That Obscure Object of Desire: Multimedia Metadata on the Web, Part 2

Frank Nack, Jacco van Ossenbruggen, and Lynda Hardman

The World Wide Web Consortium (W3C) and the International Standards Organization (ISO) have developed technologies that define structures for describing media semantics. Although both approaches are based on XML, a number of syntactic and semantic problems hinder their interoperability. In Part 2 we discuss these problems as well as ontological issues for media semantics and the problems of applying theoretical concepts to real-world applications.

In media production environments, as we argued in Part 1 of this article, metadata must accompany and document the entire production process. Creating such annotations—either manually or (semi)automatically—is difficult, labor intensive, and subjective. Nonetheless, we need flexible, collective sets of descriptions that grow over time and are collected during different stages of the working process: media generation, restructuring, representing, resequencing, repurposing, and redistribution.

To support the development of such descriptions, Part 2 analyzes the differences and similarities of the International Standard Organization's MPEG-7 and the World Wide Web's Semantic Web approaches, and their ability to define structures for describing media semantics.

Language for describing multimedia content

Machine-processable content is the main prerequisite for the more intelligent Web services constituting the Semantic Web. To build tools that are aware of the semantics of multimedia content and context, we need a language that makes the semantics of media units explicit. As we discussed in Part 1, to facilitate the description of multimedia content, such a language should

- be platform and application independent and human and machine readable;
- support a definition language for media content description structures at various levels of detail, including a rich set of syntactic, structural, cardinality, and multimedia data-typing constraints;
- support the definition of the various spatial, temporal, and conceptual relationships between the media items in a commonly agreed-upon format;
- facilitate a diverse set of linking mechanisms between the annotations and the data being described, including a way to segment temporal media.

Ultimately, describing multimedia content on the Web requires a suitable language. Despite the different representational goals in the International Standards Organization (ISO) and the World Wide Web Consortium (W3C) approaches, both use the same serialization language: XML. However, the two approaches differ widely in how they use XML to describe multimedia content.

Interoperability of the Semantic Web and MPEG-7

Both the Semantic Web and MPEG-7 experience various problems related to the syntactic interoperability between the main languages used—namely XML, the MPEG-7 document-description language (DDL), the resource description framework (RDF), RDF Schema (http://www.w3.org/TR/rdf-schema), and Web Ontology Language (OWL, http://www.w3.org/TR/owl-ref). Semantic interoperability, particularly related to defining and mapping semantic-based descriptions, also faces several obstacles. Both technologies address the expressiveness of media units to facilitate the process of audiovisual (AV) signification of multimedia, with varying success.

MPEG's DDL versus XML Schema

The MPEG-7 DDL[1] addresses the language requirements listed earlier, providing basically the same structure-oriented language elements as XML Schema (http://www.w3.org/TR/xmlschema-1). It adds only the ability to define arrays and matrices and to provide two more data types—`basicTimePoint` and `basicDuration`—which

1070-986X/05/$20.00 © 2005 IEEE

Published by the IEEE Computer Society

```
<?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/yuplifestyle.mpg</MediaUri>
        </MediaLocator>
        <MediaTime>
           <MediaTimePoint>T00:00:00</MediaTimePoint>
           <MediaDuration>PT0H08M00S</MediaDuration>
        </MediaTime>
```

(a)

```
           …
           <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>
```

(b)

Figure 1. Examples of an MPEG-7 sequential description: (a) metadata linking to a video fragment and (b) the actual annotations describing the video's content.

allow specific temporal descriptions.[1] Available MPEG-7 parsers consequently address only these extensions, in addition to XML Schema-based language constructs.

Figure 1 gives examples of MPEG-7 metadata. Figure 1a addresses a target video fragment. In addition to the URI, the `MediaTime` element identifies the first 8-minute segment of the video

file to which this piece of metadata applies. Moreover, Figure 1a shows how the MediaFormat Descriptor can describe the AV component's coding format. In the figure, the video description covers the video's aspect ratio, frame size, and frame rate per second.

Figure 1b illustrates how we can segment a video into scenes and subscenes using TemporalDecomposition, with `MediaTimePoint` providing the AV segment's temporal start point based on a Gregorian time scheme and `MediaDuration` specifying the segment length. As the figure shows, we use the `semantic` element to add semantic annotations for a particular sequence.

The XML syntax underlying the DDL facilitates platform and application independence and human and machine readability. However, because it merely adopts XML Schema syntactic elements to represent structures in schemata form, the DDL lacks particular media-based data types. The data types in the examples in Figure 1 are either standard XML Schema data types (integers, for example) or media-specific data types, defined in Part 5 of the MPEG-7 standard, the multimedia description schemes (MDS).[2]

Moreover, the DDL doesn't facilitate a diverse set of linking mechanisms between descriptions and the data being described, including, in particular, a means of segmenting temporal media. Again, the locating and segmentation techniques in the figure are plain URIs combined with time segment descriptors, also defined in the MDS.

Unlike RDF Schema or ontology-based modeling, such as the DARPA Agent Markup Language + ontology inference layer (DAML+OIL, http://www.daml.org/2001/03/reference.html) or OWL, the MPEG-7 DDL doesn't support the definition of semantic relations. Semantics of relations between syntax constructs, such as those in Figure 1a, are often only defined by English prose in the standard's text, and hence lack the formal semantics of the Semantic Web languages.

The DDL's strength lies in its support of defining and adapting schemata. MPEG-7 uses the DDL to define normative schemata that not only provide the necessary syntax but also facilitate the description of the semantics of a single multimedia object or collections of objects in the form of a multimedia unit. These schemata, however, are part of the MDS, not the description language. The MDS provides a plethora of structures for

- specific data types for describing a media expression's form and substance,[2, pp. 49-103] with extensions provided in the standard's sections on visual[3] and audio[8] features;
- linking, identification, and localization tools, based mainly on XML Path Language (XPath) but extended with temporal constructs, letting us establish references within a description and link them to the associated multimedia data;[2, pp. 74-103]
- relation graphs in which, as in RDF, basic relation units are built on conceptual triples, letting us establish named relations between description parts (the organization of relations is restricted to a defined set of 11 topological and set-theoretic graph-relation types;[2, pp. 179-191]
- forms of spatial, temporal, and spatiotemporal segmentations for video, audio, AV, multimedia, and ink content, including a set of temporal and spatial relations;[2, pp. 251-400 and 458-540] and
- a set of 45 semantic relations allowing the description of narrative structures.[2, pp. 401-457]

The syntactic description of general multimedia data types 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. Because standardized schemata define the essential semantic aspects of describing multimedia, developers must use them as provided, and any modification, including combining schemata, will be outside the standard's scope. More crucially, modifying language-related schemata will alter not only the description's semantics but also the description language itself. However, such modifications are unavoidable because many schemata describe solutions for the problems of more than one multimedia application. Moreover, dispersing language elements into description schemata begs an evaluation complexity near validation level, which no parser can cope with. In fact, at the time of writing, there is no MPEG-7 validator that can handle all the existing structures.

Thus, the MPEG-7 approach of fusing language syntax and schemata semantics is problematic and only a first step toward a language with which we can establish semantic descriptions of multimedia syntactically. Identifying semantically relevant syntax elements in the

semantic-related schemata and including them in the DDL is an open issue and would allow the Semantic Web to use the implicit semantics of low-level MPEG-7 binary descriptors.

On the other hand, the lack of explicit MPEG-7 semantics is to some extent inherent in XML's direct use. We can use XML for self-description only to the extent that XML can define the syntax of a language's elements. It can't represent anything but the document's hierarchical and syntactical structure. We therefore need a way to specify the semantics to be communicated by the syntactical XML document structures.[4] To make these semantics explicit, and to communicate them in a machine-understandable way, XML in itself is insufficient. It requires other layers built on top of it. Within the Semantic Web, however, upper-layer semantics are RDF based and should be as machine readable as possible. As a result, we have a syntactical-interoperability problem between the two main developments: XML Schema in MPEG-7 and RDF Schema for the Semantic Web.

RDF versus XML

Both the Semantic Web and MPEG-7 metadata build syntactically on top of XML. Unfortunately, this doesn't solve the syntactic interoperability issues for applications using both approaches simultaneously. In particular, the use of RDF in most Semantic Web applications causes interoperability problems. Although we often take the decision to build the Semantic Web on top of RDF for granted, doing so can result in numerous low-level, pure syntax-oriented interoperability problems (the type of problems XML was intended to solve).

Suppose we published the lifestyle video fragment from the example scenario in Part 1 of this article on the Web, distributing it under an open publication license. That this Web resource is open content could, by interpreting the surrounding text on the HTML page linking to it, be obvious to human readers, but not to a machine. We could state this fact explicitly in RDF and attach the statement as metadata to the Web page. In RDF triple terminology, the page's URL (say, yup_lifestyle.mpg) would denote the resource, the dc:rights label the property, and the string "OPL" the value. Figure 2 shows the common graph notation.

Although RDF is syntax neutral, it defines both an XML serialization syntax for interchange and an abbreviated form. Hence, we can serialize even the simple, single triple defined in Figure 2 to XML in two ways, as Figure 3 shows.

Figure 2. Simple graphical representation of an RDF triple. The property (dc:rights) links the resource (yup_lifestyle.mpg) to the value (OPL).

```
<!- Serialization syntax: ->
<rdf:Description rdf:about="yuplifestyle.mpg">
   <dc:rights>OPL</dc:rights>
</rdf:Description>
(a)
<!- Abbreviated syntax: ->
<rdf:Description rdf:about="yuplifestyle.mpg"
     dc:rights="OPL" />
(b)
```

Figure 3. Two XML serializations of the same RDF statement: (a) XML serialization syntax and (b) abbreviated form.

Because applications are expected to implement both forms, annotators can mix the two. In practice, many RDF files use both forms, making them difficult to process using generic XML tools. For example, writing an XSL transformations (XSLT) style sheet for any but the most trivial RDF documents is almost impossible. The fact that the order in which RDF triples are serialized is irrelevant for most RDF applications exacerbates this problem. In XML applications, however, the order is significant. Similarly, an RDF application might serialize descriptions in a nested form, which doesn't change the RDF semantics. Again, in XML, element nesting is significant and thus can't be changed.

So, although RDF technically uses XML, using generic XML tools for RDF processing is difficult. Unfortunately, the reverse also holds: it's difficult to make RDF tools process generic XML.[5] Suppose that in addition to our video fragment's RDF metadata, our application has access to the MPEG-7 metadata in Figure 1. Despite its XML encoding, most RDF-based Semantic Web applications couldn't even parse it on a syntactic level without a nonstandardized translation from MPEG's XML-based syntax to RDF, as Hunter[6] advocates.

The syntactic problems between the two major approaches in multimedia content description aren't alone in making it difficult to merge multiple sets of metadata, however.

Defining semantics

The Semantic Web doesn't define any multimedia-specific semantics and relies on third-party specifications for defining application-specific semantics. The Dublin Core Metadata Initiative (http://dublincore.org), for example, defines the dc:rights property in Figure 2. To attach RDF metadata to a particular video segment (as in the MPEG-7 example in Figure 1b), we need a way to address the specific fragment. Specifying such an addressing scheme is considered outside the scope of RDF and other Semantic Web languages. Instead, a third party must develop such a scheme.

Defining semantics on the Semantic Web is accomplished through relatively thin, but generic, layers defining increasingly complex semantic structures, leaving the definition of domain- and application-specific ontologies to third parties. In contrast, the MPEG-7 standard defines metadata syntax and semantics as well as the framework (including the DDL) and the actual ontologies. Many MPEG-7 schemata establish ontological structures, as most schemata are inspired by the broadcasting and AV-based entertainment domains (see, for example, the VideoEditingSegment, AgentDS, PlaceDS, or user-preference description schemata in the MDS[2]).

The large number of schemata (often describing similar aspects of the same semantic problem) and their interlocked nature indicate the ontological role of the MDS. However, because abstraction attempts to achieve domain independence, it's impossible to use the schemata as ontology items. Nevertheless, MPEG-7 has a large vocabulary of description terms developed specifically for AV material. Unfortunately, the result is monolithic, with structures that are difficult to reuse outside the MPEG-7 context.

Thus, with respect to the Semantic Web's aim to use third-party specifications, the MPEG-7 schemata are most relevant. As outlined earlier, we must remove some language barriers before full integration is possible. Moreover, opening MPEG-7 through further modularization will allow easier access to the available schemata.

Even if we can resolve these issues, more remains to be considered. As we stated in Part 1, annotations are necessarily imperfect, incomplete, and preliminary because they accompany and document the dynamic progress of understanding a concept, which usually introduces questions of subjective interpretation. Thus, we need mechanisms to establish collective sets of descriptions that grow over time. The problem we then face is mapping semantics to such structures.

Mapping semantics

The question we must ask, with regard to the Semantic Web, is whether an ontology layer should use RDF Schema as its serialization syntax, or whether we should develop a (more concise) syntax directly in XML. With the RDF-based approach, we risk complicating integration with current and future XML-based approaches. Needless to say, most current Web applications are XML-based; even the MPEG-7 metadata framework is based on XML rather than RDF. In addition, when building incremental syntax layers on top of RDF syntax, we must also make sure we can similarly layer the underlying semantics. For example, consider the potential problems of a pure RDF application interpreting the semantics of an OWL document using the RDF serialization syntax. Ideally, the RDF application's conclusions should be a subset of the conclusions an OWL application would make, but the two shouldn't contradict each other.

On the other hand, by building the ontology layer directly on XML, we risk developing 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. However, the XML versus RDF question is closely related to one of the big controversies surrounding the Semantic Web in general: Do the advantages of developing a common Semantic Web language stack, as Tim Berners-Lee proposed (http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html), really outweigh the more pragmatic approach of defining knowledge interchange formats directly in XML on a per-application domain and per-user community basis? Many e-business initiatives are taking the latter approach.

In theory, a Semantic Web-based approach requires less a priori commitment between user groups and promotes the use of generic (free and commercial) tools. The Semantic Web would standardize more levels of the information stack, leading to agreement about the semantics defined by these levels and allowing the use of off-the-shelf tools "for free." In the XML-based alternative, users from specific communities must first agree on the levels and then develop their own tools. When new users add their own sets of knowledge bases and tools, the Semantic

Web promises a better infrastructure for interoperability between the two worlds. The extent to which these promises prove realistic in practice remains to be seen.

MPEG-7's approach to semantic interoperability is also critical because the standard acts as an ontology definition language as well as an ontology. This ambiguous conceptual state results from the decision to model the DDL on XML Schema rather than RDF Schema. This choice was mainly political, because at that time, as at the time of writing, RDF Schema wasn't a W3C recommendation and thus wasn't preferable (further insights on the relationship between XML Schema and RDF Schema are available elsewhere[4,7]).

Choosing XML Schema as the serialization syntax has far-reaching consequences. As a syntax-oriented language, the DDL provides inadequate reasoning services, particularly in subsumption-based reasoning on class and property hierarchies. This required formal semantics to be established elsewhere, resulting in the semantic description tools section in the MDS.[2] Although defining its own ontology environment enhanced MPEG-7's internal interoperability, it created hurdles for interoperability with other ontologies. The environment doesn't include necessary mechanisms to connect into a source, such as an ontology, and MPEG-7's available linking mechanisms for external sources cover only other MPEG-7 documents or media items.

A potential solution to the ontology interoperability problem is the *classification schema*, which facilitates the organizational wrapper for a controlled vocabulary built out of terms and their relations. The relations organize the terms in a hierarchy, indicating whether one term's meaning is broader or narrower than another, when terms are synonyms, or, in a given set of relations, which term is most relevant. Thus, a classification schema includes aspects of a thesaurus. The MPEG-7 classification schema lets us incorporate other classification schema, although it doesn't indicate whether it allows inclusion only of other MPEG-7 classification schemata or also the insertion of or connection to other ontologies. Unfortunately, the standard doesn't explain how to map previously unconnected terms to the schema. Thus, mapping high-level media semantics remains an unsolved problem; whether the MPEG-7 approach of profiling schemata provides a suitable solution is questionable.

A final issue is the focus of both the current document Web and the future Semantic Web on textual XML and page-based layout. Few of today's Web metadata and linking technologies address multimedia's special needs. The MPEG metadata framework specifically addresses multimedia issues.

Semantics for media expressiveness

In Part 1 we outlined the special attention the multimedia object AV signification process requires when it comes to semantics. The perceptual information that objective measurements provide (information based on image or audio processing or pattern recognition, for example) plays an important role in a multimedia unit's aesthetics and consequently its subjective interpretation.

Consider the incorporated video sequences from our business authoring scenario in Part 1. Adding the videos to the presentation strengthened its logical flow by conveying the lifestyle of the new product's target audience; however, the material must also express the audience's expectations and fit the presentation's overall style.

Supporting the form of expression requires a rich set of presentation models. We focus on MPEG-7 because it represents the form and substance of media expression, which are out of scope of the current W3C recommendations.

Despite the semantic relations introduced earlier, MPEG-7 also suggests schemata that provide structures for multimedia summaries, viewpoints, partitions and variations,[2] and various forms of collections on a probabilistic, analytical, or classification level.[2] Although these schemata are detailed, they impose specific semantics on the user. In fact, the W3C approach of representing a textual serialization of temporal and spatial aspects for multimedia presentations—as the Synchronized Multimedia Integration Language (SMIL) illustrates—seems more promising because it's less rigid and thus more easily applicable.

Similarly problematic is the MPEG-7 approach for representing an expression's *substance*—that is, the semantics of low-level visual[3] (color, texture, shape, and so on) and audio[8] (such as waveform, spectrum, harmonicity, and silence) features as specified in the standard's visual and audio parts. The concepts described in the standard aren't the problem. Rather, the dilemma results from attempting to represent the dynamic nature of AV semantics by providing both a binary (algorithmic) and a textural (schema) description structure. Both representational forms aim to provide the same information to

meet the MPEG-7 system specification 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 bidirectional lossless mapping between the textual and the binary representation is possible.[9]

This turns out not to be the case, however. Both parts provide many semantic descriptions relevant to the interpretation of a schema's individual binary format. In the `ColorStructureType`,[3, pp. 50-56] for example, the descriptor specifies both color content (similar to that of a color histogram) and color content structure. Long textual and graphical descriptions giving detailed information about the extraction algorithm, requantization, color space and quantization, and the raw ColorStructure histogram accumulation accompany the binary format. To understand the meaning of every element (bin) specified in the ColorStructure descriptor array of 8-bit integer values, $h(m)$ for $m \in \{0, 1, \ldots, M-1\}$, we need all of this information.

None of this, however, made it into the textual description. In fact, the schema provides only the resulting space structure—that is, the size of the matrix containing the extraction algorithm results (see the DDL representation syntax in the Multimedia Content Description Interface standard, Part 3,[3] p. 51). Developers of the audio and visual schemata assumed agents would know about the bin semantics in the ColorStructure schema and could react accordingly. Consequently, they hid the array's semantics in the standards document. However, for real analytic parity of AV media within the Semantic Web, a media unit's semantics must be explicit, especially because an XML-based parser can't evaluate the current array content's binary or quasibinary representation. Although this problem might seem 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 automatically annotating media.

Having analyzed the two standards, we believe that, despite the fact that both build on XML, their significant incompatibilities make it difficult to establish a general framework for describing the semantics of AV information units in a machine-accessible way. Yet, both approaches address general metadata production problems. However, we haven't yet really addressed the main metadata production issue, namely its labor intensiveness.

Applicability of semantic structures

In Part 1, we argued that a future media-aware Semantic Web can only emerge if people have tools to support the dynamic nature of AV media as well as the various data representations and their combinations. As we demonstrated, current technologies to support the instantiation and maintenance of the dynamic structures are still in their infancy. This leaves us with a question: Can the methodologies provided by the two major approaches support the emergence of a media-aware Semantic Web as desired?

Earlier in the article we demonstrated that the W3C's layered approach seems to address the flexibility of descriptive structures (the essential requirement for intelligent media and metaproduction), better than MPEG-7's "universal" description schema for a domain. However, current Semantic Web technology is intrinsically biased toward describing XML-encoded content.

Although MPEG-7 provides a better means of describing (streaming) media content, its structural complexity obstructs its adoption. For example, MPEG-7 basically forces a media item description into one document (see the root element definition in the MDS[2]). Attaching the instantiation of a complete description structure to the relevant media items generates descriptions that are consistent and interoperable within MPEG-7, even if the descriptions vary in their instantiated depth. Although the schema's structure can be complex, once created and used in instantiations, it can't be altered. Any modification would cause inconsistencies with existing documents.

Another problem caused by the standard's complexity is that links in MPEG-7 provide no 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. Again, this encapsulates the required network structure in a single document.

Numerous abstract elements for establishing class structure are available. MPEG-7 implicitly addresses the fundamental problem of determining when an instance should also be a class. (This problem is also part of the language problem described earlier.) However, abstract elements

can't appear in instantiations. When a schema declares an element to be abstract, a member of that element's substitutable class must appear in the instance document. The XML namespace mechanism (xsi:type) is used to indicate that a derived type isn't abstract. Properly using these mechanisms requires a thorough understanding of schemata development. This makes instant schemata development for distinct domains difficult, especially if the required schemata should cover simple descriptions, which actually require no theoretical overhead.

Another issue is the interlocked nature of schemata, which provides an ontology-like yet general set of schemata for describing media semantics, making it difficult for users to identify the appropriate schemata and use them in isolation. It's still not clear how the current MPEG-7 profile/level version 2 profiling will address this problem.

In addition, because no fundamental data model exists, the available structures show inconsistencies and duplications, making manual schemata generation difficult.

Compensating for MPEG-7's structural complexity would require tools to support the complex process of schema development and maintenance, but few tools for manually generating new schemata exist. The situation is bleaker with respect to semiautomated tools, such as technology that can handle (locate, transfer, integrate, and so on) multimedia segments and fragments using the annotations described in Part 1. Tool support for W3C technology in commercial media production environments is also scarce, indicating that both standardization activities still operate on a theoretical level, with everyday use having a lower priority in the development agenda. In the short term, the development of these standards is analogous to the early work on the WWW, where introducing user-applicable graphical tools turned the predominantly academic infrastructure into a public environment.

Some real-world projects, such as the TV Anytime Forum (http://www.tv-anytime.org), indicate how media-aware semantic structures will be used in the future. The TV Anytime Forum is an association of organizations that develops specifications to enable AV and other services based on mass-market, high-volume digital storage.

The semantic structures, all written in XML Schema, are proprietary and cover the essential aspects of media description—that is, content description, content referencing and location, rights management and protection, systems, and transport. Although the TV Anytime schemata are similar to the equivalent structures in MPEG-7, their organizational structure is simpler. TV Anytime includes, for example, the MPEG-7 schemata on user modeling but without the complete MPEG-7 organizational overhead. Because TV Anytime uses MPEG-7 as a namespace, it incorporates only the required schemata.[10]

Other media-based standards that would benefit from a standardized approach to reusable multimedia semantics include

- Dynamic Metadata Dictionary-Unique Material Identifiers (UMIDs),[11] which link content (video, audio, graphics, stills, and so on) to metadata and generate a time code and date (time-axis) for synchronizing this data;
- Digital Video Broadcasting (http://www.dvb.org/) Project's Multimedia Home Platform (MHP), a series of measures for promoting the harmonized transition from analog TV to a digital interactive multimedia future;
- The P/Meta standard, developed by the Production Technology Management Committee of the European Broadcasting Union, which uses the standard media exchange framework (SMEF) by the British Broadcasting Corporation and SMPTE outputs to provide a common exchange framework and format (http://www.ebu.ch);
- Dublin Core Metadata Initiative (DCMI), is an organization dedicated to fostering the widespread adoption of interoperable metadata standards and promoting the development of specialized metadata vocabularies for describing resources to enable more intelligent resource discovery systems. The Dublin Core Metadata Element Set (DCMES) was the first metadata standard deliverable out of the DCMI. DCMES provides a semantic vocabulary for describing the "core" information properties, such as "Description," "Creator," and "Date" (http://www.dublincore.org/);
- NewsML (http://www.newsml.org), an XML-based standard that represents and manages news throughout its life cycle, including production, interchange, and consumer use;

Table 1 . Comparison of MPEG-7 and the Semantic Web.

Feature	MPEG-7	Semantic Web
Syntax	XML	XML/RDF
Schema/ontology language	MPEG-7 document description language (DDL)/XML Schema	RDF Schema/Web Ontology Language (OWL)
Composition	Monolithic	Many small layers
Extensibility	Aiming at completeness	Designed for extensibility
Multimedia ontologies	Part of the specification	Third party
Linking into media items	Part of the specification	Media dependent, incomplete
Available tools	None; not even a complete parser	Open-source tools
Real-life applications	None available	Mainly RDF with a few RDF Schema/OWL

- Gateway to Educational Materials project (http://www.thegateway.org), a US Department of Education initiative that expands educators' capability to access Internet-based lesson plans, curriculum units, and other educational materials; and

- The Getty Research Institute's Vocabulary Databases (the Art and Architecture Thesaurus, Union List of Artist Names, and the Getty Thesaurus of Geographic Names, http://www.getty.edu/research/tools/vocabulary), which contain terminology and other information about the visual arts, architecture, artists, and geographic places.

Future research

Neither the MPEG-7 nor the Semantic Web approach satisfies our requirements for a media-aware Semantic Web. Indeed, although both approaches are XML-based, the philosophical and implementational differences (summarized in Table 1) are substantial enough to make merging the two complicated from a technical perspective and virtually impossible from a political perspective. Before we can attempt true, large-scale, Web-based interoperability, we must overcome these incompatibilities.

The problems within MPEG-7 regarding the fusion of language syntax and schemata semantics show how a closed approach hinders the required modularity for description design, obstructing the needed interoperability on syntactic and semantic levels. Specific modules of our desired media description language could adopt several description constructs from MPEG-7's visual and audio parts. We could use these to describe media aspects only, which 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 moving toward modularity. At the moment, whether MPEG-21 should be the last standard in the series of ISO multimedia standards is an open question. Having closed the standardization work, the MPEG group would function in an advisory role, providing domains with tailored multimedia schemata libraries or showing them how to develop them.

Further developments toward a robust media-aware Semantic Web depend on *resolution* technology—that is, technology that can handle multimedia segments and fragments via annotations. No such technology yet exists on a sufficiently large scale—a lack with greatest consequences for a robust multimedia Web, where the lack of appropriate technology is currently the major obstacle to swift development. The main task is thus to provide real-world cases showing 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 designed, and hence must be syntax-neutral and modular. It should also provide tool support for human creativity. It must support designers in creating the best material for the required task while helping them extract the significant syntactic, semantic, and semiotic aspects of the content they're developing. For this to happen, however, technology developers must better understand the domains for which they're developing.

Our overview of the relevant problems has hinted at some strategies for tackling them. As a research community, we must investigate the basic conceptual, perceptual, and processable elements of multimedia information, building the funda-

mental framework correctly the first time. Then we can exploit the evolutionary process of semantic-based multimedia information exchange. **MM**

Acknowledgments

The Dutch National Toegankelijkheid en Kennisontsluiting in Nederland 2000/Cultural Heritage in an Interactive Multimedia Environment (Token2000/CHIME) and Nederlandse Organisatie voor Wetenschappelijk Onderzoek/Networked Adaptive Structured Hypermedia (NWO/NASH) projects, as well as Ontoweb (a thematic network of the European Commission) funded part of the research described here. We thank in particular Wolfgang Putz from Fraunhofer Institut für Integrierte Publikations und Informationssysteme (FHG-IPSI) in Darmstadt and Jane Hunter from Distributed Systems Technology Centre (DSTC) in Brisbane for insightful discussions and helpful comments. We also thank our colleague Lloyd Rutledge for useful discussions during this work's development. Finally, we thank *IEEE MultiMedia*'s anonymous reviewers for their detailed and valuable comments.

References

1. *Information Technology—Multimedia Content Description Interface, Part 2: Description Definition Language*, ISO/IEC JTC 1/SC 29/WG 11 N4288, ISO/IEC, 2001, pp. 9-14.
2. *Information Technology—Multimedia Content Description Interface, Part 5: Multimedia Description Schemes*, ISO/IEC JTC 1/SC 29/WG 11/N4242, ISO/IEC, 2001.
3. *Information Technology—Multimedia Content Description Interface, Part 3: Visual*, ISO/IEC JTC 1/SC 29/WG 11/N4358, ISO/IEC, 2001.
4. S. Decker et al., "The Semantic Web: The roles of XML and RDF," *IEEE Internet Computing*, vol. 15, no. 3, Sept./Oct. 2000, pp. 63-74; http://www.computer.org/internet/ic2000/w5063abs.htm.
5. P. Patel-Schneider and J. Siméon, "The Yin/Yang Web: XML Syntax and RDF Semantics," *Proc. 11th Int'l World Wide Web Conf.*, ACM Press, 2002, pp. 443-453; http: //www2002.org/CDROM/refereed/231/.
6. J. Hunter, "Adding Multimedia to the Semantic Web—Building an MPEG-7 Ontology," *Proc. Int'l Semantic Web Working Symp.* (SWWS), 2001; http://www.semanticweb.org/SWWS/program/full/paper59.pdf.
7. J. Hunter and C. Lagoze, "Combining RDF and XML Schemas to Enhance Interoperability between Metadata Application Profiles," *Proc. 10th Int'l World Wide Web Conf.*, ACM Press, 2001, pp. 457-466; http://www10.org/cdrom/papers/572/.
8. *Information Technology—Multimedia Content Description Interface, Part 4: Audio*, ISO/IEC JTC 1/SC 29/WG 11/N4224, ISO/IEC, 2001, pp. 50-56.
9. *Information Technology—Multimedia Content Description Interface, Part 1: Systems*, ISO/IEC JTC 1/SC 29/WG 11/ N4285, ISO/IEC, 2001, p. 10.
10. TV-Anytime Forum, Specification Series S-3, *Metadata Corrigenda 1 to S-3 v. 1.1*, COR1 SP003v1.1, TV-Anytime Forum, 2001.
11. Soc. Motion Picture and Television Engineers (SMPTE), *Standard 330M-2000 for Television-Unique Material Identifier* (UMID), SMPTE, 2000.

Frank Nack is a senior researcher at the Center for Mathematics and Computer Science (CWI), currently working in the multimedia and human–computer interaction group. The main thrust of his research is on the representation, retrieval, and reuse of media in distributed hypermedia systems, and computational applications of media theory and semiotics that enhance human communication and creativity. Nack has a PhD in applied artificial intelligence from Lancaster University, UK. He is an associate editor in chief of *IEEE MultiMedia*.

Jacco van Ossenbruggen is a senior researcher at CWI, currently working in the multimedia and human–computer interaction group. His current interests include synchronized multimedia on the Semantic Web and the automatic generation of user-tailored hypermedia presentations. Van Ossenbruggen has a PhD in computer science from the Vrije Universiteit Amsterdam.

Lynda Hardman is the head of the multimedia and human–computer interaction group at CWI and part-time professor at the Technical University of Eindhoven. Hardman has a PhD from the University of Amsterdam.

Readers may contact Frank Nack at CWI, Kruislaan 413, PO Box 94079, 1090 GB Amsterdam, The Netherlands; Frank.Nack@cwi.nl.