audio and visual schemata was that an agent would know about the semantics of a bin in the ColorStructure Schema and thus could react accordingly. The result is that the semantics of the array are not made explicit but are hidden in the standards document. However, for real analytic parity of audio-visual media within the Semantic Web it is of utter importance that the semantics of a media unit are made explicit, in particular as an XML-based parser is not able to evaluate the binary representation or the quasi binary representation of the current array content. 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.

Having analyzed the conceptual ideas of the two standards it seems that, despite the fact that both build on XML, their significant incompatibilities make it very difficult to establish a general framework for describing the semantics of audio-visual information units in a machine accessible way. Yet, both approaches provide relevant solutions to address the general problems of metadata production. However, the major issue within metadata production, namely its labor intensitivity, was not really addressed yet in our discussion. In part I of this article we, however, clearly stated that a Semantic Web can only emerge if the abstract idea of the media-aware Semantic Web can be turned into an environment that integrates the instantiation and maintenance of the dynamic structures into the actual working process. The next section reflects on these issues.

4 Applicability of semantic structures

In the first part of this article we argued that a future media-aware Semantic Web, where a great variety of media will be constantly generated, manipulated, analyzed, and commented on, can only emerge if people are provided with tools that support the dynamic nature of audio-visual media and the variety of data representations and their combinations. We also showed that current technologies to support the instantiation and maintenance of the dynamic structures are still in their infancy. The question is: are the methodologies provided by the two major approaches capable of supporting the emergence of a media-aware Semantic Web as desired?

Our discussion on syntactic and semantic interoperability in sections 2.1 – 2.4 already demonstrated that the layered approach used in W3C technology seems to address the flexibility of descriptive structures, the essential requirement for intelligent media- and metaproduction, better than the philosophy of the "universal" description schema for a domain as provided by MPEG-7. However, the current state of Semantic Web technology is intrinsically biased towards describing XML-encoded content.

Though MPEG-7 provides a better means of describing (streaming) media content, its crucial dilemma is its structural complexity that obstructs the take-up of the standard. Instances of the complexity problem within MPEG-7 are:

- A description of a media item is basically forced into one document (see the definition of the root element in the MDS ([9], pp. 17-19). The instantiation of a complete description structure can be attached to the relevant media items and, naturally, the resulting descriptions are consistent and interoperable within MPEG-7, even if the descriptions vary in their instan-

[7] Examples of features are: color, texture, shape, motion, or localization

[8] Examples of features are: series types (scalable, scalar, vector etc.), waveform, power, spectrum, harmonicity, silence, sound, spoken content, etc.

tiated depth. 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[9].

- Links in MPEG-7 do not provide any information about the semantics of the relationship between documents. MPEG-7 relations, which supply the desired semantics through the introduction of `relationship` elements, can only be applied within a document. This again results in encapsulating the required network structure in a single document.

- There are a great number of `abstract` elements, which are used to establish class structure[10]. 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 theoretical overhead is actually not required.

- The interlocked nature of schemata, providing an ontology-like yet general set of schemata for describing media semantics, makes it very difficult for a user to identify the appropriate schemata and to use them in isolation. At the moment it is still not clear how the currently discussed MPEG-7 profile/level version 2 profiling will address this problem

- Due to the lack of a fundamental data model the structures provided show inconsistencies and duplications, which makes manual schemata generation difficult.

Compensating the structural complexity would require support tools that help during the complex process of schema development and maintenance, but few support tools exist for the manual generation of new schemata. The situation is more bleak with respect to semi-automated tools, such as technology that can handle (e.g., locate, transfer, integrate) multimedia segments and fragments, using the annotations, as described in the first part of this article. Note that tool support for W3C technology in commercial media production environments is also scarce. This fact indicates for both standardization activities that they still operate on a theoretical level where the everyday use does not have the highest priority in the development agenda. In the short term we see an analogous development as at the beginning of the WWW, where only the introduction of user applicable graphical tools turned the predominantly academic infrastructure into a public environment.

There are, however, projects in real world domains, such as the TV Anytime Forum, that give an indication of how media-aware semantic structures, such as those 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 proprietary 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 their organizational structure. TV Anytime includes, for example, the MPEG-7 schemata on user-modeling, though without incorporating the complete MPEG-7 organizational overhead. Rather, TV Anytime uses MPEG-7 as a namespace and is thus able to incorporate only the required schemata [12].

Examples of other media-based standards that would benefit from a standardized approach to re-usable multimedia semantics are the following:

- the Dynamic Metadata Dictionary-Unique Material Identifiers (UMIDs) [11]. UMIDS provides the link between the content (video, audio, graphics, stills etc.) and the metadata and generates a time code and date (time-axis) for synchronizing this data;

- the Multimedia Home Platform (MHP) as part of the Digital Video Broadcasting (DVB) Project[11]. MHP is a series of measures designed to promote the harmonized transition from analogue TV to a digital interactive multimedia future;

- the P/Meta Standard developed by the Production Technology Management Committee (PMC) of the European Broadcasting Union (EBU), using the Standard Media Exchange Framework (SMEF) by the British Broadcasting Corporation (BBC) and SMPTE outputs, provides a common exchange framework and a format between members (and others)[12];

- the TV Anytime Forum[13] is an association of organizations that develops specifications to enable audio-

[9] 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.

[10] 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 earlier

[11] `http://www.dvb.org/latest.html`

[12] `http://www.ebu.ch/`

[13] `http://www.tv-anytime.org/`

h!

	MPEG-7	Semantic Web
Syntax	XML	XML/RDF
Schema/ontology language	MPEG-7 DDL/XML Schema	RDF Schema/OWL
Composition	– – monolithic/big	+ – (too) many small layers
Extensibility	– – aiming at completeness	+ + designed to be extended
Multimedia ontologies	+ + part of the spec	– – third party
Linking into media items	+ + part of the spec	– – media dependent, incomplete
Available tools	– not even a complete parser	+ (including open source)
Real life applications	– –	+/– mainly RDF, few RDFS/OWL

Table 1: Multimedia metadata: MPEG-7 vs Semantic Web.

visual and other services based on mass-market, high-volume digital storage;

- the Dublin Core Metadata Initiative [2];
- NewsML[14] is an XML-based standard to represent and manage news throughout its life cycle, including production, interchange and consumer use;
- the Gateway to Educational Materials project[15]. A U.S. Department of Education initiative that expands educators' capability to access Internet-based lesson plans, curriculum units and other educational materials;
- The Getty Research Institute's Vocabulary Databases (the Art & Architecture Thesaurus®, the Union List of Artist Names®, and the Getty Thesaurus of Geographic Names™[16]), contain terminology and other information about the visual arts, architecture, artists, and geographic places.

5 Conclusions and future research

In part I of this article we argued that in media production environments metadata needs to accompany and document the entire production process. Creating such annotations, either manually or (semi-)automatically, is difficult, labor intensive and subjective. In spite of this, we argued the need for flexible, collective sets of descriptions that grow over time and are collected during different stages of the working process: generation, restructuring, representing, resequencing, repurposing and redistribution of media.

In order to provide support for this, in part II we analyze the differences and similarities of the approaches taken in MPEG-7 and the Semantic Web. These are summarized in Table 1. Our analysis shows that neither approach satisfies our requirements for a media-aware Semantic Web. Indeed, although both approaches are XML-based, the differences on a philosophical and implementation level are substantial enough to make a merge between the two complicated from a technical perspective and virtually impossible from a political perspective. The incompatibilities first need to be overcome before true, large scale, Web-based interoperability can be attempted. We do show, however, that both approaches provide potential techniques for establishing a media-aware Semantic Web.

The problems within MPEG-7 regarding the fusion of language syntax and schemata semantics demonstrates that a closed approach hinders the required modularity for description design, obstructing the needed interoperability on a syntactic and semantic level. Specific modules of our desired media description language could adopt a number of description constructs from the visual and audio parts from MPEG-7. These could then be used to describe media aspects only and would allow linking into conceptual and contextual descriptions expressed in semantic languages such as RDF, RDF Schema or OWL. It seems, however, that MPEG is taking steps in the direction of modularity. At the moment it is discussed whether MPEG-21 should be the last standard in the series of ISO multimedia standards. Having closed the standardisation work the MPEG group would then function in an advisory role that provides domains with taylored multimedia schemata libraries for their needs or advises them how to develop them. This would be an interesting direction and it is worth keeping an eye on further developments within MPEG.

Further developments towards a robust media-aware Semantic Web depend on "resolution" technology, i.e., technology that can handle (e.g., locate, transfer, integrate) multimedia segments and fragments, via the annotations. Such technology does not yet exist on a sufficiently large scale. This has the greatest consequences for a robust multimedia web, where the lack of appropriate technology is currently the major obstacle for swift development. The main task is

[14] `http://www.newsml.org/`
[15] `http://www.thegateway.org/`
[16] `http://www.getty.edu/research/tools/vocabulary/`

thus to provide real world cases that show the application of semantic-enabled technology, including maintenance tools and technology that facilitates the use of established semantic descriptions.

An ideal media-aware metadata language should be applicable beyond the context for which it was initially designed. For this it needs to be syntax-neutral and modular. In addition, tool support for human creativity is needed. Designers need to be supported in the creation of the best material for the required task and while doing so be assisted in extracting the significant syntactic, semantic and semiotic aspects of the content they are developing. For this to happen, however, those who develop the technology require a better understanding of the domains for which they are developing.

In parts I and II of this article we gave an overview of the relevant problems have given some hints as to how they can be tackled. Our goal as a research community is to investigate the basic conceptual, perceptual and processable elements of that volatile thing called multimedia information to build the fundamental framework right first time. This will then allow us to exploit the evolutionary process of semantic-based multimedia information exchange.

Acknowledgments

Part of the research described here was funded by the Dutch national Token2000/CHIME and NWO/NASH projects, and Ontoweb, a thematic network of the European Comission. The authors wishes to thank in particular Wolfgang Putz from FHG-IPSI in Darmstadt and Jane Hunter from DSTC in Brisbane for insightful discussions and helpful comments. We also wish to thank our colleague Lloyd Rutledge for useful discussion during the development of this work. Finally, we wish to thank the anonymous reviewers of IEEE MM for their detailed and valuable comments.

References

[1] S. Decker, S. Melnik, F. V. Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks. The Semantic Web: The roles of XML and RDF. *IEEE Internet Computing*, 15(3):63–74, October 2000. `http://www.computer.org/internet/ic2000/w5063abs.htm`.

[2] Dublin Core Community. Dublin Core Element Set, Version 1.1, 2003. `http://www.dublincore.org/documents/dces/`. ISO Standard 15836-2003 (February 2003), http://www.niso.org/international/SC4/n515.pdf; NISO Standard Z39.85-2001 (September 2001), http://www.niso.org/standards/resources/Z39-85.pdf; CEN Workshop Agreement CWA 13874 (March 2000), http://www.cenorm.be/isss/cwa_download_area/cwa13874.pdf.

[3] J. Hunter. Adding Multimedia to the Semantic Web — Building an MPEG-7 Ontology. In *International Semantic Web Working Symposium (SWWS)*, Stanford University, California, USA, July 30 - August 1, 2001. `http://www.semanticweb.org/SWWS/program/full/paper59.pdf`.

[4] J. Hunter and C. Lagoze. Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles. In *The Tenth International World Wide Web Conference*, pages 457–466, Hong Kong, May 1-5, 2001. IW3C2, ACM Press. `http://www10.org/cdrom/papers/572/`.

[5] ISO/IEC. Text of ISO/IEC 15938-1/FDIS Information Technology - Multimedia Content Description Interface - Part 1: Systems. ISO/IEC JTC 1/SC 29/WG 11/ N4285, Singapore, March 2001.

[6] ISO/IEC. Text of ISO/IEC 15938-2/FDIS Information Technology - Multimedia Content Description Interface - Part 2: Description Definition Language. ISO/IEC JTC 1/SC 29/WG 11 N4288, Singapore, September 2001.

[7] ISO/IEC. 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, Sidney, July 2001.

[8] ISO/IEC. Text of ISO/IEC 15938-4:2001(E)/FDIS Information Technology - Multimedia Content Description Interface - Part 4: Audio. ISO/IEC JTC 1/SC 29/WG 11/N4224, Sydney, July 2001.

[9] ISO/IEC. Text of ISO/IEC 15938-5/FDIS Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes. ISO/IEC JTC 1/SC 29/WG 11/N4242, Singapore, September 2001.

[10] P. Patel-Schneider and J. Siméon. The Yin/Yang Web: XML Syntax and RDF Semantics. In *The Eleventh International World Wide Web Conference*, Honolulu, Hawaii, May 7-11, 2002. IW3C2, ACM Press. `http://www2002.org/CDROM/refereed/231/`.

[11] Society of Motion Picture and Television Engineers (SMPTE). Standard 330M-2000 for Television-Unique Material Identifier (UMID).

Standard 330M-2000 for Television-Unique Material Identifier (UMID), SMPTE, White Plains, N.Y., 2000.

[12] The TV-Anytime Forum. Specification Series: S3 On: Metadata Corrigenda 1 to S-3 V1.1. COR1_SP003v1.1, December 2001.