Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

INS

Information Systems

INformation Systems

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

Jacco van Ossenbruggen, Frank Nack, Lynda Hardman

REPORT INS-E0308 DECEMBER 3, 2003

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

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

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

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

ISSN 1386-3681

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

ABSTRACT
This article discusses the state of the art in metadata for audio-visual media in large semantic networks, such as the Semantic Web. Our discussion is predominantly motivated by the two most widely known approaches towards machine-processable and semantic-based content description, namely the Semantic Web activity of the W3C and ISO's efforts in the direction of complex media content modeling, in particular the Multimedia Content Description Interface (MPEG-7). We explain that the conceptual ideas and technologies discussed in both approaches are essential for the next step in multimedia development. Unfortunately, there are still many practical obstacles that block their widespread use for providing multimedia metadata on the Web. Based on a scenario to explain our vision of a media-aware Semantic Web, we derive in Part I a number of problems regarding the semantic content description of media units. We then discuss the multimedia production chain, in particular emphasizing the role of progressive metadata production. As a result we distill a set of media-based metadata production requirements and show how current media production environments fail to address these. We then introduce those parts of the W3C and ISO standardization works that are relevant to our discussion. In Part II of this article, we analyze their abilities to define structures for describing media semantics, discuss syntactic and semantic problems, ontological problems for media semantics, and the problems of applying the theoretical concepts to real world problems. Part II concludes with implications of the findings for future action with respect to the actions the community should take.

1998 ACM Computing Classification System: H.3.1,H.3.7,H.5.1,H.5.4.,I.7.2.
Keywords and Phrases: Semantic Web, metadata production, multimedia production process, XML, XML Schema, RDF, RDF Schema, MPEG-4, MPEG-7, MPEG-21
Note: This work was carried out under the Dutch national NWO/NASH and ToKeN2000 CHIME projects.

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

Jacco van Ossenbruggen, Frank Nack, Lynda Hardman

Abstract

This article discusses the state of the art in metadata for audio-visual media in large semantic networks, such as the Semantic Web. Our discussion is predominantly motivated by the two most widely known approaches towards machine-processable and semantic-based content description, namely the Semantic Web activity of the W3C and ISO's efforts in the direction of complex media content modeling, in particular the Multimedia Content Description Interface (MPEG-7).

We explain that the conceptual ideas and technologies discussed in both approaches are essential for the next step in multimedia development. Unfortunately, there are still many practical obstacles that block their widespread use for providing multimedia metadata on the Web.

Based on a scenario to explain our vision of a media-aware Semantic Web, we derive in part I a number of problems regarding the semantic content description of media units. We then discuss the multimedia production chain, in particular emphasizing the role of progressive metadata production. As a result we distill a set of media-based metadata production requirements and show how current media production environments fail to address these. We then introduce those parts of the W3C and ISO standardization works that are relevant to our discussion.

In part II of this article, we analyze their abilities to define structures for describing media semantics, discuss syntactic and semantic problems, ontological problems for media semantics, and the problems of applying the theoretical concepts to real world problems. Part II concludes with implications of the findings for future action with respect to the actions 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 Example scenario

Imagine, five years from now, you are the head of a lab that develops mobile communication devices and you would like to develop the company's new product line. You first need to 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 "strictly speaking, we have no budget for this").

Your multimedia presentation authoring tool (i.e. 2009's integrated successor of the PowerPoint/Director family) finds relevant media assets (including product related texts, pie charts and still and moving images) on the corporate network. Based on these assets, and their associated metadata, it generates a first preview of the presentation. Though the automatically generated story line is coherent and succeeds in conveying many of the important semantic relations among the retrieved media items, you are, however, not content with the storyline's progression and the lack of tension buildup.

You fire up the tool's storyboard editor and start to improve upon the automatically generated storyline. Your edited version of the presentation now includes some scenes that are intended to convey the lifestyle of the new product's target audience. A search on the corporate network returns no suitable footage, nor a fitting soundtrack. You are reluctant to start a search on the public peer-to-peer file sharing network: finding appropriate material won't be a problem, but dealing with the copyright issues is likely to involve more time and money than you can afford. Because you have no other option, you give it a try anyway.

Quickly, the p2p search tool shows you some (scaled-down quality) 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 that has reasonable licensing costs when the material is not used in public. You select a few clips and order your digital rights managment agent to deal with the legal issues and pay the required fees (all in anonymous mode: you do not want your competitors to be able to trace these transactions). At the end of the day, you have a presentation of sufficient quality to use for tomorrow's board meeting.

The scenario above is challenging, in particular because the tools involved need to operate, to some extent, on the semantics of the media items involved. This semantic understanding of media is required for both major tasks to be addressed during multimedia production: namely for the more traditional retrieval tasks (finding relevant media items) and the more innovative tasks, such as generating a coherent sto-

ryline from a set of media items. Additonally, semantics implies context and hence an understanding of the technical and social context is required, including information about copyrights, provenance, etc.

To be able to build tools that are aware of the semantics of both the content and the context of multimedia, we need to make these semantics explicit. Making semantics explicit is usually achieved through *metadata*. Thus, the goal is to find te appropriate ways to annotate audio-visual information units with metadata in such a way that humans and machines can handle the information. The machine readable aspect is of importance as we aim to provide tools thateither support the multimedia production process or that generate multimedia generations autonomously.

2 Metadata in the multimedia production chain

Audio-visual (AV) media production, such as the business presentation from the scenario, is a complex process. Metadata could improve the production process by making information explicit that is implicit in the AV content. An obvious approach is to incrementally store relevant metadata during the production process, and make it accessible to all the tools involved in the process.

In the remainder of this section, we first discuss the requirements for this approach and sketch the impact it would have on the production environments.

2.1 Media metadata production: requirements

Although media production is often a rather iterative and organic process, for convenience it is traditionally divided into three stages:

- **preproduction**, which is concerned with determining the main ideas and that form the core of the production (scripting, storyboarding, etc.),
- **production**, where the main task is the acquisition of media material (shooting, sound recording), and,
- **post-production**, which is oriented towards editorial decisions based on reviewing the material, editing, sound mixing, presenting, and archiving.

It must be emphasized that the different production stages are highly interdependent. Especially if each step would not only produce and modify data, but also, and even on a larger scale, produce and modify metadata. A tool used during preproduction could not only produce or update scripts, but also export the notes about the rationale underlying certain script modifications. Shooting could produce raw footage, but also explicit descriptions of activities on the set. Post-production tools could produce edited and mixed material, but also production schedules, editing lists with decision descriptions and other organizational information. In each step, changes to the metadata could directly impact material produced in other stages, and these dependencies could even be made explicit and supported accross the different tools involved.

This type of metadata is also important because it represents the progression through the various alterations on a technical, structural, and a description level. It defines the original context of the individual media items used. Today this type of information is often lost after the production is finished, and needs to be re-engineered when needed later.

This is precisely the problem with traditional approaches for metadata production. They primarily address the end product, characterizing audio-visual information on a conceptual (keyword) and on a perceptual level by using objective measurements based on image or sound processing, pattern recognition, etc. [5, 26, 18, 15, 17]. What these retrospective approaches naturally miss is important cognitive, content and context based information, describing the intermediate stages in the production and the decisions taken.

With the advance of DVDs that include extra (meta) material and the many "making of . . . " productions, high quality metadata has already become an economic asset. In most cases, however, production of all this extra information would still remain unrealistic if it would necessitate manual annotation — such an expensive endeavor would normally not be covered by the production or archival budget. Instead, we need high level tool support that is integrated into the production environment and does not hinder the creative and improvisatory processes that are so important in media production.

Note that within such an environment, the produced media item, on a micro (e.g. shot) and macro-level (e.g. complete business presentation) may still be of a linear nature. The overall collection of material, however, including all the intermediate physical AV data, as well as the creative decisions made during the media production process and other contextual information, would establish a non-linear and complex semantic network. It is important to stress, though, that this network just reflects the purposes and intentions, or in other words the context, of the particular production.

The idea of saving the complete production process is not new, though the consequent implementation in a digital environment remains difficult. It requires standardized representational structures that reflect the constant changes the AV material undergoes during its production, but also dynamic semantic structures that allow the representation of conceptual developments over time.

In addition, the deeper impact of digital media is to redefine the traditional forms of media, to blur the boundaries between traditional production steps and to alter the information flow from producers to consumers. Consequently, we have to introduce an additional step into the production process:

- metaproduction, which comprises processes such as restructuring, representing, resequencing, repurposing and redistributing media.

The scenario as described in section 1 is a first-class example of this type of production, as most of the material to be used was produced beforehand, for a different purpose and in a different context.

The outcome of any metaproduction process is an extension of an existing semantic network: it provides additional production information and describes a different context of use for existing material. Note that a piece of metadata can change its role and turn into a piece of media that needs to be described. For example, imagine a film theoretician who would like to demonstrate the referential quality within the work of a particular director. One could use the original sequence of the referenced film and link it together with the sequence that acts as the referrer. The latter media item acts in this relation as the metadata (typical examples for such references are demonstrated by the station scene from De Palma's 'Untouchables' and the arrest scene in Gilliam's 'Brazil', which both refer to the 'Odessa steps' scene in Eisenstein's 'Battleship Potemkin').

Summarizing the above discussion, the following requirements for a media-aware semantic network are:

- As a given component of media exists independently of its use in any given production, sufficient linking mechanisms are required to establish context.
- Annotation and production are basically different sides of the same coin, as a media item can play various roles (data and metadata) depending on the context it is used in. Therefore, flexible description schemata need to be developed that reflect these roles.
- Even within a single production, annotations are necessarily imperfect, incomplete, and preliminary because they accompany and document the dynamic progress of understanding a concept. This usually opens up questions of aesthetics and subjective interpretation. Thus, semantic, episodic, and technical representation structures that are able to change and grow are required. There is no such thing as a single and all-inclusive content description.
- When metadata is to be reused across multiple productions, one needs even more expressive mechanisms for encoding metadata and makin it accessible in a controlled way. There is no such thing as a single and all-inclusive content description.
- The generation of semantic annotations can best be achieved during the media production process, which requires the support of the activities associated within the production phases.

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

2.2 Media production environments

On a future media-aware Semantic Web, one expects a great variety of media to be constantly generated, manipulated, analyzed, and commented on. Such a Web could, however, 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. At the same time, these tools need to integrate with the (still mainly text-oriented) environment of current Semantic Web technology.

However, today's media production is mainly oriented towards one-time design and production. This means that important sources of metadata are lost after the production is finished, as exemplified in tools for multimedia production, such as Director, Premiere, Photoshop, Flash, FrontPage, PowerPoint and others. These tools work with proprietary data structures which makes it nearly impossible to use the internal content representation structures outside the application or for a different purpose. 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 while there are more possibilities than ever to assist in the creative development and production processes of media, we still lack environments which serve as an integrated information space for use in distributed productions, research, restructuring (e.g. by software agents) or in direct access and navigation by the audience.

On the other hand, there are first attempts in research as well as in industry to demonstrate how extra semantics could be added automatically or semi-automatically to audiovisual material during production or metaproduction without interfering with established workflows [20, 25, 29, 31]. The advantage of these tools is that they all use standardized XML-based description mechanisms and follow the paradigm of intelligent tools that rely on the existence of supportive descriptional structures.

All of these prototypes suffer, in some way or another, from their experimental nature with respect to real applicability and scaling. They are not more than a small first step towards the intelligent use and reuse of media production

material. Nevertheless, these prototypical examples provide insight into the generation of interactive media documents in particular, and research into media representation in general. The most interesting aspect of these works is their potential to cooperate when common representation structures become available and implemented.

Based on the previous discussions and our arguments on problematic aspects of metadata and the role of metadata in media production, we are now in the position to provide a more detailed look at the two most relevant common formats for machine-processable and semantic multimedia content description, namely the Semantic Web activity of the W3C and ISO's Multimedia Content Description Interface (MPEG-7).

3 Approaches to metadata: the Semantic Web vs MPEG

Machine-processable content is the main prerequisite for the more intelligent Web services that constitute the "Semantic Web" as envisioned by Tim Berners-Lee and others [1, 2] and the intelligent media applications thought about by the MPEG community [7, 14]. Metadata plays a key role in realizing these visions and the high level technical goals of both communities are very similar: providing a *general* metadata framework. The approaches to provide such a framework, however, differ radically. This section first provides a short historical overview to explain the conceptual roots of both approaches, and then provides a brief introduction to both methodologies.

3.1 Historical background

Metadata-related issues touch the core of all information sciences. Models and technology for processing metadata have been influenced by many communities, in particular, the digital library (DL) community, the knowledge representation (KR) community and the part of the AI community that interprets, manipulates or generates audio-visual media (MM-AI). The Semantic Web, as seen from W3C, can be understand as an attempt to make results of the research in the DL and KR communities applicable to the Web. The prospective of MPEG tries to incorporate aspects from all three communities.

To understand W3C's Semantic Web, one needs to understand the different views of both the DL community and the KR community.

Within the DL community, metadata is, first of all, seen as a way of supporting cataloging and retrieving information in large documents collections. This has resulted in standards that address such issues, most notably the *Dublin Core* [3]. The Dublin Core basically standardizes a set of 15 commonly agreed upon metadata elements of the type that one can expect to find in every library catalog, including title, subject, creator, language, creation date, etc.

The metadata and document-centered focus of the DL community can be contrasted with the information modeling approach of the knowledge representation (KR) community, where the focus is on representing the underlying content rather than describing the document that contains 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 considered a secondary issue.

When it comes to sharing and communicating explicit knowledge, a key concept is the notion of an *ontology*. Within KR, ontologies are often defined as a "specification of a conceptualization", that is, an explicit and commonly agreed-upon definition of the objects and concepts that play a role in a certain domain. These are specified along with the relations among them and the rules that limit the interpretation of the concepts. Given an ontology about a certain domain, parties that need to share and communicate knowledge do this by making an *ontological commitment*: a statement that both people and applications (agents) will use the terminology specified in the ontology according to the specified rules.

Despite the differences between the DL and KR approaches, many applications need elements from both worlds. Ontologies, for example, are often used to control the terminology used in metadata. By making a commitment to a specific ontology, users can be assisted in making annotations in a more systematic and consistent way [27]. In addition, applications may use the "background knowledge" specified by the ontology in addition to the metadata itself. For example, when the metadata of a particular page about a painting only specifies that the painting is painted by Rembrandt van Rijn, a query for "17th-century Dutch masters" will not return the page. When the metadata is combined with an ontology stating that Rembrandt is indeed classified as a 17th-century Dutch master, the page can be returned in response to the query.

The view of MPEG on metadata is similar to that of the W3C, only that for MPEG, a document is typical a complex audio-visual unit and thus the standardization in 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 aspects such as interoperability and globalization of metadata resources and flexibility of data management. For this purpose MPEG-7 had to reconcile the approaches in favor of the different communities. On the one hand, the DL, KR and MM-AI communities stressing the

need for high-level descriptions of audio-visual content. On the other hand, the signal processing community, which had primarily focused on image analysis, saw success in only standardizing 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 and, as will be shown, are reflected in the structure of the standard.

The following section should provide a brief overview of the W3C and ISO approaches towards the description of semantics within media. The goal is to facilitate a better understanding of the basic aspects covered by both worlds.

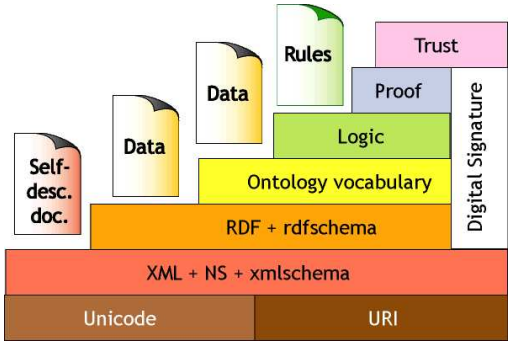

Figure 1: The layers of the Semantic Web envisioned by Tim Berners-Lee, as presented during a talk at XML 2001 (see `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`).

3.2 Metadata on the the Semantic Web

We give a summary of the current Semantic Web by using Tim Berners-Lee's (in)famous[1] "layer cake" depicted in Figure 1, because it depicts the key components of the Semantic Web and provides an intuitive perspective on the layering of these components. The "trust" layer at the top of the figure depicts the ultimate goal of the Semantic Web: machines should be able to not only find and use relevant information, but they should also be able to assess to what extent information found is accurate and can be trusted. In order to reach this level of sophistication, more complex tasks are carried out by increasing the number of cooperating layers of languages and processing tools. We will give a short summary of each layer, starting from the bottom layer.

[1]Note the figure has often been criticized because it is unclear what it actually means to stack a language layer on top of the other, and what the syntactic and semantic implications of this stacking model are [23].

3.2.1 URIs and Unicode

The basis of the whole Web pyramid is still the uniform naming scheme provided by the concept of the URI. The importance of the URI is often overlooked, but is, to a certain extent, the defining characteristic of the Web. Anything that wants to be part of the Web needs to have a URI, and, *vice versa*, anything that has a URI is by definition part of the Web. Note that this does *not* imply that something needs to be available electronically over the Internet to be part of the Web.

Also note that, while it is common to use *fragment identifiers* in conjunction with the URI to indicate that the URI addresses a specific fragment of resource (instead of the entire resource), the semantics of these fragment identifiers are media dependent and not defined by the URI specification[2]. For example, when the URI points to an HTML page, HTML defines that the fragment identifier points to the anchor element with that name. For XML documents, XPointer[3] provides a framework for defining fragment identifier semantics. For many multimedia document types, however, the semantics of fragment identifiers is still undefined, which makes it hard to hyperlink into them, or to attach metadata to specific portions of a resource.

The other ingredient of the bottom layer is the Unicode standard [30]. While earlier versions of HTML used to have a Western-European bias by only allowing the ISO Latin-1 character set, the current Web infrastructure now supports a wide variety of other languages by allowing the full range of characters specified by Unicode.

3.2.2 XML

On top of the URI/Unicode layer, the current, XML-based, "document web" is built. This layer includes not only XML itself, but also XML schema[4] and XML namespaces[5]. Other XML related languages, such as XPath[6], XPointer and XLink[7] could also be classified as part of this layer. The current Web uses the syntactic rules specified by this layer, on top of which *self-describing* document languages such as XHTML[8], SMIL[9] and SVG[10] are defined. These documents are called self-describing because they have a text-based syntax with markup that is meaningful to human readers. For example, just by looking at its raw encoding, the content of a well-written HTML document could be interpreted by a human reader even when there is no HTML displaying software

[2]`http://www.ietf.org/rfc/rfc2396.txt`
[3]`http://www.w3.org/TR/xptr-framework/`
[4]`http://www.w3.org/TR/xmlschema-0/`
[5]`http://www.w3.org/TR/REC-xml-names`
[6]`http://www.w3.org/TR/xpath`
[7]`http://www.w3.org/TR/xlink`
[8]`http://www.w3.org/TR/xhtml1`
[9]`http://www.w3.org/TR/smil20`
[10]`http://www.w3.org/TR/SVG/`

available (compare this with most proprietary binary document formats whose content will become lost when the associated applications are no longer available).

3.2.3 RDF

As outlined in the discussion on metadata production, there is no absolute boundary between data and metadata. On a practical level, however, metadata benefits from having languages and tools that are especially designed to facilitate the encoding and processing of metadata. This is the motivation behind the development of RDF (Resource Description Framework)[11]. Built as a layer on top of XML, RDF itself was also designed from the beginning as a layer on top of which more specific metadata languages could be built.

The fundamental building block of RDF is the *statement* that is used to define a *property* of a specific *resource*. The *value* of each property is either another resource (specified by a URI) or a literal (a string encoded conforming to syntax rules specified by XML). The *name* of a property can be any (namespace qualified) XML name. In short, each RDF statement is basically a *triple*, consisting of the resource being described, the name of a certain property and the value of this property.

RDF triples can be linked, chained and nested. Resources can be *linked* because they can be the subject of multiple triples as well as being reused as the value of multiple triples. *Chains* can be formed by using the value of the first triple as the object of the following triple. Triples can be (arbitrarily) *nested*, so that any triple can be treated as an object (this is termed *reification*) and reused as a resource. Together, these combinations allow the creation of arbitrary graph structures.

Note that while RDF does not cater especially for multimedia applications, it is, in itself, not specific to text. In an RDF statement, both the subject and the value of the property could refer to a multimedia resource on the Web.

3.2.4 RDF Schema

While RDF allows complex graphs of metadata to be encoded, RDF itself does not associate any specific semantics to these graphs other than the three roles implied by subject/predicate/value triple.

However, just as it is often useful in a specific XML context to define the element and attribute names that may be used and in what syntactic combination, in RDF it is often useful to define, for a specific application, what set of semantic concepts the application is supposed to recognize, and what basic semantic relations hold among those concepts. RDF Schema[12] defines a language on top of RDF that supports this. By predefining a small RDF vocabulary for defining other RDF vocabularies, one can use RDF Schema to specify the vocabulary used in a particular application domain. RDF Schema extends the RDF datamodel by allowing organization of properties in a hierarchical fashion, that is, one can declare one property to be a `subPropertyOf` another property. In addition, one can group resources that belong to the same type in a `Class`.

RDF schema structures give sufficient information to allow basic queries in terms of the semantics of the concepts and their relationships in the application domain. For example, one could select all paintings that are painted by a specific painter. Such queries are much harder when they have to be phrased in terms of the XML syntax structure used to encode the information.

While the need for formal semantics and inference models may be less urgent for the more classical metadata applications for which RDF was initially developed, they are critical ingredients for the upper layers of the Semantic Web (e.g. the logic, proof and trust layers in Figure 1). At the time of writing, such a formal semantics[13] is being developed for both RDF and RDF Schema.

3.2.5 Ontology languages: OWL and beyond

Ontologies are used to explicitly specify a set of (domain-specific) concepts and the relations among them. While ontologies are not new in knowledge-based applications, the topic received much wider attention when people began to realize that Web applications will not be able to communicate unless they agree on the terminology used.

At the time of writing, W3C is developing an ontology language[14] for the Web (OWL). The development of OWL draws heavily on the experience and lessons learned during the development of earlier Web-oriented ontology languages, most notably DAML+OIL [32]. DAML+OIL, on it turn, draws heavily upon one of the major results of the European On-To-Knowledge project: the Ontology Inference Layer[15] and the associated Ontology Interchange Language, both known under the acronym *OIL*. OIL combines the efficient reasoning support and formal semantics from Description Logics, rich modeling primitives commonly provided by Frame languages and a standard for syntactical exchange notations based on the languages discussed above. Further work on the language was carried out jointly by both European and American researchers in the context of DARPA's Agent Markup Language project[16], and the language was renamed to *DAML+OIL*.

[11] `http://www.w3.org/TR/REC-rdf-syntax`
[12] `http://www.w3.org/TR/rdf-schema/`
[13] `http://www.w3.org/TR/rdf-mt/`
[14] `http://www.w3.org/TR/owl-ref/`
[15] `http://www.ontoknowledge.org/oil/TR/oil.long.html`
[16] `http://www.daml.org`

One of the lessons learned during the development of both OIL and DAML+OIL was the need for formal semantics to provide adequate tool support. The OWL specification is distributed over several documents, of which one is entirely devoted to the semantics of the language.

3.3 Metadata within the MPEG framework

The Moving Pictures Expert Group within ISO is in charge of developing standards for coded representation of digital audio and video. It aims to provide a framework for interoperable multimedia content-delivery services. Important standardization activities with respect to the representation of semantics are the Extensible MPEG-4 Textual Format (XMT), the Multimedia Content Description Interface (MPEG-7) and the MPEG-21 Multimedia framework, which we summarize briefly in the following sections.

3.3.1 MPEG-4 - XMT

In MPEG-4 [6], the standard for multimedia on the Web, the Extensible MPEG-4 Textual format (XMT) [16] provides content authors with a textual syntax for the MPEG-4 Binary Format for Scenes (BIFS) to exchange their content with other authors, tools, or service providers. XMT is an XML-based abstraction of the object descriptor framework for BIFS animations. Moreover, it respects existing practices for authoring content, such as Synchronized Multimedia Integration Language (SMIL), HTML, or Extensible 3D by allowing the interchange of the format between a SMIL player, a Virtual Reality Modeling Language player, and an MPEG player. It does this using the relevant language representations such as XML Schema, MPEG-7 DDL, and VRML grammar. In short, XMT serves as a unifying framework for representing multimedia content where otherwise fragmented technologies are integrated and the interoperability of the textual format between them is bridged.

3.3.2 MPEG-7

The goal of MPEG-7 [8, 9, 10, 11, 12] is to provide a standardized means of describing audiovisual data content in multimedia environments. Its scope is to facilitate the description of content of multimedia data, so that this data can be searched for, browsed, filtered or interpreted either by search engines, filter agents, or any other program.

MPEG-7 offers a set of audiovisual description tools in the form of descriptors (Ds) and description schemata (DS) describing the structure of the metadata elements, their relationships and the constraints a valid MPEG-7 description should adhere to. These structures form the basis for users to create application specific content descriptions, i.e. a set of instantiated description schemata and their corresponding descriptors. Figure 2 portrays the main MPEG-7 elements.

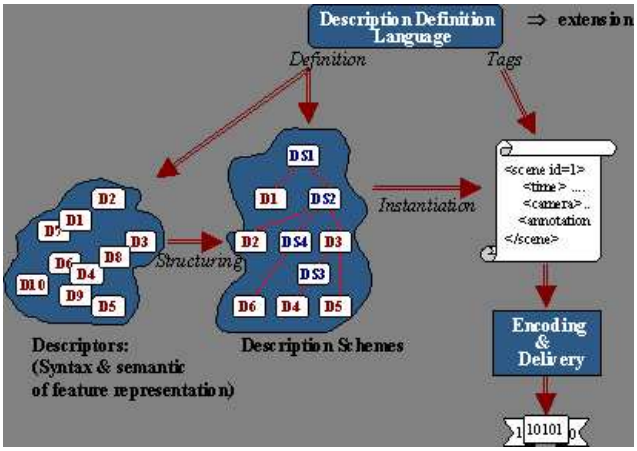

Figure 2: The main MPEG-7 elements.

The standard is organized in 8 parts, each responsible for a particular aspect of the functionality:

Systems specifies the tools for preparing descriptions for efficient transport and storage, compressing descriptions, and allowing synchronization between content and description. It is important to mention that MPEG-7 descriptions may be delivered independently of, or together with, the content they describe [8].

The Description Definition Language (DDL) specifies the language for defining the standard set of description tools (Description schemata (DS), descriptors (Ds), and datatypes) and for defining new description tools. The main parser requirements are defined here [9]. Note that additional essential datatypes are defined in the parts Audio, Video and, in particular, the MDS (see below).

Visual consists of structures and descriptors that cover basic visual features, such as color, texture, shape, motion, localization, and face recognition. The syntax of the descriptors and description schemata is provided in normative DDL specifications and the corresponding binary representations. Moreover, normative definitions of the semantics of all the components of the corresponding descriptors and description schemata are provided [10].

Audio specifies a set of low-level descriptors for audio features (e.g., spectral, parametric, and temporal features of a signal), and high-level description tools that are more specific to a set of applications. Those high-level tools include general sound recognition and indexing schemata, such as for instrumental timbre, spoken content, audio signature and melody. Moreover,

normative definitions of the semantics of all the components of the corresponding descriptors and description schemata are provided [11].

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

- the basic elements for building a description (this section also defines additional datatypes used in the visual and audio part, which are not covered by the DDL datatype definitions),
- the tools to describe content and relate the description to the data and
- the tools to describe content on organization, navigation and interaction level [12].

As the MDS is the largest part of the standard, Figure 3 displays the overall organization structure of the MDS.

Reference Software provides reference software to the standard [14].

Conformance specifies the guidelines and procedures for testing conformance of implementations of the standard [14].

Extraction and use specifies the guidelines and procedures for testing conformance of implementations to the standard [14].

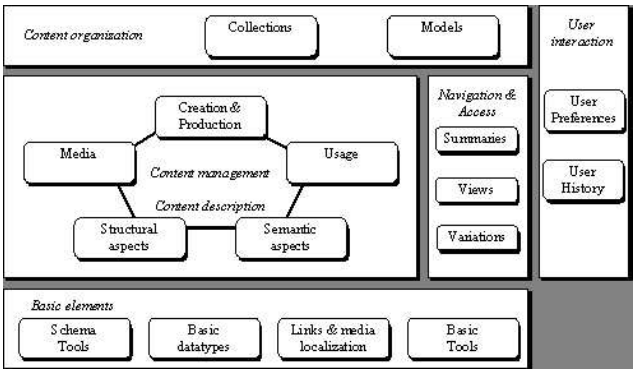

Figure 3: Overall organization of MPEG-7 MDS.

Clearly, the standard addresses a broad spectrum of representational problems, from high-level conceptual descriptions of the content itself and its production down to details on a low-level feature level. However, the attempt of providing a highly interoperable standard also establishes the fundamental problems in MPEG-7, as will be shown in the detailed discussion in part II of this article.

3.3.3 MPEG-21

The general goal of MPEG-21 [13] activities is to describe an open framework which allows the integration of all components of a delivery chain necessary to generate, use, manipulate, manage, and deliver multimedia content across a wide range of networks and devices.

The MPEG-21 multimedia framework will identify and define the key elements needed to support the multimedia delivery chain, the relationships between and the operations supported by them. Within the parts of MPEG-21, MPEG will elaborate the elements by defining the syntax and semantics of their characteristics, such as interfaces to the elements. 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, etc. The seven key elements defined in MPEG-21 are:

- Digital Item Declaration (a uniform and flexible abstraction and interoperable schema for declaring Digital Items);
- Digital Item Identification and Description (a framework for identification and description of any entity regardless of its nature, type or granularity);
- Content Handling and Usage (provide interfaces and protocols that enable creation, manipulation, search, access, storage, delivery, and (re)use of content across the content distribution and consumption value chain);
- Intellectual Property Management and Protection (the means to enable content to be persistently and reliably managed and protected across a wide range of networks and devices);
- Terminals and Networks (the ability to provide interoperable and transparent access to content across networks and terminals);
- Content Representation (how the media resources are represented);
- Event Reporting (the metrics and interfaces that enable Users to understand precisely the performance of all reportable events within the framework).

Some of the metadata aspects covered in MPEG-21 are specifically interesting for audio-visual content description, for example the content handling and usage, and that is why we provided the short overview here. As already outlined in section 4.2 we will, however, not analyze this part of the standard in our ongoing discussion, as this would expand the paper excessively.

Having provided an overview on the two main standard activities for the semantic representation of media, we are

now in the position to evaluate both. Part II of the article analyzes both approaches in detail with respect to the requirements outlined in 2.1. The aim is to identify the strength and weaknesses, as we are mainly interested in the advances on and remaining problems of representing essential conceptual aspects of a multimedia unit. Part II also concludes with implications of the findings for future action with respect to the actions the community should take.

4 Conclusion

4.1 General problems with metadata

Though the general problems with both concepts, namely metadata and annotation, are discussed in great detail in the literature, it seems to be useful to briefly summarize them seperately to establish the basis for our ongoing investigation on the state of the art in metadata for audio-visual media.

High quality metadata is essential for supporting many multimedia applications, including those sketched in our business presentation scenario. Unfortunately, multimedia metadata comes with a number of significant problems that apply to metadata in general.

Costs Obtaining high quality metadata is expensive and time consuming. Although text analysis and feature extraction can be used to obtain metadata descriptions of some low level features automatically, most applications depend on higher level annotations that still require human labour. Because human annotation is both important and expensive, it is crucial that it is done "right" the first time: most organizations simple cannot afford a second round of annotation when it turns out that first round did not yield the desired results.

Subjectivity Having humans make annotations is not only expensive and time consuming, the results are also highly subjective. Even with good tool support, documents are often interpreted differently by different human annotators, resulting in inconsistencies within a single document collection. Even worse, annotators often have a specific view on content and in what context it is supposed to be used. When the annotations are actually used, possibly many years later, the end-user's context is likely to differ radically from everything the annotators could imagine at annotation time.

Restrictiveness Highly formalized metadata schemata may provide machines with more appropriate information, but are often perceived as too restrictive by human annotators. On the other hand, less restrictive schemata (e.g. free text fields etc) often yield results in which the terminology is used subjectively and inconsequently to the extent that is hardly of any value for processing by a machine.

Longevity While longevity is a problem for many electronic documents, it may be even worse for their annotations. It is very hard to design annotation schemata that are applicable both in the short and long term, and that are both sufficiently specific to be useful within their original domain *and* sufficiently generic to be used across domains. Such schemata require extreme flexibility in tool support for extensions, modifications, version tracking, etc. that extends the current state of the art.

Standardization The tools used by the annotators are often not the same as the tools used by the end user, so a relatively high degree of standardization is needed to provide the required interoperability. On the syntax level, to ensure that one tool can parse the formats produced by the other, but also on the semantic level, to make sure that tools can figure out to what shared concepts the terms used by different parties refer to. In practice, semantic interoperability requires a certain degree of automatic inferencing. As a minimum requirement, tools need to be capable to find out when different terms are equivalent and when terms are related to each other by a subsumption relation.

Privacy Metadata might provide privacy or security sensitive information that needs to be handled with particular care. Examples include medical documents, annotated with personal information about the patient, or digital reproductions of artwork, annotated with the insurance value of the original artifacts.

Despite these general metadata problems, there are also a number of issues that are specific to the use of metadata in a multimedia context.

4.2 Multimedia-specific metadata problems

Granularity The goal of multimedia production is to provide interesting and relevant information by the composition of different audio-visual information units. The problem is that each audio-visual unit might provide a limited amount of information but still contains a wealth of meaning. For example, a single image shown in isolation may provide an identifiable semantics. The same image presented in a sequence, however, might appear with a modulated semantics because the order created new levels of meaning. The same effect appears in sequences of shots and scenes

in film, only that process of signification is more complex here. Hence, the essential aspects of audio-visual production is to get the relationship between the two representational systems, i.e. the image (space) and order (time) right, because these relationships communicate a significant part of the meaning. Making these kinds of relationships explicit in metadata requires descriptions on multiple levels that go beyond direct content description, and also cover implicit connotations, narrativity and discourse relationships, relations describing the rhetorical argument, etc. (see [21] for a general overiew and [4, 19] for computational semantic-based representation models of audio-visual data). Making this diverse range of implicit meaning explicit by capturing it in metadata annotations is a non-trivial challenge. For multimedia, it is also common to attach metadata to objects that appear in the media stream, e.g. an object in a video. That metadata might apply to any frame featuring that object. Specification of such frames is hard because it is often independent of shot or scene boundaries. Different units of metadata may address different frame ranges, requiring a stratified approach [22, 28]. Even within a specific frame, identifying the target object is often not trivial.[17]

Audio-visual interpretation The specific subjectivity of human annotators is often a more serious obstruction when the semantics of non-textual documents need to be interpreted. The basis for this problem is rooted in the myriad of perceptional, cognitive and cultural codes buried in audio-visual material, which are the source for varying interpretations and thus annotations between individuals. The goal is to facilitate metadata generation that allows to establish different views on the same audio-visual material.

Work Flow Management Nevertheless, a lot of high quality metadata is produced during the different phases of the multimedia production process. Examples include scenarios, scripts, storyboards, edit decision lists, etc. Additionally, semantic-loaded low-level features are captured during the multimedia production process. Many digital camera's already record a continuous stream of information about the camera's settings (zoom, focus and other information about the lens, shutter speed, white balance, wall clock time, etc) along with the video signal. Unfortunately, most of this metadata is no longer available in the final version that is distributed to the end-user. The challenge lies thus in controlling this flow of metadata during the entire production chain and making the relevant parts accessible to the people and applications authorized to use it.

Repurposing Repurposing of media items into a new, coherent story is for multimedia even more challenging than for text. Here the aim is to provide the description of alternating contexts, where in each context the particular media item might play a different rhetoric role.

Data quantity and streaming The sheer bulk of digital multimedia content often makes a complete download of the material before playback undesirable, giving rise to streaming content delivery. Similar arguments apply to bulky multimedia metadata, that will need to be delivered in a streaming fashion without disrupting the stream of AV content.

Digital Rights Management Multimedia's more complex production process also makes digital rights management more complex than for text. Several parties (directors, producers, scenario writers, actors, etc.) may exercise their rights on a single media item.

All of the above problems need addressing to make the vision of a media-aware Semantic Web possible. The excessive nature of a discussion that combines all the various problems in one argument is, however, beyond the scope of this paper. Thus, we will focus in this article on those problems that are directly associated with the semantics of non-textual media.

Based on the scenario as described earlier, the essential problem we face is to supply means that allow the production and maintenance of high quality metadata. It is quality metadata that provides the significant syntactic, semantic, and semiotic aspects of the media's content necessary to establish the new, persuasive contexts that are required for effective restructuring, representing, resequencing, or re-purposing of existing content.

In the next section we investigate, therefore, the main stages of the media production process to indicate where (semi)automatic mechanisms can provide the required machine-readable descriptions based on standardized languages, as provided by W3C or ISO.

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.

[17] Addressing the target of metadata in multimedia is similar to the identification of the source and target of links in time-based hypermedia, which is still seen as a open issue, especially from a standardization point of view.

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] T. Berners-Lee. *Weaving the Web*. Orion Business, 1999.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001.

[3] W. Cathro. Metadata: An Overview. In *Standards Australia Seminar: Matching Discovery and Recovery*, August 1997. See also http://dublincore.org.

[4] M. Davis. *Media Streams: Representing Video for Retrieval and Repurposing*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1995.

[5] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, Inc., 1999.

[6] International Organization for Standardization/International Electrotechnical Commission. MPEG-4 Overview - (V.18 - Singapore Version). ISO/IEC JTC1/SC29/WG11 N4030, Singapore, March 2001.

[7] International Organization for Standardization/International Electrotechnical Commission. MPEG-7 Requirements Document V.15. ISO/IEC JTC1/SC29/WG11/N4317, Sydney, July 2001.

[8] International Organization for Standardization/International Electrotechnical Commission. 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.

[9] International Organization for Standardization/International Electrotechnical Commission. 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.

[10] International Organization for Standardization/International Electrotechnical Commission. 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.

[11] International Organization for Standardization/International Electrotechnical Commission. 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.

[12] International Organization for Standardization/International Electrotechnical Commission. 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.

[13] International Organization for Standardization/International Electrotechnical Commission. MPEG-21 Overview v.5. ISO/IEC JTC1/SC29/WG11/N5231, Shanghai, October 2002.

[14] International Organization for Standardization/International Electrotechnical Commission. Overview of the MPEG-7 Standard (version 8). ISO/IEC JTC1/SC29/WG11/N4980, Klagenfurt, July 2002.

[15] S. Johnson, P. Jourlin, K. S. Jones, and P. Woodland. Audio Indexing and retrieval of Complete Broadcast News Shows. In RIAO 2000 [24], pages 1163–1177.

[16] M. Kim, S. Wood, and L.-T. Cheok. Extensible MPEG-4 Textual Format (XMT). In *Proceedings of the eighth ACM Multimedia Conference*, Los Angeles, California, October 30 - November 4, 2000. ACM Press.

[17] M. Melucci and N. Orio. SMILE: a System for Content-based Musical Information Retrieval Environments. In RIAO 2000 [24], pages 1246–1260.

[18] T. Mills, D. Pye, N. Hollinghurst, and K. Wood. At&TV: Broadcast Television and Radio Retrieval. In RIAO 2000 [24], pages 1135–1144.

[19] F. Nack. *AUTEUR: The Application of Video Semantics and Theme Representation in Automated Video Editing*. PhD thesis, Lancaster University, 1996.

[20] F. Nack. The Future of Media Computing - From Ontology-based Semiotics to Computational Intelligence. In C. Dorai and S. Venkatesh, editors,

Media Computing - Computational Media Aesthetics, pages 159–196. Kluwer Academic Publishers, Boston, Dordrecht, London, 2002.

[21] F. Nack and L. Hardman. Denotative and Connotative Semantics in Hypermedia: Proposal for a Semiotic-Aware Architecture. Technical Report INS-R0202, CWI, March 2002.

[22] A. Parkes. Settings and the Settings Structure: The Description and Automated Propagation of Networks for Perusing Videodisk Image States. In N. J. Belkin and C. J. van Rijsbergen, editors, *SIGIR '89*, pages 229–238, Cambridge, MA., 1989.

[23] 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.

[24] *RIAO' 2000 Conference proceedings*, volume 2, Collége de France, Paris, France, April 2000.

[25] J. Ryu, Y. Sohn, and M. Kim. MPEG-7 Metadata Authoring Tool. In *Proceedings of the tenth ACM International Conference on Multimedia*, pages 267–270, Juan-les-Pins, France, December 1 - December 6, 2002. ACM Press.

[26] S. Santini and R. Jain. Integrated Browsing and Querying for Image Databases. *IEEE Multimedia*, 7(3):26–39, July 2000.

[27] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001.

[28] T. G. A. Smith and G. Davenport. The Stratification System. A Design Environment for Random Access Video. In *ACM workshop on Networking and Operating System Support for Digital Audio and Video.*, San Diego, California, 1992.

[29] The Styrian Competence Center for Knowledge Management. Knowledge Retrieval and Knowledge Visualization - Caliph & Emir Download Page, 2002.

[30] The Unicode Consortium. The Unicode Standard, Version 3.0. Reading, Mass. Addison-Wesley Developers Press, 2000.

[31] The Video Wizard Consortium. The Video Wizard Home Page, 2002.

[32] F. van Harmelen, P. F. Patel-Schneider, and I. Horrocks. Reference description of the DAML+OIL (March 2001) ontology markup language, 2001.