That Obscure Object of Desire: Multimedia Metadata on the Web

Jacco van Ossenbruggen, Frank Nack, Lynda Hardman

1 Introduction

In this article we discuss the advances in, and remaining problems of, making use of audio-visual media in a semantic-based environment, such as the Semantic Web, facilitated through media-aware and ontology-based metadata.

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 [2, 4] and ISO's efforts in the direction of complex media content modeling, in particular the the Multimedia Content Description Interface (MPEG-7) [30, 31, 32, 33, 34]. We chose these two approaches as they provide the potential techniques to establish a media-aware Semantic Web, even though at the time of writing the approaches seem to be diverging rather than converging.

The Semantic Web should bring machine-processable content to Web pages, thus being an extension of the current Web. The aim is to add ontology-based metadata to Web resources to improve Internet search and provide means for machine-based reasoning about the content. A major drawback of the current semantic Web developments, however, is its media-agnostic view on Web resources. The specific needs of dynamic audio-visual media with its variety of data representations is not recognized. That is, however, precisely what the intention of MPEG is, partially addressed in MPEG-4 [27] and MPEG-21 [35] and fully in MPEG-7.

In this paper, we explain that the conceptual ideas and technologies discussed in both approaches are essential for the next step in Web-based multimedia development. Unfortunately, there are still many practical obstacles that block their widespread use for providing multimedia metadata on the Web. We show that a media-ware Semantic Web will blur the boundaries between traditional categories like preproduction, production, and postproduction, with far-reaching effects on concepts such as data, metadata, consumer and producer.

The paper is structured as follows. We first provide a scenario to explain our vision of a media-aware Semantic Web and derive from it 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. 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.

2 Example scenario and problem statement

> Imagine, five years from now, you are the head of a lab that develops mobile communication devices and you would like to develop the 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. 2008'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. You are pleasantly surprised by the fact that the automatically generated story line is actually coherent and that the presentation 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 DRM 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.

What problems need to be solved to make this scenario work? Many of these problems arise because the tools involved need, to some extent, operate on the semantics of the media items involved. Tools need to know what the media is about, both for the more traditional retrieval tasks (finding relevant media items) and the more innovative tasks (generating a coherent storyline from a set of media items).

Another common theme is that tools need information about the context of media items: in what context have media items originally been produced, and to what extent will they fit into the current context? While part of this is directly related to content semantics, it also involves understanding the semantics of the technical and social context, 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. Information that is designed to make the semantics of other information explicit, in a machine readable way, is often referred to as *metadata*. The process of adding metadata to existing information is known as *annotation*.

2.1 General problems with metadata

High quality metadata is essential for supporting many multimedia applications, including those sketched in the scenario above. 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, as yet, only humans can make to some extent reliably. 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.

Granularity A related problem is that 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.

2.2 Multimedia-specific metadata problems

There are also a number of problems that are specific to the use of metadata in a multimedia context.

Audio-visual interpretation The subjectivity of human annotators is often a more serious obstruction when the semantics of non-textual documents need to be interpreted. For text, annotators might base the terms used for their descriptions upon terms found in the text, which is not possible in non-textual media. For example, even though an image might provide a limited amount of visual information, it contains a wealth of meaning. This functionality is based on the two formal structures within visuals (single or moving), the content (realized through the image) and its spatial and temporal relationship with other media items. The resulting combinatorial possibilities form the basis for the subjective interpretation for each viewer.

Fragmentation While the granularity of annotations might also be an issue for text documents, in practice most metadata applies to either the document as a whole, or to a fragment with boundaries that are inherent in the text's structure, e.g. metadata that relates to specific chapter, paragraph or sentence (note that addressing the target of metadata in text is similar to the identification of the source and target of links in hypertext). For multimedia, it is 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 [48, 57]. Even within a specific frame, identifying the target object is often not trivial (note that addressing the target of metadata in multimedia is similar to the identification of the source and target of links in time-based hypermedia).

Work Flow Management Paradoxically, another problem relates directly to the high quality metadata that is produced during the different phases of the multimedia production process. Examples include scenarios, scripts, storyboards, edit decision lists, etc. 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, because of continuity problems, undesired side effects of montage, etc.

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. However, the excessive nature of a discussion that combines all the various problems in one argument is beyond the scope of this paper. Thus, we will focus in this paper 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 provide 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.

The problem is that the semantics of the media item, are ultimately drawn from the context of the current application and the context of the overall placement of the information in the domain the application is applied to. The understanding of audio-visual material relies both on its denotative and connotative aspects and neither can be deduced on the mere basis of the others' description. Thus, we hope to show that the traditional approaches in research and industry are too limited because they only target the final media product to characterize 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. [1, 20, 11, 53, 42, 37, 40, 39].

The advantage of these approaches is that they can be automatically applied to allow indexing. In art, for example, we intuitively see the importance of physical features by using them to identify styles, which is in particular helpful in cases where no metadata for in-depth interpretation is available. Nevertheless, a mere low-level description cannot provide more than an indication of what style type an image might have and it depends on the application up to which certainty it would accept such a retrieved media item. Hence, we have to incorporate annotation methods that go beyond the mere use low-level perceptual descriptors with limited semantics for content representation (approach over several semantic levels are described among others by [8, 25, 47]).

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.

3 Metadata in the multimedia production chain

Audio-visual (AV) media production, such as for news, documentaries, interactive games, virtual environments, or business presentations as described in the scenario, is a complex process, that differs from text production in many aspects.

An author of a text, who wants to communicate a message, can state this message explicitly in the text. A producer of a piece of AV material, however, usually needs to package the message implicitly within the AV content. Therefore, creating metadata that makes this type of communication explicit is even harder for AV material than it is for text.

Another issue that complicated media annotation is defining the part of the AV content that the annotation is about. For text, these techniques for pointing into a larger text are well established, both in

the traditional context of literary annotation and cross referencing, and in the more recent digital context of identification of hyperlink start and end points. For AV media, we still need to develop commonly accepted techniques to attach an annotation to, for example, a certain actor in a certain scene in a certain film.

The role of composition in creating the intended message is also more prominent in the case of AV media production, where the goal is to provide interesting and relevant information by the composition of different audio-visual information units. 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. [56, 45]. Again, making this diverse range of implicit meaning explicit by capturing it in metadata annotations is a non-trivial challenge.

So to a large extent, multimedia annotation is about making explicit the relevant information that is implicit in the AV content. This is hard to do retroactively (i.e. after the production is finished), and much of required information is already made explicit during the production. An obvious approach would thus aim at storing this metadata during the production phase, and making this accessible after production in a controlled way.

In the remainder of this section, we first discuss the requirements for such an approach. We sketch the role of standardization and syntactic interoperability. We then address the issue of semantic interoperability, and conclude by discussing the problems related to practical applicability.

3.1 Media metadata production: requirements

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

- preproduction, which is concerned with determining the main ideas and logic that forms the core of the production,
- production, where the main task is the acquisition of media material, and, finally,
- postproduction, which is oriented towards editorial decisions based on reviewing the material, editing, sound mixing, presenting, and archiving.

Traditionally, metadata production is directed primarily towards describing the final end product. The most striking difficulty with such an retrospective approaches is, that it does not provide important cognitive, content and context based information (see [15, 41] for a theoretical analysis of audio-visual signification, and [9, 43] for computational semantic-based representation models of audio-visual data). This type of semantic information, describing also the intermediate stages in the production and the decisions taken, is required to enrich the automatically extracted metadata and retrospective annotation.

It must be emphasized that the interrelationships between different production stages is extremely cross influential. Especially if each step would not only produce new media units (such as a script, a shot, or an edited shot including temporal changes and applied effects), but also, and even in a larger scale, relevant semantic data in the form of notes of script alterations, explicit descriptions of activities on the set, production schedules, editing lists with decision descriptions, and organizational information. This type of information is important because it represents the progression through the various alterations on a technical, structural, and a descriptional level that affects the nature of the final result, but also the original context of the individual media items used. Today this type of information is merely used to harmonize the workflow but is often lost after the production is finished.

Although with the advance of DVDs this information has already become an economic asset, in many cases production of all this extra information would be unrealistic if it would necessitate manual annotation — such an expensive endeavor would normally not be covered by the production or archival budget.

Instead, we would need a high level of 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 result including all the intermediate physical AV data, as well as the creative decisions made during the production of media and other contextual information, would be a non-linear, complex semantic network structure of relationships between the signs of the audio-visual information units and the ideas they represent, according to the creator's intention. Such a semantic-aware media production generates semantic, episodic, and technical representation structures and the relations between them organized as a network of specialized content description documents. It is important to stress, though, that a network produced during one production just reflects the purposes and intentions, or in other words the context, of that particular production. This would imply the use of domain-dependent constraints, the use of particular content description schemata, etc.

For example, the methodology involved in the creation of a dramatic film, a documentary, and a news program is very different. Commercial dramatic film production is typically a highly planned and linear process, while documentary making is much more iterative, with the story structure often being very vague until well into the editing process. News production is structured for very rapid assembly of material from very diverse sources. Media information systems must be able to accommodate all these styles of production, providing a common framework for the storage of media content and assembly of presentations, resulting in a set of specialized tools designed to assist the particular needs of a user. Again, it is essential that the tools should not place extra workload on the user: who should concentrate on the creative aspects of the work.

The idea of saving the complete production process is not new[1], though the consequent implementation in a digital environment remains difficult. It requires representational structures that reflect the constant changes the AV material undergoes during its production, but also dynamic semantic structures that allow the representation of modifying conceptual developments.

In addition, the deeper impact of digital media is to redefine the forms of media, to blur the boundaries between traditional categories like preproduction, production, and postproduction, and to alter the structure of 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 2 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.

Note that 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. More importantly, however, is the fact that the augmentation provides additional connotations about potential intrinsic meanings of the material gathered. These would depend on the circumstances and presuppositions of the receiver at the time of perception, along with the various legitimated codes and subcodes the receiver uses as interpretational channels. Note that the receiver of the existing media context and the producer of the new context are here identical. Also note that a piece of meta-data 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. The easiest way is to use the original sequence of the referenced film and link it together with the sequence that acts as reference. 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:

[1] The domain that demonstrates best what is possible today is news. In particular the technology of Avstar Systems LLC, the market leader in newsroom computing systems, shows how automatic indexing enables news journalists, editors, producers and directors to effectively search, browse and pre-edit all of their incoming videos from the desktop.

- 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 meta-data) depending on the context it is used in. Therefore, flexible description schemata need to be developed that reflect these roles.
- The nature of annotations is necessarily imperfect, incomplete, and preliminary because they accompany and document the dynamic progress of understanding a concept, which usually open up questions of aesthetics and subjective interpretation. Thus, semantic, episodic, and technical representation structures are required with the capability to change and grow.
- There is no such thing as a single and all-inclusive content description. Thus, there is a need for mechanisms to establish collective sets of descriptions growing over time (i.e. every degree of cognition might be illuminative for other interests and should remain accessible).
- 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. However, no such environment does exist to the current day, as the following analysis of todays media production environments will show.

3.2 Media production environments

The emphasis of the above discussion is based on the assumption that a media-aware Semantic Web, where a great variety of media is constantly generated, manipulated, analyzed, and commented on, can only emerge if people are provided with tools that recognize the dynamic nature of audio-visual media as well as the variety of data representations and their mixes, by supporting at the same time the integration of these data representations into the textually-oriented environment of existing 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.

Current professional IT tools exacerbate information loss during production. These applications assist in the processes of transforming ideas into scripts (e.g. a text editor, Dramatica, etc.), digital recording (e.g. Sony's 24p Camera, HDreel from director's friend), digital/analog editing (Media 100, Media Composer, FAST blue, df-cineHD and HDreel from director's friend, etc.), presentation design (Flash, Director/Shockwave, Flash, or Dreamweaver), or production management (Media-PPS, SAP R/2, SESAM, etc.).

The tools are often based on incompatible and proprietary architectures. Hence, it is not easy to establish an automatic information flow between different tools, or to support the information flow between distinct production phases. Additionally problematic is that most of these tools also 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 [9, 44, 52, 54, 59, 61]. 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 and thus are not more than a small step towards the intelligent use and reuse

of media production material. In particular, they omit the addressing of relevant representational problems on a subjective level, such as:

- what are the intrinsic media items that trigger an experience;
- how can we capture multi-sensory data during experiences;
- how can we capture context;
- how can we represent and adjust recollections in memory over time.

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 urge to use standardized representation structures.

Based on this urge and with reference to the discussions on the problematic aspects of metadata (see section 2) and the role of metadata in media production (see section 3.1), the rest of the paper provides a more detailed look at the two most relevant approaches towards machine-processable and semantic-based content description, namely the Semantic Web activity of the W3C and ISO's Multimedia Content Description Interface (MPEG-7). We analyze in particular their abilities to define structures for describing media semantics, discuss inherent syntactic and semantic problems in both approaches, their ways to address ontological problems for media semantics, and the problems of applying the theoretical concepts in both approaches to real world problems.

4 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 [2, 4] and the intelligent media applications thought about by the MPEG community [28, 29], and metadata plays a key role in realizing these visions. To implement their visions, the high level technical goals of both communities are very similar, and boil down to 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, before we then briefly introduce both methodologies to provide their genuine nature.

4.1 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 large documents collections. This has resulted in standards that address such issues, most notably the *Dublin Core* [6]. 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 itself. 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 [55]. 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 the understanding of what a resource document encompasses is different. 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"). In this context, MPEG-7 addresses aspects such as interoperability and globalization of data resources and flexibility of data management. For this purpose MPEG-7 had to reconcile the approaches that the DL, KR and MM-AI community favor, namely the general need for high-level descriptions of audio-visual content. However, due to its past, additionally those techniques and methodologies of the signal processing community needed to be embraced. The signal processing community, which had primarily focused on image analysis, saw success in only standardizing the representation of the content features (standardizing descriptor terminology). 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.

4.2 Metadata on the the Semantic Web

We give a summary of the current Semantic Web by using Tim Berners-Lee (in)famous[2] "layer cake" depicted in Figure 1 on the following page, 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.

4.2.1 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 the 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

[2]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 [49].

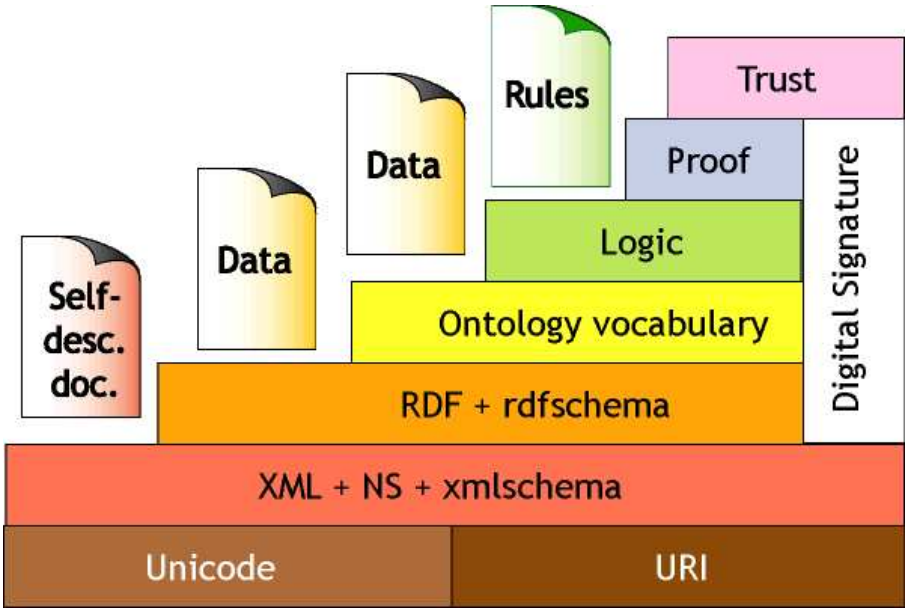

Figure 1: The layers of the Semantic Web envisioned by Tim Berners-Lee, as presented during a talk at XML 2001 (see `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`).

these fragment identifiers are media dependent and not defined by the URI specification [3]. 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 [72] 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 [60]. 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.

4.2.2 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 [68] and XML namespaces [5]. Other XML related languages, such as XPath [7], XPointer [13] and XLink [12] 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 [67], SMIL [65] and SVG [16] 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 available (compare this with most proprietary binary document formats whose content will become lost when the associated applications are no longer available).

4.2.3 RDF

As outlined in the discussion on meta-data 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) [66]. 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.

4.2.4 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 [70] 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 is being developed for both RDF and RDF Schema [21].

4.2.5 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 for the Web (OWL [71]). 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 [64]. DAML+OIL, on it turn, draws heavily upon one of the major results of the European On-To-Knowledge project: the Ontology Inference Layer [22] 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[3], and the language was renamed to *DAML+OIL*.

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.

4.3 Metadata within the MPEG framework

The Moving Pictures Expert Group is in charge of developing standards for coded representation of digital audio and video, and 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.

4.3.1 MPEG-4 - XMT

In MPEG-4 [27], the standard for multimedia on the Web, XMT [38] 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.

4.3.2 MPEG-7

The goal of MPEG-7 [30, 31, 32, 33, 34] 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. 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 [30].

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 [31]. 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 [32].

[3] See `http://www.daml.org`

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 [33].

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 [34].

Reference Software provides reference software to the standard [36].

Conformance specifies the guidelines and procedures for testing conformance of implementations of the standard [36].

Extraction and use specifies the guidelines and procedures for testing conformance of implementations to the standard [36].

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 section 5.

4.3.3 MPEG-21

The general goal of MPEG-21 [35] 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 and that is why we provided the short overview here. As already outlined in section 2.2 we will not analyze this part of the standard in our ongoing discussion, as this would expand the paper excessively. However, a detailed description of MPEG-21 can be found in the article by Jane Hunter in the same issue of this journal [24].

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. The following section analyzes both approaches in detail with respect to the requirements outlined in 3.1. 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

5 Semantic Web versus MPEG-7: A Language Analysis

The requirements for a language that facilitates the description of multimedia content for purposes as described in the scenario of section 2 and the production environment sketched in section 3.1 can be summarized as follows:

1. support the definition of syntactic rules to express and combine description structures at various levels of detail, which results in the provision of a rich set of syntactic, structural, cardinality and datatyping constraints

2. state spatial, temporal and conceptual relationships between the components of a description and between descriptions, so that a meaningful discourse about, or with, descriptions, through algebraic, logic, or functional means, is possible,

3. facilitate a diverse set of linking mechanisms between descriptions and the data that is described, which includes, in particular, means of segmentation for temporal media.

4. be platform and application independent and human- and machine-readable.

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

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

5.1 Syntactic Interoperability: XML Schema vs MPEG's DDL

Within MPEG-7, the DDL is intended to address the language requirements listed above. The language provides basically the same structure-oriented language elements as XML-Schema [69]. The only extensions to XML-Schema cover the ability to define arrays and matrices and to provide two additional datatypes, `basicTimePoint` and `basicDuration`, which allow specific temporal descriptions (see [31], pp. 9 – 14). Any available MPEG-7 parser addresses consequently only these extensions in addition to the other XML Schema-based language constructs.

Figure 2 on the next page and Figure 3 on page 17 provide a small example of a piece of MPEG-7 metadata. The first halve of the example shows how to address the target video fragment. Note that in addition to the URI, the `MediaTime` is used to identify the first eight minute segment of the video file this piece of metadata applies to. Moreover, this part of the example shows how the coding format of the audio-visual component can be described by using the `MediaFormat Descriptor`. Here the description of the video covers its aspect ratio, the frame size and the frame rate per second.

Figure 3 on page 17 exemplifies how the segmentation of the video in scenes and subscenes (`TemporalDecomposition`) can be achieved, where the `MediaTimePoint` provides the temporal start point of the audio-visual segment based on a Gregorian time scheme and the `MediaDuration` describes the temporal period of the segment. Moreover, the illustration also exemplifies a simple way to provide extra semantic annotations for a particular sequence supported by the `semantic` element.

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

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

Figure 2: Example of a MPEG-7 sequential description (I): linking to the video fragment.

1. lacks particular media-based datatypes. The datatypes used in the example are either standard XML Schema datatypes (such as integers etc) or media-specific datatypes defined by the MDS.

2. does not facilitate a diverse set of linking mechanisms between descriptions and the data that is described, which includes, in particular, means of segmentation for temporal media. Again, the locating and segmentation techniques used in the example are plain URLs combined with descriptors for time segments, also defined in the MDS.

3. does not facilitate definition of semantic relations, as does RDF Schema [70], or ontology-based modeling, such as DAML+OIL [64] or OWL [71]. The semantics of the relations between the syntax constructs used in the example are only defined by the English prose in the text of the standard, and lack the formal semantics that the Semantic Web languages have.

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

- specific datatypes required for the description of form and substance of media expression [ISO MPEG-7 2001e, pp. 49–103]. Extensions are provided in the parts Visual [32] and Audio [33];

- linking, identification and localization tools, mainly based on XPath but extended with particular temporal constructs, that provide a basic means of establishing references within a description and linking to the associated multimedia data ([34], pp. 74–103);

- graphs of relations, where the basic unit of a relation is built, similar to RDF, on a conceptual triple that allows the establishment of named relations between the parts in a description. The organization

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

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

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

Figure 3: Example of MPEG-7 sequential description (II) the actual annotations

of relations is restricted to a defined set of 11 topological and set-theoretic graph-relation types ([34], pp. 179–191);

- forms of spatio, temporal and spatio-temporal segmentations for video, audio, audio-visual, multimedia, and ink content, including a set of temporal and spatial relations ([34], pp. 251–400 and 458–540);
- a set of 45 semantic relations that allow the description of narrative structures ([34], pp. 401–457).

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

Moreover, dispersing language elements into description schemata asks for an evaluation complexity close to a validation level no parser can cope with. In fact, at the time of writing there is no MPEG-7 validator that can handle all the existing structures.

The lack of explicit semantics in MPEG-7 is, to some extent, inherent to the direct use of XML. The XML level of "self description" is limited to the extent that XML is only able to define the *syntax* of the elements in a language. There is no understanding of anything other than the hierarchical, syntactical structure of the document. What is needed is some way of specifying the semantics that is supposed to be communicated by the syntactical XML document structures [10]. Currently, the implicit semantics of an XML document can, if the author of the document employed markup using well-designed self-descriptive tag names, be perfectly clear to a human reader. However, to make these semantics explicit, and to communicate them in a machine understandable way, XML in itself is insufficient. Other layers, built on top of XML, are required to accomplish this. The semantics of the XML constructs used in MPEG-7 are defined by English prose. Within the Semantic Web, however, the semantics of the upper layers are RDF-based, and envisioned to be themselves machine-readable as much as possible.

5.2 Syntactic interoperability: RDF vs XML

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

Suppose that the "lifestyle video" fragment from the example scenario is published on the Web, distributed under an "Open Publication License". That this Web resource is indeed open content could, by interpreting the surrounding text on the HTML page from where it is linked to, be obvious to human readers, but not to a machine. To make this explicit, once could state this explicitly in RDF, and attach this statement as metadata to the Web page. In RDF triple terminology: the URL of the page (say, the (relative) URL `yup_lifestyle.mpg`) would denote the resource, the "dc:rights" label the property, and the string "OPL" the value. Figure 4 shows the common graph notation.

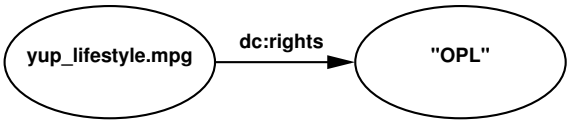

Figure 4: Simple graphical representation of an RDF triple

While RDF in itself is syntax neutral, it defines a XML serialization syntax for interchange. In addition, it defines an abbreviate form. As a result, even the simple, single triple defined above can be serialized to XML in two ways, as shown in Figure 5.

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

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

Figure 5: Example of two XML serializations of the same RDF statement.

Applications are expected to implement both forms and annotators are thus free to mix the two. In practice, many RDF files indeed use both forms simultaneously, which makes it hard to process RDF using

generic XML tools (e.g. it is almost impossible to write an XSLT stylesheet for any but the most trivial RDF documents). In real life, this problem is made even worse by the fact that, in most cases, the order in which RDF triples are serialized is irrelevant for RDF applications (while it is relevant for XML applications). Similarly, an RDF application might decide to serialize descriptions in a nested form without changing the RDF semantics (while in XML, the nesting of elements is usually consider relevant and can thus not be changed).

So while RDF technically uses XML, it makes it very hard to use generic XML tools for RDF processing. Unfortunately, the reverse also holds. In practice, is also very hard to make RDF tools process generic XML [49]. Suppose that in addition to the RDF metadata of our video fragment, our application has also access to the MPEG-7 metadata shown in Figure 2 and Figure 3. Despite the fact that it is encoded in XML, most RDF-based Semantic Web applications will not even be able to parse this on a syntactic level, unless one uses a non-standardized translation from MPEG's XML-based syntax to RDF, as advocated by Hunter [23].

5.3 Semantic interoperability: defining semantics

Another important difference between the Semantic Web and MPEG-7 approach is the way the semantics are defined. Note that in the case of RDF, the only semantics defined by RDF itself are the roles in the triple of the resource, property and value. For a generic RDF parser, `yup_lifestyle.mpg` is just an arbitrary URI of a subject resource, `dc:rights` just a property label, and "OPL" just a string literal denoting the property value. Defining what concepts may be used in a particular RDF graph, and how these concepts relate to one another is left to the application or to other languages such as RDF Schema or OWL.

For example, RDF Schema allows typing by introducing the notion of a class. Classes themselves can be organized in a `subClassOf` hierarchy. Web resources can be declared to be of a certain class by using the `type` property. One can define constraints on properties by defining the `domain` and `range` of each property to be of a specific class.

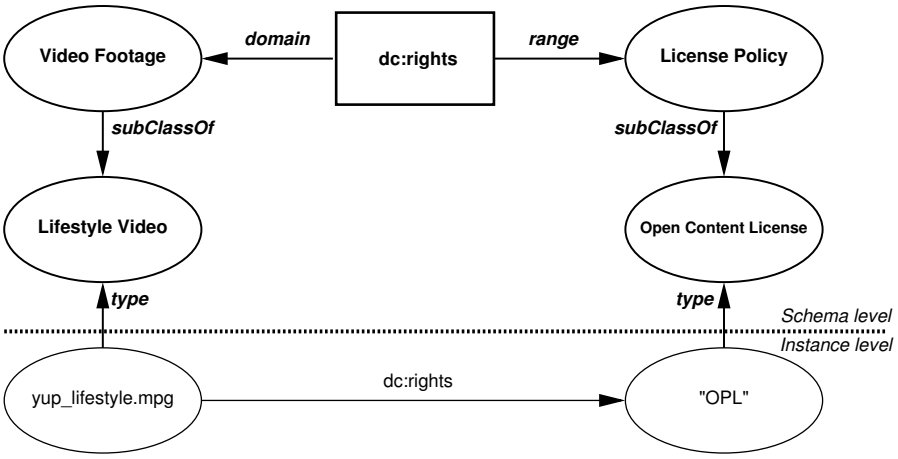

Figure 6: Fragment of the RDF Schema associated with the triple of Figure 4.

For example, a (fragment of) an RDF Schema for our example RDF triple is depicted in Figure 6. It defines an `Video Footage` class and `License Policy` class, with `Lifestyle Video` and `Open Content License` as their subclasses. In addition, this application constrains the `dc:rights` property, by only allowing instances of type `Video Footage` in its domain, and type `License Policy` in its range.

The RDF Schema above may provide some useful background context for applications that need to interpret our triple. More complex ontological relations could be added using languages such as OWL. All these relations remain, however, rather generic (that is, defined in terms of relations such as subClassOf etc.). For defining application-specific semantics the Semantic Web relies on third-party specifications. The meaning of the "dc:rights" property, for example, is defined by the Dublin Core Metadata Initiative.

Similarly, the Semantic Web stack itself does not define any multimedia specific semantics. For example, attaching RDF metadata to a particular segment of a video (as is done in the MPEG-7 example in Figure 3) requires a way to specify that specific fragment. Specification of such an addressing scheme is not considered to be within the scope of RDF or the other Semantic Web languages. Instead, it is left to a third party to develop such a scheme.

So the approach of defining semantics on the Semantic Web is to provide relative thin but generic layers that define increasingly complex semantic structures, and to defer the definition of domain and application specific ontologies to third parties. This approach can be contrasted to the MPEG-7 approach, which defines metadata syntax and semantics within the MPEG-7 standard. It also defines both the framework (including the DDL) and the actual ontologies. A large number of schemata in MPEG-7 establish ontological structures, as most schemata are inspired by the domain of broadcasting and audiovisual- based entertainment (see for example the VideoEditingSegment, the AgentDS, PlaceDS, or the user preference description schemata in the MDS). The large number of schemata, often describing similar aspects of the same semantic problem, and their interlocked nature, indicate the ontological role at least of the MDS. However, the attempt of abstraction to achieve domain independence makes it impossible to use those schemata as ontology items. Nevertheless, the advantage of the approach taken by MPEG-7 is that it provides a large vocabulary of description terms, developed especially for describing audiovisual material. A disadvantage is that the result is rather monolithic, and it is hard to reuse parts of the standard outside the MPEG-7 context. This problem cannot be underestimated, as the definition of semantics also points to the tribulation of mapping semantics.

5.4 Semantic interoperability: mapping semantics

In section 5.2 we already discussed the difficulties regarding the syntactic interoperability issues for applications that need to use XML and RDF simultaneously. Similar problems appear on the level of mapping semantics. The question to be asked, with respect to the Semantic Web, is: should an ontology layer use RDF(S) as its serialization syntax, or is it better to develop a (more concise) syntax directly in XML? In the RDF-based approach, one runs the risk of making integration with current and future XML-based approaches harder. Needless to say, the majority of current Web applications is XML-based, and even the MPEG-7 metadata framework is based on XML, not RDF. In addition, by using RDF syntax, and building incremental syntax layers on top of that, one also need to make sure that the underlying semantics can be layered in a similar fashion (for example, consider the potential problems when a pure RDF application interprets the semantics of a OWL document using the RDF serialization syntax. Ideally, the conclusions of the RDF application should be a subset of the conclusions an OWL application would make, but the two should not contradict one another).

On the other hand, by building the ontology layer directly on XML, one runs the risk of the development of two incompatible Semantic Webs: an XML/ontology-based “knowledge” Web versus an RDF/RDF Schema-based “metadata” Web. Clearly, the OWL Working Group chose the RDF-based approach. But the XML vs RDF question is closely related to one of the big controversies surrounding the Semantic Web in general: the question of whether the advantages of developing a common Semantic Web language stack such as shown in Figure 1 on page 10 really outweigh the more pragmatic approach of defining knowledge interchange formats directly in XML on a per application domain and per user community basis. The latter is the approach many E-business initiatives are currently taking. In theory, a Semantic Web-based approach would require less *a priori* commitment between the different user groups, and would promote the use of generic (free and commercial) tools. The Semantic Web would standardize more levels of the information stack, agreement about the semantics defined by these levels and the possibility of using off-the-shelf tools would thus come “for free”. In the XML-based alternative, users from a specific community would need to agree on these levels first, and then develop their own tools. Second, when new users would join in, adding their own set of knowledge bases and tools, the Semantic Web promises a better infrastructure for interoperability between the two worlds. It still remains to be seen to what extent these promises prove to be realistic in practice.

MPEG-7’s approach towards semantic interoperability is also critical because the standard acts as an ontology definition language as well as an ontology (see comments in section 5.1). This ambiguous conceptual state is a result of the decision to model the DDL on XML Schema rather than on RDF Schema.

This choice was mainly political, as RDF Schema was at that time, and still is not at the time of writing, a W3C recommendation and was thus not referable (for more insights w.r.t. the relationship between XML Schema and RDF Schema, see [10, 26]). The choice for XML-Schema as serialization syntax had far reaching consequences. As a mere syntax oriented language, the DDL could not provide the basis for some basic reasoning services, mainly subsumption based reasoning on the class and properties hierarchies. This required formal semantics that had to be established elsewhere, resulting in the `Semantic description tools` section in the MDS ([34], pp. 401-457). That MPEG-7 defined its own ontology environment is an asset with respect to interoperability within MPEG but it turned out as being a hurdle for the interoperability with other ontologies, as necessary mechanisms to connect into a source such as an ontology were not developed and the available linking mechanisms in MPEG-7 into external sources only cover other MPEG-7 documents or media items.

A potentially solution to overcome this problem of ontology interoperability is the `Classification schema`.The Classification schema facilitates the organizational wrapper for a controlled vocabulary built out of terms and the relations between them. The relations organize the terms in the form of a hierarchy, indicating if one term is broader or narrower in its meaning than another, a synonym or, in the given set of relations, the one of highest relevance. Thus, a classification schema in some sense covers aspects of a thesaurus. The classification schema allows the incorporation of other classification schema, though no indication is given, if this feature only takes account of the inclusion of other MPEG-7 classification schemata or also the insertion of or connection to other ontologies. Unfortunately, there is no information provided about how the mapping from previously unconnected terms should be achieved. Thus, the problem of mapping high-level media semantics is not solved yet and it remains questionable if the MPEG-7 approach of profiling schemata provides suitable solutions.

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

5.5 Semantics for media expressiveness

As stated earlier, it the process of audio-visual signification of multimedia objects that require special attention when it comes to semantics. This process, though based on common human knowledge and thematic structures (expression form), provides its own temporal-spatial realities based on patterns of juxtaposition of the media intrinsic parts (expression substance). The information provided on a perceptual level using objective measurements, such as those based on image or audio processing or pattern recognition, play an important role regarding the aesthetics of a multimedia unit and consequently its subjective interpretation. For our example of the business presentation this means that the added video sequences should not only strengthen the logical flow within the presentation by conveying the lifestyle of the new product's target audience but the material also has to express the expectations of the audience (in the example the board of managers) and has to fit into the overall style of the presentation.

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

Despite the semantic relations, already introduced in section 5.1, MPEG-7 additionally suggests a set of schemata that provides structures for multimedia summaries, points of view, partitions and variations ([34], pp. 458 – 540) and various forms of collections on a probabilistic, analytical or classification level ([34], pp. 541 – 600).

These schemata are very detailed, but they impose particular semantics on the user. In fact, the approach taken by the W3C, as exemplified through SMIL, representing a textual serialization of temporal and spatial aspects for multimedia presentations seems more promising because it is less rigid and thus more easily applicable.

Similar problematic is the the approach taken by MPEG-7 for representing substance of expression, i.e., the semantics of low-level audio and visual features as manifested in the parts Visual [32] and Audio [33]. It must be clearly stated that it is not so much the conceptual ideas to describe features for video such

as color, texture, shape, motion, or localization; or audio features such as series types (scalable, scalar, vector etc.), waveform, power, spectrum, harmonicity, silence, sound, spoken content, etc.

The dilemma is rather caused by the attempt to solve the challenge of representing the dynamic nature of audiovisual semantics by providing a binary (algorithmic) and textural (schema) description structure. The intention is that both representational forms provide the same information, since a requirement for the system specification of MPEG-7 is that "MPEG-7 data can be represented either in textual format, in binary format or a mixture of the two formats, depending on application usage. A bi-directional loss-less mapping between the textual and the binary representation is possible." ([30], p. 10).

This, however, turns out not to be the case. Both, Visual and Audio part provide many semantic descriptions relevant for the interpretation of the individual binary format of a schema. Take the `ColorStructureType` ([32], pp. 50-56) as an example. The descriptor specifies both color content (similar to that of a color histogram) and the structure of this content. It does this via the use of a structuring element. Its main function is image-to-image matching and its intended use is for still-image retrieval, where an image may consist of either arbitrarily shaped, possibly disconnected, regions or a single rectangular frame.

The binary format is accompanied by long textual and graphical descriptions giving detailed interpretational information about the extraction algorithm, re-quantization, color space and color quantization, and the `raw ColorStructure histogram accumulation`. All of that information is required to understand the meaning of every single element (bin) specified in the ColorStructure descriptor array of 8-bit integer values, $h(m)form \in \{0, 1, ..., M-1\}$.

None of this, however, made it into the textual description. In fact, the schema merely provides the structure of the result space, that is the size of the matrix that contains the results of the extraction algorithm (see the DDL representation syntax on [32], p. 51). The assumption during the development of the audio and visual schemata was that an agent would know about the semantics of a bin in the ColorStructure Schema and thus could react accordingly. The result is that the semantics of the array are not made explicit, what is to be expected from the textual description, but they are hidden in the standard document. However, for real analytic parity of audio-visual media within the Semantic Web it is of utter importance that the semantics of a media unit are made explicit, in particular as an XML-based parser is not able to evaluate the binary representation or the quasi binary representation of the current array content. While this problem may appear trivial, it has far reaching consequences because the use of low-level features for semantic-based descriptions is one of the few mechanisms available for the automatic annotation of media.

Having analyzed the conceptual ideas of the two standards it is obvious that there is potential to establish a media-aware Semantic Web. Despite the fact that there is still a lot of work ahead to provide the right framework, there is still the unanswered question on how applicable the provided technologies are. In section 3 we clearly stated that a Semantic Web can only emerge if the abstract idea of the media-aware Semantic Web can be turned into an environment that integrates the instantiation and maintenance of the dynamic structures into the actual working process. The next section reflects on theses issues.

5.6 Applicability of semantic structures

As pointed out in section 3.2 current technologies to support the instantiation and maintenance of the dynamic structures are still in their infancy, as is the research for the Semantic Web.

Our discussion on syntactic and semantic interoperability in sections 5.1 - 5.4 already demonstrated that the layer approach used in W3C technology seems to address the flexibility of descriptive structures, the essential requirement for intelligent media- and metaproduction (see section 3) better than the philosophy of the "universal" description schema for a domain as provided by MPEG-7.

The crucial dilemma of the MPEG-7 approach is that a description of a media item is basically forced into one document (see the definition of the root element in the MDS ([34], pp. 17-19). The instantiation of a complete description structure can be attached to the relevant media items and naturally, the resulting descriptions are consistent and interoperable within MPEG-7, even if the descriptions vary in their instantiated depth. However, though the structure of the schema can be complex, once it is created and used in instantiations, its structure cannot be altered. Any modification would cause inconsistencies with existing

documents[4]. However, new semantic needs could be expressed in an additional schema and its instantiations could then refer to the existing documents. Such a solution would require typed links that facilitate semantic relations between documents. However, links in MPEG-7, do not provide any information about the semantics of the relationship between documents. MPEG-7 relations, which supply the desired semantics through the introduction of `relationship` elements, can only be applied within a document, which again results in encapsulating the required network structure in a single document.

The solution to the above problem can be provided by a layered approach which facilitate the deconstruction of a description into an open semantic-oriented network structure. Various levels of clustering allow structure without losing flexibility and the potential for temporal growth.

Other instances of the complexity problem within MPEG-7 are:

- There are a great number of `abstract` elements, which are used to establish class structure[5]. However, abstract elements cannot appear in instantiations. When an element is declared to be abstract, a member of that element's substitutable class must appear in the instance document. To indicate that the derived type is not abstract, the XML namespace mechanism is used (xsi:type). Thus, a thorough understanding of schemata development is required, which makes instant schemata development for distinct domains hard, especially if the required schemata should cover simple descriptions, where the theoretically founded overhead is actually not required.

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

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

The structural complexity of MPEG-7 is obstructing the take-up of the standard, especially since few support tools exist for the manual generation of new schemata. The situation is more bleak with respect to semi-automated tools as described in section 3.2. Note, that support of tools for W3C technology is in every day production environments is on a similar spare level. This fact indicates for both standardization activities that they still operate on a theoretical level where the everyday use has not the highest priority in the development agenda. For the nearer future we see an analog development as at the beginning of the WWW, where only the introduction of user applicable graphical tools turned the pure academic infrastructure into a public environment.

However, there are projects in real world domains, such as the TV Anytime Forum [17], that give an indication of how media-aware semantic structures, such as those provided by MPEG-7, will be used in the future. The TV Anytime Forum develops specifications for services based on consumer digital storage devices. The semantic structures, all written in XML Schema, are self-developments and cover the essential aspects of media description, i.e. content description, content referencing and location, rights management and protection, systems and transport. Though the TV Anytime schemata are similar to the equivalent structures in MPEG-7, they are less complex in their organizational structure. TV Anytime includes, for example, the MPEG-7 schemata on user-modeling, though without incorporating the complete MPEG-7 organizational overhead. Rather, TV Anytime uses MPEG-7 as a namespace and thus be able to incorporate just the required schemata [18].

5.7 Summary

The problem of describing the semantics of multimedia in a network-based environment is complex. The discussion of the two main standards that address the difficult issues involved, [reflects this complexity. Table 1 recapitulates the similarities and differences between both approaches.

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

[5]The fundamental problem of class and instance, where sometimes an instance should also be a class, is implicitly addressed in MPEG-7 and also forms part of the language problem described earlier

	MPEG-7	Semantic Web
Syntax	XML	XML/RDF
Schema/ontology language	MPEG-7 DDL/XML Schema	RDF Schema/OWL
Composition	monolithic/big	small layers
Extensibility	? (version problems?)	designed to be extended
Multimedia ontologies	++	– (third party)
Linking into media items	++	– (media dependent)
Tool support	-	+
Real life applications	-	-

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

Our analysis showed that a ideal media-aware metadata language should be applicable outside the context is was initially designed for. Therefor, it should have a syntax-neutral basis and modular design. In particular the problems within MPEG-7 regarding the fusion of language syntax and schemata semantics clearly exemplified that a closed approach hinders the required modularity for description design, obstructing the needed interoperability on a syntactic and semantic level. Specific modules of such an ideal media description language, could adopt a number of description constructs from the visual and audio parts from MPEG-7. These could then be used to describe media aspects only and would allow the linking into conceptual and contextual descriptions expressed in semantic languages such as RDF, RDF Schema or OWL.

The flexible syntactic properties of this language, would facilitate re-use and inferencing about material for specific purposes, potentially leading to interoperability with other description formats, used in media-based standards, such as

- the Dynamic Metadata Dictionary-Unique Material Identifiers (UMIDs) [58]. A Unique material Identifier provides a reference for all captured audio-visual content units in a clip or shot, so that particular a content unit can always be located either locally, remotely or on a archive storage medium. The UMIDS provides for the link between the essence (video, audio, graphics, stills etc.) and the metadata and generates a time code and date (time-axis) for synchronizing this data.

- the Multimedia Home Platform (MHP) as part of the Digital Video Broadcasting (DVB) Project [50]. MHP is a series of measures designed to promote the harmonized transition from analogue TV to a digital interactive multimedia future. Based around a series of Java APIs (Application Programming Interfaces) for DVB set-top-boxes, MHP promises to provide a domestic platform, which will facilitate convergence like no other DVB specification.

- the P/Meta Standard developed by the Production Technology Management Committee (PMC) of the European Broadcasting Union (EBU), using the Standard Media Exchange Framework (SMEF) by the British Broadcasting Corporation (BBC) and SMPTE outputs, provides a common exchange framework and a format between members (and others) [63].

- the TV Anytime Forum [17], The TV Anytime Forum is an association of organizations that develops specifications to enable audio-visual and other services based on mass-market, high-volume digital storage.

- the Dublin Core Metadata Initiative [14],

- NewsML [46] is An XML-based standard to represent and manage news throughout its life cycle, including production, interchange, and consumer use.

- the Gateway to Educational Materials project [62]. A U.S. Department of Education initiative, the Gateway to Educational Materials SM (GEM) expands educators' capability to access Internet-based lesson plans, curriculum units and other educational materials.

- The Getty Research Institute's Vocabulary Databases (the Art & Architecture Thesaurus(r), the Union List of Artist Names(r), and the Getty Thesaurus of Geographic Names(tm)) [19], contain terminology and other information about the visual arts, architecture, artists, and geographic places.

The most crucial activity, however, is to leave the laboratories and provide real world cases that show the applicability of the technology. Related to that is the provision of development and maintenance tools, but also technology that allows to make use of the established semantic descriptions.

6 Conclusion and future research

In this article we argued that the traditional linear approach of generating information and thus meaning is far too restrictive, as any form of information is necessarily imperfect, incomplete and preliminary. Metadata accompanies and document the progress of interpretation and understanding of a concept. Consequently, we described the need for flexible, collective sets of descriptions growing over time and being collected during the actual working process, including the generation, restructuring, representing, resequencing, repurposing or redistributing of media.

We discussed the state of the art in content description for audio-visual media and the Semantic Web, by analyzing the two approaches towards machine-processable and semantic-based content description, namely the Semantic Web activity of the W3C and ISO's Multimedia Content Description Interface (MPEG-7). We showed that the two approaches provide the potential techniques to establish a media-aware Semantic Web. We illustrated that, though both approaches are XML-based, the differences on a philosophical and implementation level are substantial enough to make a merge between the two complicated. We focused in particular on the problems emerging from syntactic interoperability, the definition and mapping of semantics, and the description of expressiveness of media and showed that the current developments in both approaches are but a small step towards the intelligent use and reuse of media-based information. Consequently we outlined that a media-aware Semantic Web can only emerge if the boundaries between traditional categories like preproduction, production, and postproduction get blurred.

In fact, far more work is required on flexible formal annotation mechanisms and structures, but also on tools that first support human creativity to create the best material for the required task and additionally use the creative act to extract the significant syntactic, semantic and semiotic aspects of the content description.

Acknowledgments

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

References

[1] P. Aigrain, P. Joly, and V. Longueville. Medium Knowledge-Based Macro-Segmentation of Video into Sequences. In M. Maybury, editor, *IJCAI 95 - Workshop on Intelligent Multimedia Information Retrieval*, pages 5–16, Montréal, Canada, August 1995.

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

[3] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform Resource Identifiers (URI): Generic Syntax, August 1998. `http://www.ietf.org/rfc/rfc2396.txt`. RFC 2396.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001. `http://www.sciam.com/2001/0501issue/0501berners-lee.html`.

[5] T. Bray, D. Hollander, and A. Layman. Namespaces in XML. W3C Recommendations are available at http://www.w3.org/TR, Januari 14, 1999. http://www.w3.org/TR/REC-xml-names.

[6] W. Cathro. Metadata: An Overview. In *Standards Australia Seminar: Matching Discovery and Recovery*, August 1997. http://www.nla.gov.au/nla/staffpaper/cathro3.html. See also http://dublincore.org.

[7] J. Clark and S. DeRose. XML Path Language (XPath) Version 1.0. W3C Recommendation, 16 November 1999. http://www.w3.org/TR/xpath.

[8] C. Colombo, A. D. Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53, 1999.

[9] M. Davis. *Media Streams: Representing Video for Retrieval and Repurposing*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1995. http://garage.sims.berkeley.edu/pdfs/1995_Marc_Davis_Dissertation.pdf.

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

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

[12] S. DeRose, E. Maler, and D. Orchard. XML Linking Language (XLink). W3C Proposed Recommendations are available at http://www.w3.org/TR, 20 December 2000. http://www.w3.org/TR/xlink.

[13] S. DeRose, E. Maler, and J. Ron Daniel. XML Pointer Language (XPointer) Version 1.0. W3C Candidate Recommendations are available at http://www.w3.org/TR, 8 January 2001. http://www.w3.org/TR/WD-xptr. Superceded by http://www.w3.org/TR/xptr-framework/ etc.

[14] Dublin Core Community. Dublin Core Element Set, Version 1.1, 1999. http://www.dublincore.org/documents/dces/.

[15] U. Eco. Articulations of the Cinematic Code. In B. Nichols, editor, *Movies and Methods*, pages 590–607. Berkeley: University of California Press, 1976.

[16] J. Ferraiolo. Scalable Vector Graphics (SVG) 1.0 Specification. W3C Recommendation, 4 September 2001. http://www.w3.org/TR/SVG/.

[17] T. T.-A. Forum. The TV-Anytime Forum Home Page. http://www.tv-anytime.org/. http://www.tv-anytime.org/.

[18] T. T.-A. Forum. Specification Series: S3 On: Metadata Corrigenda 1 to S-3 V1.1. COR1_SP003v1.1, December 2001.

[19] Getty Research Institute. Art & Architecture Thesaurus (Online). http://www.getty.edu/research/tools/vocabulary/aat/, 2000. http://www.getty.edu/research/tools/vocabulary/aat/. Version 2.0.

[20] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40:71–79, 1997.

[21] P. Hayes. RDF Semantics. Work in progress. W3C Working Drafts are available at http://www.w3.org/TR, 23 Janaury 2003. http://www.w3.org/TR/rdf-mt/.

[22] I. Horrocks, D. Fensel, J. Broekstra, M. Erdman, C. Goble, F. van Harmelen, M. Klein, S. Staab, R. Struder, and E. Motta. The Ontology Inference Layer OIL. http://www.ontoknowledge.org/oil/TR/oil.long.html, 2000. http://www.ontoknowledge.org/oil/TR/oil.long.html.

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

[24] J. Hunter. MPEG-21. *IEEE Multimedia*, tentative reference. FIX ME XXX.

[25] J. Hunter and L. Armstrong. A Comparison of Schemas for Video Metadata Representation. In *The Eighth International World Wide Web Conference*, pages 353–373, Toronto, Canada, May 11-14, 1999.
http://archive.dstc.edu.au/RDU/staff/jane-hunter/www8/paper.html.

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

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

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

[29] International Organization for Standardization/International Electrotechnical Commission. Overview of the MPEG-7 Standard (version 6.0). ISO/IEC JTC1/SC29/WG11/N4509, Pattaya, December 2001.

[30] International Organization for Standardization/International Electrotechnical Commission. Text of ISO/IEC 15938-1/CD Information Technology - Multimedia Content Description Interface - Part 1 Systems. ISO/IEC JTC 1/SC 29/WG 11/ M704, Singapore, March 2001.

[31] International Organization for Standardization/International Electrotechnical Commission. Text of ISO/IEC 15938-2/FCD Information Technology - Multimedia Content Description Interface - Part 2: Description Definition Language. ISO/IEC JTC 1/SC 29/WG 11 N4288, Singapore, September 2001.

[32] 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.

[33] 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/N424, Sydney, July 2001.

[34] International Organization for Standardization/International Electrotechnical Commission. Text of ISO/IEC 15938-5/FCD Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes. ISO/IEC JTC 1/SC 29/WG, Singapore, September 2001.

[35] International Organization for Standardization/International Electrotechnical Commission. MPEG-21 Overview v.5. ISO/IEC JTC1/SC29/WG11/N5231, Shanghai, October 2002.
http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm.

[36] International Organization for Standardization/International Electrotechnical Commission. MPEG-7: Overview (version 8). ISO/IEC JTC1/SC29/WG11 N4980, http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm, July 2002.
http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

[37] S. Johnson, P. Jourlin, K. S. Jones, and P. Woodland. Audio Indexing and retrieval of Complete Broadcast News Shows. In RIAO 2000 [51], pages 1163–1177.

[38] 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. `http://www.acm.org/sigs/sigmm/MM2000/ep/michelle/`.

[39] K. Lemstroem and J. Tarhio. Searching Monophonic Patterns within Polyphonic Sources. In RIAO 2000 [51], pages 1261–1279.

[40] M. Melucci and N. Orio. SMILE: a System for Content-based Musical Information Retrieval Environments. In RIAO 2000 [51], pages 1246–1260.

[41] C. Metz. *Film Language: A Semiotic Of The Cinema.* New York: Oxford University Press., 1974.

[42] T. Mills, D. Pye, N. Hollinghurst, and K. Wood. At&TV: Broadcast Television and Radio Retrieval. In RIAO 2000 [51], pages 1135–1144.

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

[44] F. Nack. The Future of Media Computing - From Ontology-baesd 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.

[45] F. Nack and L. Hardman. Denotative and Connotative Semantics in Hypermedia: Proposal for a Semiotic-Aware Architecture. Technical Report INS-R0202, CWI, March 2002. `http://ftp.cwi.nl/CWIreports/INS//INS-R0202.pdf`.

[46] NewsML. The NewsML Home Page. http://www.newsml.org/, 2000. `http://www.newsml.org/`.

[47] F. Pachet and D. Cazzly. A Taxonomy of Musical Genres. In RIAO 2000 [51], pages 1238–1245.

[48] 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.

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

[50] T. D. V. B. Project. The Digital Video Broadcasting Project Home Page. http://www.dvb.org/latest.html. `http://www.dvb.org/latest.html`.

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

[52] 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. `http://www.acm.org/sigs/sigmm/MM2000/ep/rehm/`.

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

[54] A. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga. Ontology-based Photo Annotation. *IEEE Intelligent Systems*, 16(3):66–74, May-June 2001. `http://www.computer.org/intelligent/ex2001/x3066abs.htm`.

[55] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June 2001. `http://www.swi.psy.uva.nl/usr/Schreiber/papers/Schreiber01a.pdf`.

[56] S. B. Shum, V. Uren, G. Li, J. Domingue, and E. Motta. Visualizing Internetworked Argumentation. In P. A. Kirschner, S. J. B. Shum, and C. S. Carr, editors, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, pages 185–204. Springer-Verlag: London, 2003. http://kmi.open.ac.uk/projects/scholonto/docs/VizNetArg2002.pdf.

[57] 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.

[58] Society of Motion Picture and Television Engineers (SMPTE). Standard 330M-2000 for Television-Unique Material Identifier (UMID). Standard 330M-2000 for Television-Unique Material Identifier (UMID), SMPTE, White Plains, N.Y., 2000.

[59] The Styrian Competence Center for Knowledge Management. Knowledge Retrieval and Knowledge Visualization - Caliph & Emir Download Page, 2002. http://www.know-center.at/en/divisions/div3demos.htm.

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

[61] The Video Wizard Consortium. The Video Wizard Home Page, 2002. http://www.video-wizard.com/index-n.htm.

[62] T. G. to Educational Materials Consortium. The Gateway to Educational Materials Home Page. http://www.thegateway.org/welcome.html, 2002. http://www.thegateway.org/welcome.html.

[63] T. E. B. Union. The European Broadcasting Union Home Page. http://www.ebu.ch/. http://www.ebu.ch/.

[64] F. van Harmelen, P. F. Patel-Schneider, and I. Horrocks. Reference description of the DAML+OIL (March 2001) ontology markup language, 2001. http://www.daml.org/2001/03/reference.html.

[65] W3C. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification. W3C Recommendation, June 15, 1998. http://www.w3.org/TR/1998/REC-smil. Edited by Philipp Hoschka.

[66] W3C. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendations are available at http://www.w3.org/TR, February 22, 1999. http://www.w3.org/TR/REC-rdf-syntax. Edited by Ora Lassila and Ralph R. Swick.

[67] W3C. XHTML 1.0: The Extensible HyperText Markup Language: A Reformulation of HTML 4.0 in XML 1.0. W3C Recommendation, January 26, 2000. http://www.w3.org/TR/xhtml1.

[68] W3C. XML Schema Part 0: Primer. W3C Recommendation, May 2, 2001. http://www.w3.org/TR/xmlschema-0/. Edited by David C. Fallside.

[69] W3C. XML Schema Part 1: Structures. W3C Recommendation, May 2, 2001. http://www.w3.org/TR/xmlschema-1/. Edited by Henry S. Thompson, David Beech, Murray Maloney and Noah Mendelsohn.

[70] W3C. RDF Vocabulary Description Language 1.0: RDF Schema. Work in progress. W3C Working Drafts are available at http://www.w3.org/TR, 12 November 2002. http://www.w3.org/TR/rdf-schema/. Edited by Dan Brickley and R.V. Guha.

[71] W3C. Web Ontology Language (OWL) Reference Version 1.0. Work in progress. W3C Working Drafts are available at http://www.w3.org/TR, 12 November 2002. http://www.w3.org/TR/owl-ref/. Edited by Mike Dean and Dan Connolly and Frank van Harmelen and James Hendler and Ian Horrocks and Deborah L. McGuinness and Peter F. Patel-Schneider Lynn Andrea Stein.

[72] W3C. XPointer Framework. W3C Proposed Recommendations are available at http://www.w3.org/TR, 13 November 2002. http://www.w3.org/TR/xptr-framework/.