That Obscure Object of Desire: Multimedia Metadata on the Web, Part 1

Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman
CWI

The Semantic Web and the Multimedia Content Description Interface (MPEG-7) are the two most widely known approaches toward machine-processable and semantic-based content description. The concepts and technologies behind the approaches are essential for the next step in multimedia development—that is, providing multimedia metadata on the Web. Unfortunately, as this article discusses, many practical obstacles block their widespread use.

Despite the increasing bandwidth and availability of many network-friendly multimedia codecs, finding the media content you are looking for on the Net is still as hard as ever. The next step in multimedia development is to provide multimedia metadata on the Web. The World Wide Web Consortium (W3C) and the International Organization for Standardization (ISO) address this need for machine-processable and semantic-based content description through the Semantic Web and the Multimedia Content Description Interface (MPEG-7).

The "Example Scenario" sidebar presents a typical business case. We consider this scenario challenging, particularly because the tools involved must operate, to some extent, on the semantics of the media items involved.

Traditional retrieval tasks (such as finding relevant media items) and more innovative tasks (such as generating a coherent story line from a set of media items) require a semantic understanding of media. Semantics also implies context, and hence the tools must understand the items' technical and social context, including information about copyrights and provenance.

Before we can build tools that are aware of the semantics of multimedia content and context, we must make the semantics explicit. Integrating the production, use, and management of metadata into the multimedia production chain can achieve this goal. A key issue in making this work is a commonly agreed-upon metadata exchange format.

We present this article in two parts. In Part 1, we identify problems and requirements regarding the semantic content description of media units. In Part 2, which will appear in the January–March 2005 issue, we analyze the ability of the W3C and ISO efforts to define structures for describing media semantics and discuss syntactic, semantic, and ontological problems, as well as the problems created when we attempt to apply theoretical concepts to real-world problems.

Metadata in the multimedia production chain

Audiovisual (AV) media production, such as the business presentation described in the "Example Scenario" sidebar, is a complex process. Metadata could improve this process by making information implicit in the AV content explicit. An obvious approach is to incrementally store relevant metadata during the production process and make it accessible to all tools involved.

Requirements for gathering metadata

Although media production is iterative and organic, it's traditionally a three-stage process:

- *Preproduction*—the production team determines the main ideas forming the production's core (scripting, storyboarding, and so on).
- *Production*—the production team acquires the media material (through shooting or sound recording).
- *Postproduction*—the production team makes editorial decisions (such as editing, sound mixing, presenting, and archiving) based on reviews of the material.

These production stages are highly interdependent, and tools manipulating the data at each stage must interoperate. When this process also involves producing and storing metadata alongside the original data, its complexity increases considerably. Preproduction tools must then not only produce or update scripts, but also export notes about the rationale underlying certain script modifications. Shooting produces raw

1070-986X/04/$20.00 © 2004 IEEE

footage and explicit descriptions of set activities. Postproduction tools produce edited and mixed material, but could include production schedules and editing lists with decision motivations. At each stage, changes to the metadata can affect material produced in the other stages. Dependencies among the stages themselves could even be made explicit in the preserved metadata.

Example Scenario

Imagine, five years from now, you head a lab that develops mobile communication devices and you'd like to develop the company's new product line. First, you must convince the board that your department has the vision, skill, and attitude needed to make the new product line a success. For this, you need a multimedia presentation fast (as in "by the end of today") and cheap (as in "we have no budget for this").

Your multimedia presentation authoring tool (that is, 2009's integrated successor of the Microsoft PowerPoint/Macromedia Director family) finds relevant media assets (including product-related texts, pie charts, and still and moving images) on the corporate network. Using these assets and their associated metadata, the tool generates a first preview of the presentation. Although the automatically generated story line is coherent and successfully conveys many of the important semantic relations among the retrieved media items, you aren't happy with the story line's progression and the lack of buildup.

You fire up the tool's storyboard editor and start improving the story line. The edited presentation includes some scenes that attempt to convey the lifestyle of the new product's target audience. A search on the corporate network returns neither suitable footage nor a fitting soundtrack. You're reluctant to search the public peer-to-peer file-sharing network; finding appropriate material won't be a problem, but dealing with the copyright issues will likely involve more time and money than you can afford. But, because you have no other option, you give it a try anyway.

The P2P search tool quickly shows you some previews of the material it found along with the relevant metadata. It even includes some open content material you can use directly, and some usable stock footage with reasonable licensing costs if you don't use it publicly. You select a few clips and order your digital rights management agent to deal with the legal issues and pay the required fees. (You do all this anonymously; you don't want your competitors to be able to trace the transactions.) At the end of the day, you have a presentation of sufficient quality to use for tomorrow's board meeting.

Metadata acquisition also represents the progression through the various editing stages on technical, structural, and descriptive levels, letting us preserve a media item's original context. Currently, we often lose this type of information after finishing a production and must reengineer it when needed later.

Traditional approaches for creating metadata only address the end product, using objective measurements based on image or sound processing, pattern recognition, and so on, to characterize AV information on conceptual (keyword) and perceptual levels.[1] Such retrospective approaches miss important cognitive-, content-, and context-based information describing decisions at intermediate stages in the production.

The introduction of DVDs with extra (meta) material and the many "making of …" productions have made high-quality metadata an economic asset. In most cases, however, creating this extra information will remain unrealistic if it requires extensive manual annotation (tight production and archival budgets don't cover such expensive endeavors). Instead, we need high-level support integrated into the production environment that doesn't hinder the creative and improvisatory processes so important in media production.

Within such an environment, the resulting media item, on a micro (such as a shot) or macro level (a complete business presentation, for example), can be linear in nature. The entire collection of material, however, including all intermediate physical AV data as well as the production decisions and other contextual information, composes a nonlinear and complex semantic network. Although the resulting information structure is potentially a richly connected network, it remains a collection of metadata in a particular context within a particular production.

The idea of saving the complete production process is not new, but implementing it in a digital environment remains difficult. It requires standardized representational structures reflecting the constant changes the AV material undergoes during its production, as well as dynamic semantic structures for representing conceptual developments over time.

Going further, digital media redefines traditional media forms, blurs the boundaries between production steps, and alters the information flow from producers to consumers. Consequently, we introduce another step into the production process: *metaproduction*, which involves restructuring, representing, resequencing, repurposing, and redistributing media.

The scenario described in the sidebar is a prototypical example of metaproduction, as most of the material was produced beforehand, for a different purpose and in a different context.

Any metaproduction process extends an existing semantic network: It provides additional production information and describes a different use context for existing material. A piece of metadata can change its role and turn into a piece of media needing description. For example, imagine a film theoretician who wants to demonstrate the referential quality within a particular director's work. The theoretician could link the original sequence of the referenced film with the referrer sequence. In this relation, the referrer sequence is the metadata. The station scene from Brian De Palma's *Untouchables* and the arrest scene in Terry Gilliam's *Brazil*, which both refer to the Odessa steps scene in Sergei Eisenstein's *Battleship Potemkin* are examples.

Thus, a media-aware semantic network requires

- sufficient linking mechanisms to establish context for a given media component, which exists independently of its use in a production;
- flexible description schemata that reflect a media item's varying roles (data and metadata), depending on the context in which it's used;
- evolving semantic, episodic, and technical representation structures to account for the fact that even within a single production, annotations are necessarily imperfect, incomplete, and preliminary because they accompany and document the dynamic progress of understanding a concept;
- expressive mechanisms for encoding metadata and making it accessible in a controlled way when it's to be reused across multiple productions; and
- support by production activities in generating semantic annotations during the media production process.

Addressing these requirements in an environment that integrates the instantiation and maintenance of these dynamic structures into the actual working process is a challenge.

Media production environments

A future media-aware Semantic Web should include a great variety of media to be constantly generated, manipulated, analyzed, and commented on. Such a Web can only emerge, however, if people have tools that support the dynamic nature of AV media and the variety of data representations and their combinations. At the same time, these tools must integrate with the (still mainly text-oriented) environment of current Semantic Web technology. Today's media production is mainly oriented toward one-time design and production, meaning that we lose important metadata sources when production ends, as is the case with multimedia production tools such as Macromedia Director, Adobe Photoshop, Macromedia Flash, and Microsoft PowerPoint. Because these tools work with proprietary data structures, using the internal content representation structures outside the application or for a different purpose is nearly impossible. The net result is little or no intrinsic compatibility across systems from different providers, and poor support for broader reuse of media content. Hence, we face the paradoxical situation that although we have more potential than ever to assist in media's creative development and production, we still lack environments that serve as integrated information spaces for use in distributed productions, research, and restructuring (for example, by software agents), or by the audience in direct access and navigation.

Groups in academia and industry have attempted to add extra semantics semiautomatically to AV material during meta- and regular production without interfering with established workflows[2,3] (also see the Caliph and Emir tool at http://www.know-center.at/en/divisions/div3demos.htm or the Video Wizard at http://www.video-wizard.com) These tools use standard XML-based description mechanisms and follow the paradigm of intelligent tools that rely on the existence of supportive descriptional structures.

Because these prototypes are experiments, they suffer, to varying degrees, with respect to real applicability and scalability. They're not more than a small first step toward the intelligent use and reuse of media production material. Nevertheless, these examples provide insight into generating interactive media documents in particular, and researching media representation in general. Most interesting is their potential to cooperate when common representation structures become available and implemented.

The W3C's Semantic Web and ISO's Multimedia Content Description Interface (MPEG-7) are the two most relevant common formats for machine-processable and semantic multimedia content description.

W3C and ISO approaches

Metadata plays a key role in providing machine-processable content, the main prerequisite for the more intelligent Web services constituting both the Semantic Web[4,5] and the MPEG community's intelligent media applications.[6,7] Both communities seek to provide a general metadata framework. Their approaches to providing such a framework, however, differ radically. (See the "History of the Semantic Web and MPEG-7" sidebar for details on both technologies' origins.)

History of the Semantic Web and MPEG-7

Metadata-related issues touch all information sciences. Many communities—in particular, the digital library (DL) and knowledge representation (KR) communities, and the part of the artificial intelligence (AI) community that interprets, manipulates, or generates audiovisual (AV) media (known as MM-AI)—have influenced models and technology for processing metadata. From the W3C viewpoint, the Semantic Web is an attempt to make results of DL and KR research applicable to the Web. MPEG incorporates aspects from all three communities.

Understanding the W3C's Semantic Web requires understanding the views of the DL and KR communities. The DL community sees metadata as a way to support cataloging and retrieving information in large document collections and has produced standards that address such issues, most notably the Dublin Core.[1] The Dublin Core basically standardizes a set of 15 commonly agreed-upon metadata elements of the type appearing in most library catalogs, including title, subject, and creation date.

The DL community's metadata and document-centered focus differs from the information modeling approach of the KR community, which focuses on representing the underlying content rather than describing the document containing the content. For KR researchers, a well-designed, powerful infrastructure for adding metadata to Web documents forms the basis for publishing explicit, formalized forms of knowledge directly on the Web. To what extent, and how this knowledge is associated with existing (informal) Web documents, is often a secondary issue.

The *ontology* concept is key in sharing and communicating explicit knowledge. The KR literature often defines ontology as a "specification of a conceptualization"—that is, an explicit and commonly agreed-upon definition of the objects and concepts in a certain domain. The ontology specifies these objects and concepts, the relations among them, and the rules limiting the concepts' interpretation. Given an ontology about a certain domain, parties needing to share and communicate knowledge make an *ontological commitment*—a statement that both people and applications (agents) will use the ontology's terminology according to the specified rules.

Despite the differences between the DL and KR approaches, many applications need elements from both worlds. Applications often use ontologies to control the terminology used in metadata, for example. Committing to a specific ontology can help users make annotations more systematically and consistently.[2] In addition, applications can use the ontology's background knowledge in addition to the metadata. For example, if a video's metadata only specifies that the video is about two young urban professionals, a query for "yuppie lifestyle" won't return the page. By combining the metadata with an ontology stating that "yuppie" is a common acronym for "young urban professional" and denotes a specific lifestyle, the same query will return the page.

MPEG's view on metadata is similar to the W3C's, but MPEG documents are typically complex AV units, and thus MPEG-7 focuses on a common interface for describing multimedia materials (representing information about the content, but not the content itself: "the bits about the bits"). MPEG-7 addresses interoperability and globalization of metadata resources and data management flexibility. For this purpose, MPEG-7 had to reconcile the different communities' approaches. On one side were the DL, KR, and MM-AI communities, who stressed the need for high-level descriptions of AV content; on the other side was the signal-processing community, who, initially focusing on image analysis, wanted to standardize only the representation of the low-level content features and feature-detection algorithms.

The different technical insights, and the different ways of formulating the challenges presented by MPEG-7, have caused the most difficulty within MPEG-7, as the standard's structure reflects.

References

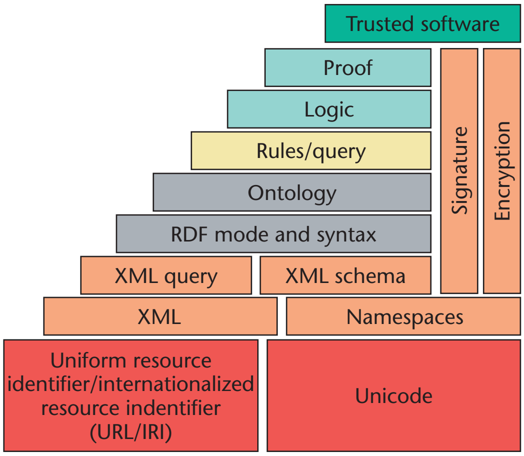
1. W. Cathro, "Metadata: An Overview," *Standards Australia Seminar: Matching Discovery and Recovery*, Aug. 1997, http://www.nla.gov.au/nla/staffpaper/cathro3.html.
2. A.T. Schreiber et al., "Ontology-Based Photo Annotation," *IEEE Intelligent Systems*, May/June 2001, vol. 16, no. 3, pp. 66-74.

Semantic Web

To summarize the current Semantic Web, we use Tim Berners-Lee's (in)famous "layer cake," depicted in Figure 1 (next page), because it depicts the Semantic Web's key components and provides an intuitive perspective on the components' layering. (Many have criticized the figure as not clarifying what it means to stack one language layer on top of another or the syntactic and semantic implications of this stacking model.[8])

The top "trust" layer depicts the Semantic Web's ultimate goal: Machines should be able to not only find and use relevant information, but also to assess the extent to which the information found is both accurate and trustworthy. To reach this level of sophistication, systems perform more complex tasks by increasing the num-

Figure 1. Semantic Web layers, as envisioned by Tim Berners-Lee and presented during a talk at XML 2001 (see http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html). Image courtesy of the W3C.



ber of cooperating layers of languages and processing tools.

Uniform resource identifiers and Unicode. The entire Web pyramid is still based on the naming scheme provided by the URI. Although many overlook the URI's importance, it is, to some extent, the Web's defining characteristic. Anything that wants to be part of the Web needs a URI, and anything with a URI is by definition part of the Web. This doesn't imply that a resource must be available over the Internet to be part of the Web. In addition, although using fragment identifiers with the URI to indicate that the URI addresses a specific resource fragment (instead of the entire resource), the semantics of these fragment identifiers are media dependent and not defined by the URI specification.[9] For example, when a URI points to an HTML page, HTML indicates that the fragment identifier is pointing to the anchor element with that name. For XML documents, XPointer (http://www.w3.org/TR/xptr-framework) provides a framework for defining fragment identifier semantics. Because for many multimedia document types the fragment identifier semantics is still undefined, hyperlinking into them or attaching metadata to specific portions of a resource is difficult.

The Unicode standard[10] is the bottom layer's other ingredient. Whereas earlier versions of HTML had a Western European bias, allowing only the ISO Latin-1 character set, the current Web infrastructure supports a wide variety of languages by allowing the full range of Unicode characters.

XML. On top of the URI/Unicode layer is the XML-based "document Web." In addition to XML, this layer includes XML schema (http://www.w3.org/TR/xmlschema-0) and XML namespaces (http://www.w3.org/TR/REC-xml-names). We can also classify other XML-related languages, such as XPath (http://www.w3.org/TR/xpath), XPointer, and XLink (http://www.w3.org/TR/xlink), as part of this layer.

The current Web uses the syntactic rules specified by this layer, on top of which it defines self-describing document languages such as XHTML (http://www.w3.org/TR/xhtml1), Synchronized Multimedia Integration Language (SMIL, http://www.w3.org/TR/smil20), and scalable vector graphics (SVG, http://www.w3.org/TR/SVG). These documents are *self-describing* because they have a text-based syntax with markup that's meaningful to human readers. For example, a human reader could interpret the content of a well-written HTML document just by looking at its raw encoding (compare this with most proprietary binary document formats, whose content becomes lost when the associated applications are no longer available).

Resource description framework. As outlined earlier, no absolute boundary between data and metadata exists. On a practical level, however, metadata needs languages and tools designed to facilitate its encoding and processing. This need motivated RDF's development (http://www.w3.org/TR/REC-rdf-syntax). Built as a layer on top of XML, RDF's design allows more specific metadata languages to be built on top of it. RDF's fundamental building block is the *statement,* used to define a specific resource's property. Each property's value is either another resource (specified by a URI) or a *literal* (a string encoded conforming to XML-specified syntax rules). The property's name can be any (namespace-qualified) XML name. In short, each RDF statement is a *triple*, consisting of the resource being described, the property's name, and the property's value. RDF triples can be linked, chained, and nested. Together, they allow the creation of arbitrary graph structures.

Although RDF doesn't cater specifically to multimedia applications, it's also not specific to text. In an RDF statement, both the property subject and value could refer to a multimedia Web resource.

RDF schema. Although RDF lets users encode complex metadata graphs, it doesn't associate specific semantics to the graphs, other than the roles implied by the triple. Thus, a user can make RDF statements without committing to a specific ontology. However, just as defining the names of the elements and attributes that might be used and their possible syntactic combinations is often useful in specific XML contexts, in RDF, it's often useful to define the set of semantic concepts a given application should recognize, as well as the basic semantic relations among the concepts. RDF schema (http://www.w3.org/TR/rdf-schema) defines a language on top of RDF that supports the definition process. By predefining a small RDF vocabulary for defining other RDF vocabularies, we can use RDF schema to specify the vocabulary for a particular application domain.

Although it doesn't define a full ontology language, RDF schema extends the RDF data model by allowing hierarchical organization of properties—that is, a user can declare one property to be a `subPropertyOf` another property. Users can also group resources belonging to the same type in a `Class`.

RDF schema structures give sufficient information to allow basic queries regarding the concepts' semantics and their relationships in the application domain. For example, a user could select all intranet documents about mobile phone models from a specific year. Such queries are much harder when they must be phrased in terms of the XML or HTML syntax structure.

Although the more classical metadata applications for which RDF was initially developed might have less of a need for these formal semantics and inference models, they are critical ingredients for the upper layers of the Semantic Web (for example, the logic, proof, and trust layers in Figure 1). At the time of this writing, developers are working on a formal semantics for both RDF and RDF Schema (http://www.w3.org/TR/rdf-mt).

Ontology languages: OWL and beyond. As we write this, the W3C is developing the Web Ontology Language (OWL, http://www.w3.org/TR/owl-ref). OWL's development draws on the experience and lessons learned during the development of earlier Web-oriented ontology languages, most notably DARPA Agent Markup Language + ontology inference layer (DAML+OIL, http://www.daml.org/2001/03/reference.html). DAML+OIL, in turn, draws heavily on one of the major results of the European On-To-Knowledge project: the ontology inference layer (http://www.ontoknowledge.org/oil/TR/oil.long.html) and the associated Ontology Interchange Language, both known under the acronym OIL.

OIL combines the efficient reasoning support and formal semantics of description logics, rich modeling primitives commonly provided by frame languages, and a standard serialization syntax based on XML, RDF, and RDF schemes. European and American researchers, in the context of the DAML project, further developed the language, which became DAML+OIL. The development of OIL and DAML+OIL highlighted the need for formal semantics to provide adequate tool support. The OWL specification is distributed over several documents, one of which is devoted to its semantics.

MPEG framework

ISO's Moving Pictures Expert Group develops standards for coded representation of digital audio and video. The MPEG standard provides a framework for interoperable multimedia content-delivery services. The extensible MPEG-4 textual format (XMT), multimedia content description interface (MPEG-7), and MPEG-21 multimedia framework are important standardization activities with respect to semantic representation.

MPEG-4 XMT. MPEG-4[11] is ISO's standard for interactive multimedia on the Web. XMT[12] gives content authors a textual syntax for the MPEG-4 binary format for scenes (BIFS), letting them exchange content with other authors, tools, or service providers. XMT is an XML-based abstraction of the object descriptor framework for BIFS animations. XMT respects existing practices for authoring content in formats such as SMIL, HTML, or extensible 3D by letting a SMIL player, a Virtual Reality Modeling Language player, and an MPEG player exchange formats using relevant language representations such as XML schema, MPEG-7 document-description language (DDL), and VRML grammar. In short, XMT serves as a unifying framework for representing multimedia content where otherwise fragmented technologies are integrated through the textual format's interoperability.

MPEG-7. MPEG-7[13] aims to standardize description of AV data content in multimedia environments. It provides descriptions of multimedia content on varying complexity levels to let users search, browse, filter, or interpret content using search engines, filter agents, or any other program. MPEG-7 offers a set of AV description tools in the form of descriptors (Ds) and description schemata (DS) describing the meta-

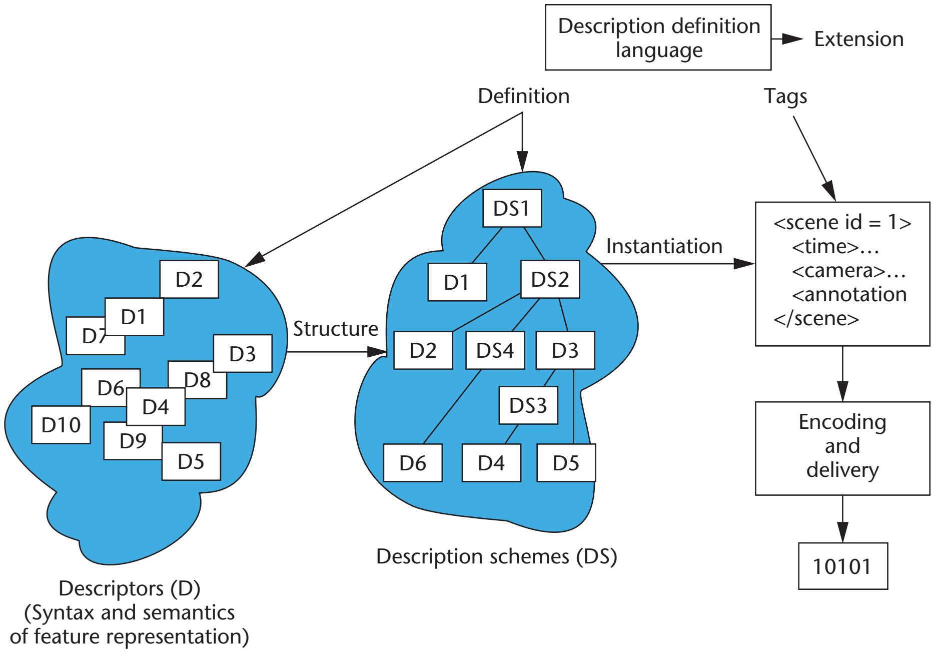


Figure 2. The main MPEG-7 elements.[7] *Content authors can use these structures to create application-specific content descriptions. Used with permission.*

data elements' structure, their relationships, and the constraints a valid MPEG-7 description should adhere to. These structures let users create application-specific content descriptions—that is, a set of instantiated description schemata and their corresponding descriptors. Figure 2 portrays the main MPEG-7 elements.

The standard has eight parts, each responsible for one aspect of the functionality:

- The *systems* component specifies the tools for preparing descriptions for efficient transport and storage, compressing descriptions, and allowing synchronization between content and description. MPEG-7 descriptions can be delivered independently of, or together with, the content they describe.[13]

- The *DDL* specifies the language for defining the standard set of description tools (description schemata, descriptors, and data types), new tools, and the main parser requirements.[13]

- *Visual* consists of schemata and descriptors covering basic visual features such as color, texture, shape, and face recognition. It provides the descriptor syntax and description schemata in normative DDL specifications and the corresponding binary representations. It also provides normative definitions of all the components of the corresponding descriptors and description schemata.[13]

- *Audio* specifies a set of low-level descriptors for audio features (for example, a signal's spectral, parametric, and temporal features) as well as high-level application-specific description tools (such as general sound recognition and indexing schemata used for instrumental timbre, spoken content, audio signature, and melody). It also provides normative definitions of all the components of the corresponding descriptors and description schemata.[13]

- *Multimedia description schemes* (MDS) specify generic description tools pertaining to multimedia, including audio and visual content. MDS covers the basic elements for building a description, the tools for describing content and relating the description to the data, and the tools for describing content on organization, navigation, and interaction levels.[13] The MDS alone forms more than half the complete standard and has its own internal structure, shown in Figure 3.

- *Reference software* provides the software corresponding to the tools defined in the standard (parts 3–5).[7]

- *Conformance* specifies the guidelines and procedures for testing an implementation's conformance to the standard.[7]

- The *extraction and use* component specifies the extraction and use software corresponding to the tools defined in the standard (parts 3–5).

MPEG-7 clearly addresses a broad spectrum of representational problems, from high-level conceptual descriptions of the content and its production to details on low-level features. However, the attempt to provide a highly interoperable standard also creates MPEG-7's fundamental problems, as we'll show in Part 2 of this article.

MPEG-21. MPEG-21's general goal is to describe an open framework that lets users integrate all delivery chain components necessary to generate, use, manipulate, manage, and deliver multimedia content across a wide range of net-

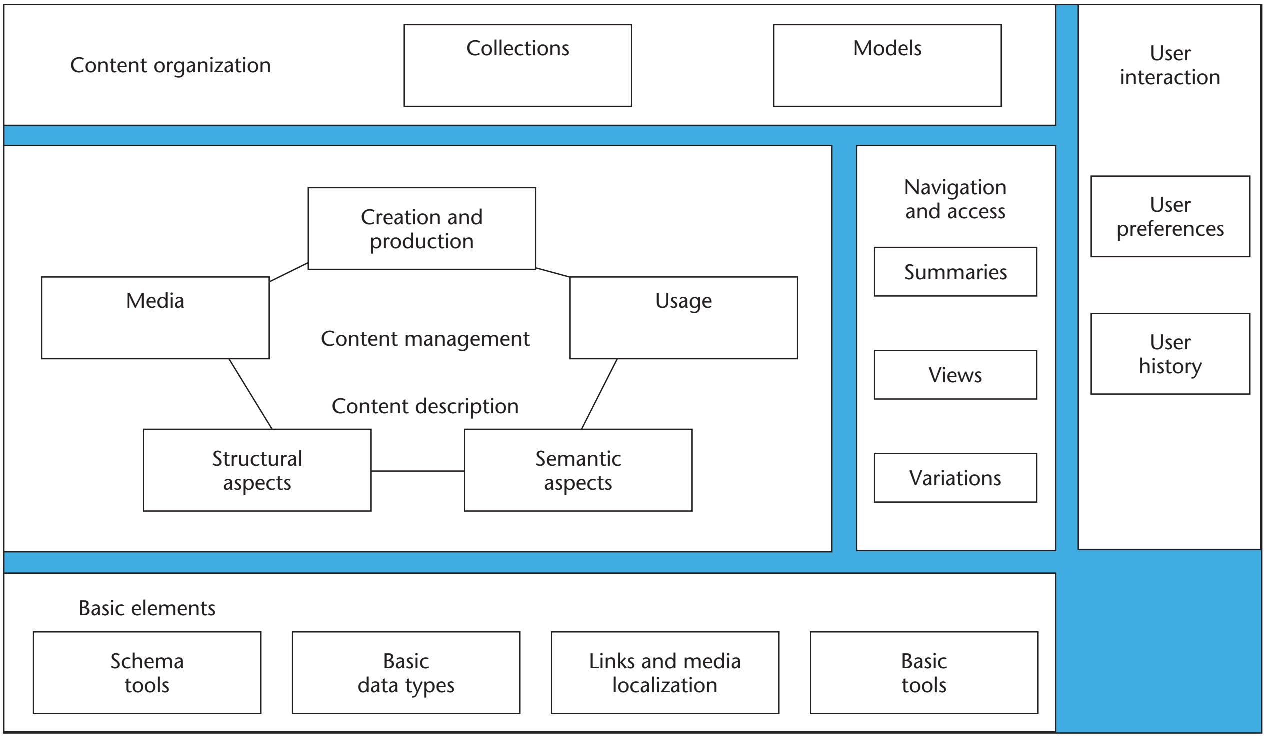


Figure 3. Overall organization of MPEG-7 multimedia description schemes.[7] Used with permission.

works and devices.[14] The MPEG-21 multimedia framework will identify and define the key elements needed to support the multimedia delivery chain, their relationships, and the operations they support. MPEG-21 will also address the necessary framework functionality such as the protocols associated with the interfaces and mechanisms to provide a repository, composition, conformance, and so on.

Key MPEG-21 elements are

- *digital item declaration,* a uniform and flexible abstraction and interoperable schema for declaring digital items;
- *digital item identification and description,* a framework for identifying and describing any entity regardless of its nature, type, or granularity;
- *content handling and usage,* which provides interfaces and protocols enabling creation, manipulation, search, access, storage, delivery, and content reuse across the content distribution and consumption value chain;
- *intellectual property management and protection,* which persistently and reliably manages and protects content across a wide range of networks and devices;
- *terminals and networks,* which provide interoperable and transparent access to content across networks and terminals;
- *content representation*, which determines how the media resources will be represented;
- *event reporting,* the metrics and interfaces that let users understand precisely the performance of all reportable events within the framework.

Some metadata aspects covered in MPEG-21 apply to AV content description (for example, content handling and usage), hence our inclusion of a short overview. Although this issue is relevant, because of space constraints we don't discuss it further.

Metadata issues: Text versus multimedia

High-quality metadata is essential to many multimedia applications. Unfortunately, multimedia metadata comes with several significant problems that apply to metadata in general:

- *Cost*. Obtaining high-quality metadata is expensive and time consuming. Although we can use text analysis and feature extraction to obtain metadata descriptions of some low-level features automatically, most applications depend on higher-level annotations that require human labor. Because human annotation is important and expensive, it must be done right the first time.

- *Subjectivity*. Having humans make annotations also leads to highly subjective results. Even with good tool support, human annotators often interpret documents differently, resulting in inconsistencies within a single document collection. Even worse, annotators often have specific views on content and the context in which it's used. Because the annotations might not be used for many years, the end user's context will likely differ radically from anything the annotators might have imagined.

- *Restrictiveness*. Highly formalized metadata schemata can provide machines with more appropriate information, but human annotators often consider them too restrictive. On the other hand, less restrictive schemata (such as free text fields) often result in subjective and inconsequential terminology to the extent that it has hardly any value for machine processing.

- *Longevity*. Longevity is a problem for many electronic documents, and it might be even worse for their annotations. Designing annotation schemata that are applicable in the short and long term, and are sufficiently specific to be useful within their domain and sufficiently generic to be used across domains, is difficult. Such schemata require flexible tool support for extensions, modifications, version tracking, and so on.

- *Privacy*. Metadata might provide private or security-sensitive information requiring particular care. Examples include medical documents annotated with personal information about the patient, or digital artwork reproductions annotated with the original's insurance value.

- *Standardization*. Annotators' tools often differ from end users' tools. Providing the required interoperability therefore often requires a relatively high degree of standardization both on the syntax level, to ensure that one tool can parse the other's formats, and on the semantic level, to make sure that tools can figure out which shared concepts a different party's terms refer to. In practice, semantic interoperability requires a certain degree of automatic inferencing. Minimally, tools must be able to determine when different terms are equivalent and when a subsumption relation links them.

Several issues specific to the use of metadata in a multimedia context also exist.

First is the problem of associating annotations with multimedia data. Although linking annotations with content might also be an issue for text documents, in practice most metadata applies to either the entire text document, or to a fragment with boundaries that are inherent in the text's structure—for example, metadata that relates to a specific chapter, paragraph, or sentence. (Addressing the metadata's target in text is similar to identifying a link's source and target in hypertext, which is generally regarded as a solved problem.)

For multimedia, it's common to attach metadata to objects in the media stream, such as an object in a video. That metadata might apply to the object's region in any frame featuring the object. Specifying such regions is hard because it's often independent of shot or scene boundaries. Different units of metadata might address different frame ranges, requiring a stratified approach.[15] Even within a specific frame, identifying the target object is rarely trivial. (Addressing the metadata's target in multimedia is similar to identifying a link's source and target in time-based hypermedia, which is still considered an open issue, especially from a standardization viewpoint.)

AV interpretation is another issue specific to multimedia metadata. A human annotator's subjectivity is often a more serious obstruction when the individual must interpret a nontextual document's semantics. This problem is rooted in the myriad perceptual, cognitive, and cultural codes buried in AV material. The aim of any standardized or proprietary annotation thus is to facilitate metadata generation to allow different views of the same AV material.

Work flow management can be a significant problem in the multimedia production process, which produces a lot of high-quality metadata such as scenarios, scripts, storyboards, and edit decision lists. Additionally, capturing devices capture semantic-loaded low-level features during the production process. Many current digital cameras record a continuous stream of information about the camera's settings (zoom, focus, shutter speed, and so on) along with the video signal. Unfortunately, most of this metadata isn't available in the version the end user receives. The challenge thus lies in controlling metadata flow throughout the production chain and making the relevant parts accessible to the people and applications authorized to use it.

Another issue, repurposing media items into a new, coherent story, is more challenging for multimedia than for text. Repurposing aims to describe alternating contexts, where a media item might play a different rhetorical role in each context.

Data quantity and streaming is another multimedia-specific issue. The sheer bulk of digital multimedia content often makes downloading all the material before playback undesirable, giving rise to streaming content delivery. Similar arguments apply to bulky multimedia metadata that must be delivered in a streaming fashion without disrupting the stream of AV content.

In addition, rapidly changing capture and playback technologies result in multimedia products quickly becoming legacy and unplayable (floppy disks and vinyl records are good examples). The resolution problem, for example the coding of an MPEG1 video, leads us to ask how annotations describing the original production process can improve a multimedia product's total reconstruction or restoration.

Finally, multimedia's more complex production process also makes digital rights management more complex than for text. Several parties (directors, producers, scenario writers, actors, and so on) may exercise their rights on a single media item.

Although we must address these problems before we can realize the vision of a media-aware Semantic Web, Part 2 of this article will focus on problems directly associated with the semantics of nontextual media.

Conclusion

Having overviewed the two main standard activities for the semantic representation of media, we are now in a position to evaluate both. In Part 2, we'll analyze the W3C and ISO approaches in detail with respect to these requirements. We'll also discuss the findings' implications for future actions. **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, and Ontoweb, a thematic network of the European Commission, funded part of the research described here. We thank 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 the development of this article. Finally, we thank the anonymous reviewers of *IEEE MultiMedia* for their detailed and valuable comments.

References

1. A. Del Bimbo, *Visual Information Retrieval,* Morgan Kaufmann, 1999.
2. F. Nack, "The Future of Media Computing—From Ontology-Based Semiotics to Computational Intelligence," *Media Computing—Computational Media Aesthetics*, C. Dorai and S. Venkatesh, eds., Kluwer, 2002, pp. 159-196.
3. J. Ryu, Y. Sohn, and M. Kim, "MPEG-7 Metadata Authoring Tool," *Proc. 10th ACM Int'l Conf. Multimedia*, ACM Press, 2002, pp. 267-270; http://www.acm.org/sigs/sigmm/MM2000/ep/rehm/.
4. T. Berners-Lee, *Weaving the Web,* Orion Business, 1999.
5. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific Am.*, May 2001; http://www.sciam.com/2001/0501issue/0501berners-lee.html.
6. *MPEG-7 Requirements Document* (v. 15), ISO/IEC Std. JTC1/SC29/WG11/N4317, ISO/IEC, 2001.
7. *Overview of the MPEG-7 Standard* (v. 8), ISO/IEC Std. JTC1/SC29/WG11/N4980, ISO/IEC, 2002; http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.
8. 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/.
9. T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform Resource Identifiers (URI): Generic Syntax," IETF RFC 2396, 1998; http://www.ietf.org/rfc/rfc2396.txt.
10. Unicode Consortium, *The Unicode Standard, v.3.0,* Addison-Wesley Developers Press, 2000.
11. *MPEG-4 Overview* (v. 18, Singapore version), ISO/IEC Std. JTC1/SC29/WG11 N4030, ISO/IEC, 2001.
12. M. Kim, S. Wood, and L.-T. Cheok, "Extensible MPEG-4 Textual Format (XMT)," *Proc. Int'l Workshop on Standards, Interoperability and Practices*

in conjunction with the 8th ACM Multimedia Conf., ACM Press, 2000; http://www.acm.org/sigs/sigmm/MM2000/ep/michelle/.

13. *Information Technology—Multimedia Content Description Interface, Parts 1–5:* ISO/IEC Std. 15938-1 to *Systems*, ISO/IEC Std. 15938-5/FDIS, ISO/IEC, 2001.
14. *MPEG-21 Overview* (v. 5), ISO/IEC Std. JTC1/SC29/WG11/N5231, ISO/IEC, 2002; http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm.
15. A. Parkes, "Settings and the Settings Structure: The Description and Automated Propagation of Networks for Perusing Videodisk Image States," *Proc. 12th Int'l Conf. Research and Development in Information Retrieval* (SIGIR), N.J. Belkinand and C.J. van Rijsbergen, eds., ACM Press, 1989, pp. 229-238.

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.

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*.

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.

For further information on this or any other computing topic, please visit our Digital Library at http://www.computer.org/publications/dlib.