INVITED PAPER

# Multimedia Retrieval and Delivery: Essential Metadata Challenges and Standards

*To allow reduced costs, technical competition and evolution, and development of sizeable markets, standards are needed for metadata about the nature, production, management and use of multimedia material.*

By Fernando Pereira, *Fellow IEEE*, Anthony Vetro, *Senior Member IEEE*,
and Thomas Sikora, *Senior Member IEEE*

**ABSTRACT** | Multimedia information retrieval (MIR) and delivery plays an important role in many application domains due to the increasing need to identify, filter, and manage growing amounts of data, notably multimedia information. To efficiently manage and exchange multimedia information, interoperability between coded data and metadata is required and standardization is central to achieving the necessary level of interoperability. In the context of this paper, the term retrieval refers to the process by which a user, human or machine, identifies the content it needs, and the term delivery refers to the adaptive transport and consumption of the identified content in a particular context or usage environment. Both the retrieval and delivery processes may require content and context metadata. This paper will argue that maximum quality of experience depends not only on the content itself (and thus content metadata) but also on the consumption conditions (thus context metadata). Additionally, the rights and protection conditions have become critically important in recent years, especially with the explosion of electronic music commerce and different "shopping" conditions. This paper will review existing multimedia standards related to information retrieval and adaptive delivery of multimedia content, emphasizing the need for such standards, and will show how these standards can help the development, dissemination, and valorization of MIR research results. Moreover, it will also discuss limitations of the current standards and anticipate what future standardization activities are relevant and needed. Due to space limitations, the paper will mainly concentrate on MPEG standards although many other relevant standards are also reviewed and discussed.

**KEYWORDS** | Interoperability; metadata standards; MPEG standards; multimedia retrieval and delivery

## I. INTRODUCTION

It is largely recognized that multimedia data and related technologies are a growing part of our lives. The increasing ease for consumers to acquire, produce, process, store, transmit, and publish multimedia data has also transformed many of us from content consumers to content creators. Multimedia content is being searched, accessed, and managed under very different conditions in terms of users, location, time, devices, networks, etc. Consequently, content retrieval and delivery have become central issues in the provision of efficient and powerful multimedia experiences since users must not only be able to quickly and effectively retrieve and filter what they want, but also get access to a version of that content which maximizes their multimedia experiences. For the purposes of this paper, the term *retrieval* refers to the process by which a user, human or machine, identifies the content it needs, while the term *delivery* refers to the adaptive transport and consumption of the identified content in a particular context or usage environment. While consumption of (coded) multimedia data is the ultimate target for users, years and experience have shown that to shorten the bridge between content and users, the so-called metadata

Manuscript received June 1, 2007.

**F. Pereira** is with the Instituto Superior Técnico—Instituto de Telecomunicações, Lisbon, Portugal (e-mail: fp@lx.it.pt).
**A. Vetro** is with the Mitsubishi Electric Research Laboratory, Cambridge, MA 02139 USA (e-mail: avetro@merl.com).
**T. Sikora** is with the Institut für Telekommunikationssysteme, Fachgebiet Nachrichtenübertragung, Technische Universität Berlin, D-10587 Berlin, Germany (e-mail: sikora@nue.tu-berlin.de).

Digital Object Identifier: 10.1109/JPROC.2008.916384

0018-9219/$25.00 ©2008 IEEE

or "data about the data" plays a central role. Metadata or description data plays the role of a "visit card" for the content; metadata may be more or less complete and sophisticated, but typically provides key information about the content it represents in a quick and simple way, which makes multimedia content as searchable as text.

Considering the current multimedia landscape, it becomes very clear that standards provide a set of reference solutions for specific interoperability needs. While these reference solutions tend to take a snapshot of available technology for a particular problem at a given time, this is largely compensated for in the interoperability that is achieved, which enables sizeable markets, reduced costs, and technical competition and evolution. Interoperability may take on different meanings at each point in the multimedia chain, but it is clearly an indispensable requirement since it is the basis that devices and applications could work together for particular functions such as identification, retrieval, delivery, management, and consumption. This can easily be confirmed by the evolution of multimedia in the last 10–15 years. For instance, the widespread growth of both MP3 and JPEG formats created a level of interoperability among digital music and image formats that essentially launched the multimedia age. Interoperability at this large scale opened up many opportunities for consumers to retrieve and consume multimedia and created a generation of people sometimes referred to as the "MP3 generation."

The biggest challenge for standards is to match the market needs with the technological capabilities coming from research; this must happen in a timely manner. In developing standards, it is also essential to minimize the amount of normative technologies so that the areas for competition could be maximized and evolution in the realm of an interoperable framework could occur. As a case in point, audio and video encoders are not standardized, while decoders are; this separation has enabled the performance of encoders to increase over time and compete in the market while still maintaining interoperability with the decoders.

Although the most obvious standardization needs in the early days of multimedia were for coding formats, the proliferation of multimedia data soon created the need for metadata standards. Both the retrieval and delivery processes may require content and context metadata to maximize the quality of the user experience. Metadata allows the full value of digital multimedia content to be realized since it plays a key role in providing machine-processable content, a central requisite for more intelligent, adaptive, and powerful multimedia services and applications.

Since metadata addresses so many aspects of content representation and delivery there is a multiplicity of metadata types that may be relevant for different industries, processes, functionalities, application domains, etc. There is also metadata that does not change in the content life cycle and metadata that changes along the value chain or if the content is modified. Metadata may be produced and consumed at various points in the multimedia chain; since it may also be modified, the issue of protecting the metadata itself (not only the content) must be considered. Different types of metadata may be produced at the various steps in the chain using different methods, e.g., manual or automatic, and with different values, e.g., semantic or legal. For the purpose of better understanding and organizing the metadata problem, Table 1 includes a list of relevant metadata types with a brief definition.

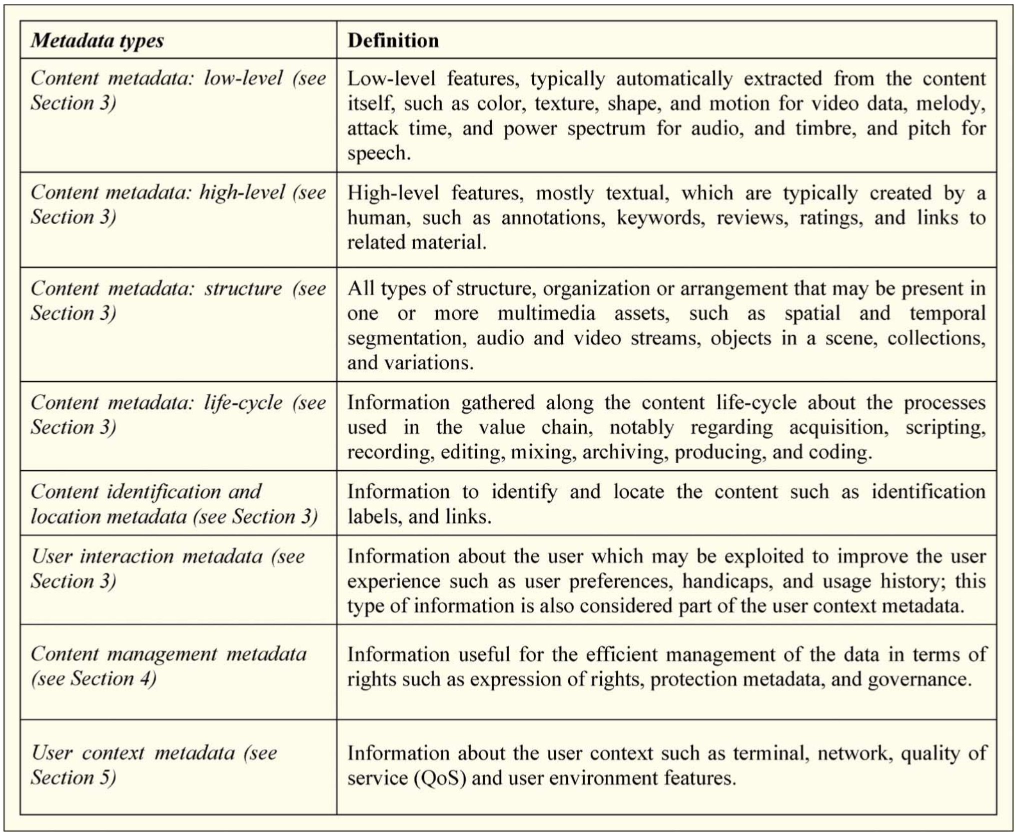
For the advanced retrieval and delivery of multimedia content, it is essential to achieve a semantic understanding of the media content. Semantics relate not only to the content itself but also to the context and thus to the social, cultural, and legal content dimensions. Because semantics are vital, the role of controlled terms such as classification schemes, taxonomies, and controlled vocabularies is also important. The set of metadata types presented in Table 1 will be used in the next sections to help structure the analysis of metadata standards. Due to length restrictions, only some of the most relevant metadata standards from those available will be considered.

The objective of this paper is to briefly review the most relevant available metadata standards, understand the development state of the field, and provide a strategic analysis of future standardization needs. Due to space limitations, the paper will give special emphasis to MPEG standards, although many other relevant standards are also reviewed and discussed. The paper also analyses the challenges of deploying metadata standards and the improvements in the standardization process that are already happening or should happen to help the deployment of metadata standards. For this, the paper is organized as follows. The next section covers the various metadata interoperability points in the production and delivery chain and discusses several practical scenarios in which metadata interoperability is critical. Section III reviews relevant standards for the description of multimedia content. Section IV provides an overview of metadata pertaining to multimedia rights and protection, while Section V describes context descriptions that characterize the usage environment where multimedia is ultimately consumed. Section VI presents a new dimension of multimedia standards targeting application formats that specify combinations of technology, notably metadata. Section VII identifies future needs and challenges in the area of metadata standards, and some concluding remarks are given in Section VIII.

## II. INTEROPERABLE METADATA SCENARIOS

While it is well accepted, and even evident for some domains like ID3 tags for digital music [1], that metadata

TABLE 1 Metadata Types With Definition



| ***Metadata types*** | **Definition** |
|---|---|
| *Content metadata: low-level (see Section 3)* | Low-level features, typically automatically extracted from the content itself, such as color, texture, shape, and motion for video data, melody, attack time, and power spectrum for audio, and timbre, and pitch for speech. |
| *Content metadata: high-level (see Section 3)* | High-level features, mostly textual, which are typically created by a human, such as annotations, keywords, reviews, ratings, and links to related material. |
| *Content metadata: structure (see Section 3)* | All types of structure, organization or arrangement that may be present in one or more multimedia assets, such as spatial and temporal segmentation, audio and video streams, objects in a scene, collections, and variations. |
| *Content metadata: life-cycle (see Section 3)* | Information gathered along the content life-cycle about the processes used in the value chain, notably regarding acquisition, scripting, recording, editing, mixing, archiving, producing, and coding. |
| *Content identification and location metadata (see Section 3)* | Information to identify and locate the content such as identification labels, and links. |
| *User interaction metadata (see Section 3)* | Information about the user which may be exploited to improve the user experience such as user preferences, handicaps, and usage history; this type of information is also considered part of the user context metadata. |
| *Content management metadata (see Section 4)* | Information useful for the efficient management of the data in terms of rights such as expression of rights, protection metadata, and governance. |
| *User context metadata (see Section 5)* | Information about the user context such as terminal, network, quality of service (QoS) and user environment features. |

standards may be important for the explosion of a business, it is also true that metadata standards have been finding a hard time to impose themselves compared with coding standards, e.g., the MPEG-7 standard as discussed later in this paper. This difficulty seems to derive from one major reason: the broader scope and meaning of metadata interoperability deriving from its coverage of a larger range of technical and business dimensions including a deeper dependence on the application domains. While coding stays at a rather low-level in the multimedia chain, metadata spans over the entire multimedia chain in terms of *industries*, e.g., television, music, and surveillance, *processes*, e.g., creation, production, packaging, management, and distribution, *application domains*, e.g., news, sports, and movies, *content representation approach*, e.g., framework, language, schema, semantics, and coding, *functionalities*, e.g., retrieval, summarization, filtering, and personalization, and *content types*, e.g., video, audio, graphics, 3-D, speech, and text [2]. The diversity and size of the metadata target is extremely broad.

In a metadata context, interoperability aims for a common exchange format between industries, application domains, processes in the multimedia chain, devices, etc., independently of the fact the metadata may be stored in some database or carried around together with the data itself. The added value of metadata interoperability may be present in many multimedia actions along the value chain (see Fig. 1):

1) allowing for aggregation of metadata provided by multiple sources such as various metadata service providers including local providers and personal metadata for content provided by a user;
2) facilitating access to content by consumers, e.g., access contents based on metadata retrieval from database, content transmission with associated metadata followed by local retrieval on consumer device, and access to video on demand services with metadata;
3) facilitating access and sharing of content among different users and user devices, e.g., locally stored on a hard-disc drive (HDD) recorder or PC in the home, or among devices of different users.

Because there are many ways and dimensions to address the metadata problem, several organizations and standardization bodies have developed several metadata standards according to different perspectives, more or less domain specific, and with different degrees of mutual complementary and harmonization. Among these organizations and bodies are the Dublin Core Metadata Initiative (DCMI) [3], the Society for Motion Pictures and Television Engineers (SMPTE) [4], the European Broadcasting Union (EBU) [5], the Moving Picture Experts

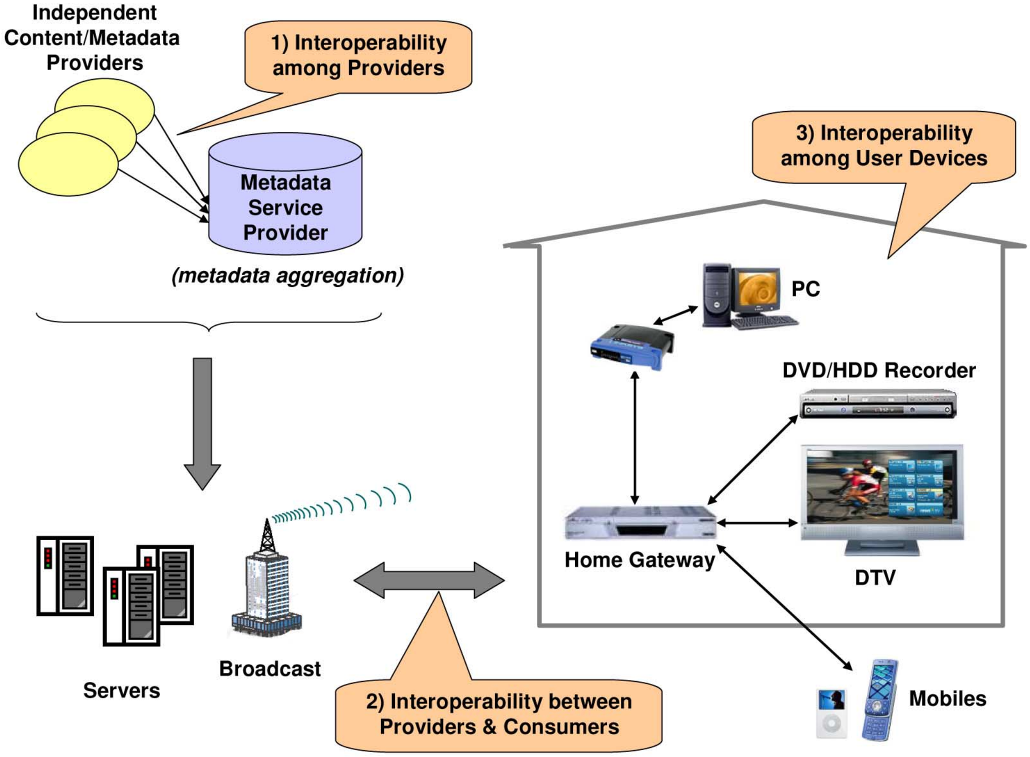

Independent Content/Metadata Providers
1) Interoperability among Providers
Metadata Service Provider
(metadata aggregation)
3) Interoperability among User Devices
PC
DVD/HDD Recorder
Home Gateway
DTV
Servers
Broadcast
2) Interoperability between Providers & Consumers
Mobiles

**Fig. 1.** *Scope of metadata interoperability.*

Group (MPEG) [6], the TV-Anytime Consortium [7], the World Wide Web Consortium (W3C) [8], and the International Press Telecommunications Council (IPTC) [9]. A brief analysis of available metadata standards demonstrates the growing importance of the Extensible Markup Language (XML) as a common definition language. However, commonalities between the various standards are very limited which makes interoperability between them very difficult, if not impossible, to achieve. In fact, the interoperability between standards is not so much based on the representation language, though this helps, but rather on the clear definition of what a particular term means and which relations can be associated with it. Thus, the strength of standards such as MPEG-7 is that the specification of description tools allows a comparison with the same or similar concepts in other standards. So, from a metadata point of view, it is the ontological dimension that provides interoperability, even more than the representation language. In practice, syntactic and semantic interoperability should go together since one without the other will always run into (insurmountable) obstacles [10].

Because addressing all the dimensions mentioned above in a single standard—industry, process, application domain, content types, technical approach, functionality, and content type—may easily prove to be an impossible task, the future of metadata standards seems to ask for a serious harmonization through a modular approach targeting complementary and application-specific specifications rather than trying to develop a super-standard addressing all possible metadata dimensions. An excellent example of the advantages of this approach is given by the adoption of XML Schema as a common schema definition language; this has allowed, for example, TV-Anytime [7] and MPEG-21 [6], [11] to reuse types from MPEG-7 [6], [12], rather than defining new types [2]. Among the main requirements relevant for a metadata standard are interoperability, modularity, extensibility, granularity, and media and format independence. Metadata standards should fulfill these requirements while overcoming some significant problems such as cost (high-quality metadata is expensive and time consuming), subjectivity (high-level annotations depend on the annotator's subjectivity), restrictiveness (tradeoff between annotator's restrictions and machine ambiguity), longevity (long-term needs are difficult to foresee), and privacy (metadata may touch individual privacy and public security) [13].

In the subsections that follow, several usage scenarios that highlight the importance of metadata interoperability in the context of Fig. 1 are discussed. In particular, information push, pull, and share scenarios are considered. Each of these scenarios demonstrates the practical interoperability needs at different points of the multimedia chain as well as the associated functionality and operations that are enabled.

## A. Information Pull

Research on multimedia database retrieval has been largely motivated by the incredible growth in digital content and the need to locate desired content quickly and effectively. This content may be available as part of very large professional archives, distributed on the Internet, or stored on consumer devices. Regardless of the location, the challenges for multimedia retrieval remain the same, which are basically to identify a specific piece of content or collections of content through an input query and related search mechanisms. As illustrated in Fig. 2, the information pull model assumes that search, browse, and filter operations are performed at the server side; in this way, only the target content is pulled from upstream locations.

Given this basic challenge for search and retrieval technology, there are varying needs for metadata interoperability depending on the content aggregation, distribution models, and level of interaction and query. Three distinct cases are outlined below.

1) At one extreme are the user-driven sites, such as YouTube, Google, and Flickr, where users upload images and video to huge multimedia galleries and provide input to assist the search and retrieval process. In the current model, users typically enter a set of textual keywords that index the uploaded content, and the search is done based on textual queries. Manual classification into a fixed set of categories could also be done to facilitate browsing by other users. One could easily imagine that back-end processes supplement the user-provided metadata with low-level content descriptors to enable improved clustering of related contents. It is noted that the metadata format, if any, would be specific to the particular site.

2) At the other extreme are more controlled forms of archiving and distribution by service providers, such as the video on demand over managed IPTV networks of telephone companies with guaranteed quality of service (QoS) and content protection. In such scenarios, various subsystems in the post-production, delivery, and consumer environments are required to interoperate, and search tasks could vary from pulling specific media clips from archive to searching an electronic program guide for a movie with a specific rating; electronic program guide distribution will be discussed further in the next section. In this case, the need to exchange metadata among various subsystems creates a strong need for metadata interoperability. However, the equipment that a

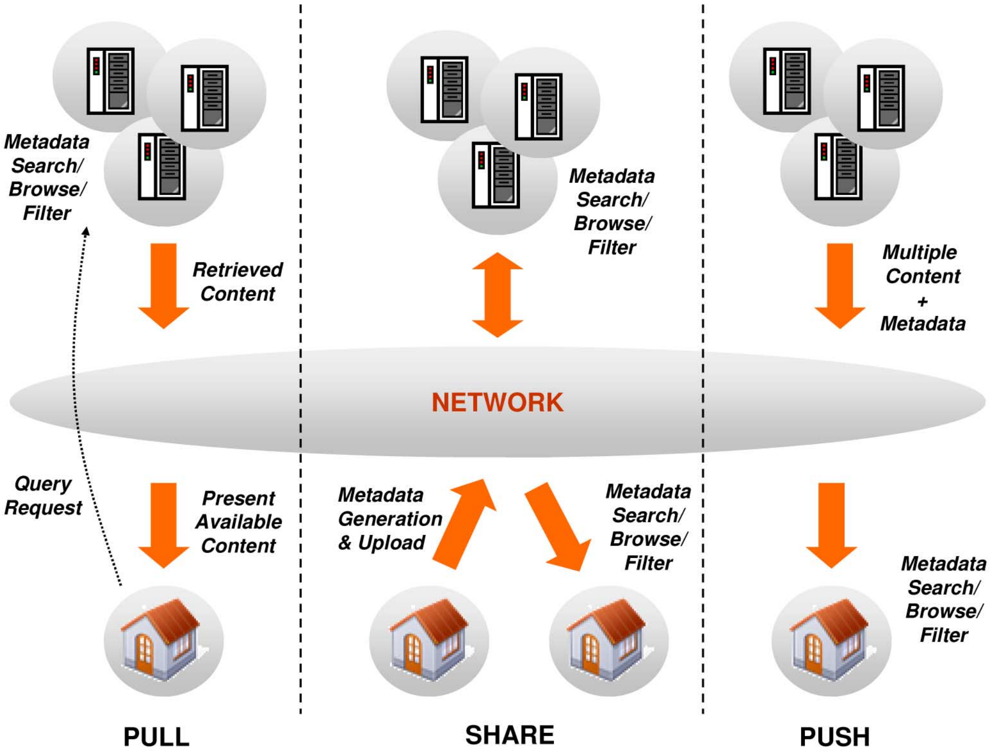

Metadata Search/ Browse/ Filter
Retrieved Content
Query Request
Present Available Content
PULL
Metadata Search/ Browse/ Filter
NETWORK
Metadata Generation & Upload
Metadata Search/ Browse/ Filter
SHARE
Multiple Content + Metadata
Metadata Search/ Browse/ Filter
PUSH

**Fig. 2.** *Pull, share, and push scenarios.*

consumer would use to search and browse content is usually tied to a particular service provider, therefore metadata interoperability between third parties is somewhat limited.

3) Between these two extremes are various web sites that publish multimedia information on a specific topic or genre, such as news (www.bbc.co.uk), sports (www.espn.com), consumer product review (www.cnet.com), etc. The query interface in this case is rather similar with the user-driven sites, while the metadata generation and exchange is closer to the service provider model.

Multimedia retrieval will vary in all of the above settings, as will the delivery of the media itself. With a purely text-based metadata and search process, XML provides a satisfactory level of interoperability and specific schemas could be defined to provide some simple structure to the metadata, e.g., author, keywords, dates, etc. However, to enable richer forms of searching for content, e.g., a similarity search based on audio and/or visual characteristics, the use of standardized metadata that specifies a particular set of useful features, is needed. The content metadata standards described in Section III could be used for this purpose.

It is also worth noting the added value that interoperable metadata brings in the use case scenarios described above. While it is true that many popular web and delivery services operate very successfully today using proprietary metadata, there is some added value when interoperable metadata is introduced. For instance, it becomes possible to search across different services and domains and possibly port content and its associated metadata more easily across devices and services.

### B. Information Push

The Electronic Program Guide (EPG) is essentially a description of programming information and a good example of the information push paradigm. In its most basic form, live television programs are described by program title, channel, and time. The display format that most people are familiar with today is a simple grid-like structure that could be used to browse the programs at a given time on a given channel. When a program is selected, the device could tune to the channel or schedule a recording of the selected program. The EPG data that is available today is significantly richer and could include a list of actors in a program, release and production information, the genre, parental guidance and rating data, as well as metadata about the media formats. Links to related material such as trailers and reviews may also be included along with unique identifiers and associated rights information.

Before any of this information reaches the consumer, various sources of information must be aggregated and prepared for delivery. As shown in Fig. 1, service providers require interoperability since the origin of different parts of metadata are likely to vary, e.g., details of media formats from one source and reviews on a program from another source. Therefore, having a standardized metadata format greatly simplifies the aggregation process. The next point of interoperability is between the service provider and the consumer, which is illustrated in Fig. 2. In this case, standardized metadata to be processed on a set-top box, PC, or gateway devices is obviously critical to enable various functionalities to be achieved on the consumer device, such as search, browse and filter, in an interoperable way; the rights metadata associated with particular content or channels also plays a vital role in the retrieval and delivery process. Finally, if permissible, redistribution of select portions of the EPG data could be used to enable access of content from networked devices within the home. With this type of information available, new applications could be developed to provide wider and more efficient access to the increasing amounts of content and improve the overall user experience.

There exist several standards that specify metadata to support EPG services. One has been developed by the TV-Anytime Forum and has been published as a European Telecommunications Standards Institute (ETSI) standard [14]. Another specification has been developed by the Consumer Electronics Association (CEA) for the U.S. market [15] and is built upon the TV-Anytime specification. It should be noted that both specifications reference data types defined by MPEG-7, including those that describe the media format. Whether such metadata is being delivered as part of an integrated content service, or provided as a separate service to users, its existence in the home will play an important role in achieving multimedia interoperability.

### C. Information Sharing

A relatively new and compelling scenario for interoperable metadata exists in the context of sharing personal multimedia metadata, i.e., metadata that is generated by the user, also known as user-generated content (UGC); see Fig. 2. YouTube, Flickr, and MySpace are prime examples of existing web services that offer users the ability to exchange multimedia content in an open and accessible environment. Metadata is associated with the content hosted by these services, thereby enabling the content to be classified, searched, and filtered.

Generally speaking, the metadata generation could be done using offline software tools or by consumer devices such as digital cameras or personal video recorders. In this information sharing scenario, it is assumed that one user would like to share or publish specific multimedia segments, such as a video clip, a compilation of images, or custom playlist of songs, with another user or within a community. This could be achieved with metadata that includes a link to the full content and segment information. Additional annotation, ratings, and related media could also be added in. For copyright material, digital rights management would be handled through appropriate mechanisms as described in Section IV.

Depending on the type of media, associated metadata, and copy protection requirements, the full multimedia package might be compatible with one of the Multimedia Application Formats (MAFs) described later in Section VI. In such a social networking scenario, metadata interoperability is required to facilitate all processes from identification and retrieval of content to its delivery and playback.

A key difference between this case and others described earlier is that the metadata to be shared is not necessarily handled by a service provider, but rather the metadata from one user is directly consumed and processed by the device of a second user. In the case that the two devices are manufactured by different companies, yet are still compatible with one another, true metadata interoperability that is not necessarily present in other cases has been achieved. It is important to note that when a service provider is in the loop, metadata is likely to be written into a particular exchange format of their choosing that might not be based on an open standard or published schema, which essentially limits interoperability between applications and devices. This point underscores the importance of open standards to maximize interoperability and create an ecosystem of applications that rely on metadata that could be exchanged based on an open format.

## III. MULTIMEDIA METADATA STANDARDS: DESCRIBING CONTENT

This section addresses some of the most important metadata standards for the description of the content itself, low- and high-level, as well as the content structure, production, and identification. Due to length limitations, this section will give special emphasis to the MPEG standards.

Those with interest in recent developments related to semantic interoperability are referred to the work of W3C's Multimedia Semantics Incubator Group [16]. In particular, this group has explored the very challenging and important problem of achieving metadata interoperability across existing metadata standards. This group has also investigated the added value of formal semantics, i.e., semantics that could be understood and interpreted more widely by both humans and machines, including the specification of richer vocabularies for describing properties and classes as well as relations between those classes.

### A. MPEG-7 Standard

The MPEG-7 standard defines standardized description tools that allow users or agents to search, identify, filter, and browse multimedia content. The pull, push, and share scenarios outlined in Fig. 2 were instrumental in defining parts of the MPEG-7 standards. Besides support for metadata and text descriptions of the multimedia content, much focus has been in the definition of efficient content-based description and retrieval specifications [6], [12], [17].

The main elements of the MPEG-7's standard are as follows.

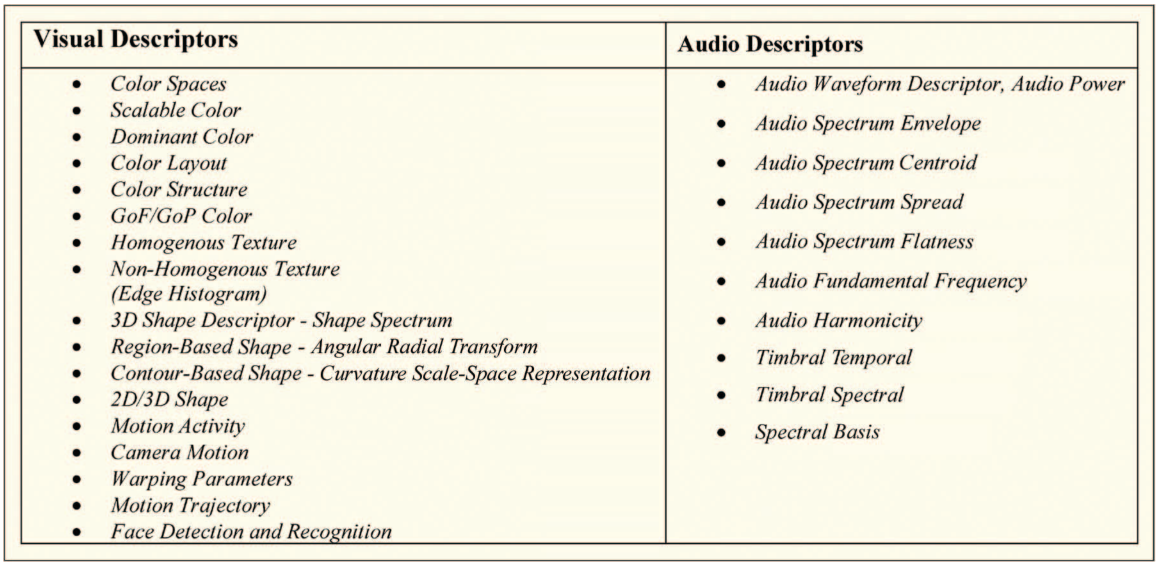
1) Descriptors (D) that define the syntax and the semantics of feature vectors and their elements. Descriptors bind a feature to a set of values.
2) Description Schemes (DS) that specify the structure and semantics of the relationships between the components of descriptors and between other description schemes.
3) Description Definition Language (DDL) to define the syntax of existing or new MPEG-7 multimedia description tools. This allows the extension and modification of description schemes and descriptors and the definition of new ones.
4) Binary coded representation of Ds or DSs. This enables efficient storage, transmission, multiplexing of Ds and DSs, synchronization of Ds with content, etc.

The MPEG-7 content descriptions (Ds and DSs) may include:

1) information describing the creation and production processes of the content, e.g., director, author, and title;
2) information related to the usage of the content, e.g., copyright pointers, usage history, and broadcasting schedule;
3) information of the storage features of the content, e.g., storage format, encoding;
4) structural information on temporal components of the content;
5) information about low-level features in the content, e.g., image color and edge information, motion in video, audio spectral energy distribution, sound timbres, and melody;
6) conceptual information of the event captured by the content, e.g., objects and events, interactions among objects;
7) information about how to browse the content in an efficient way;
8) information about collections of objects;
9) information about the interaction of the user with the content, e.g., user preferences, usage history.

Table 2 provides an overview of some of the more prominent low-level descriptors in MPEG-7. Fig. 3 outlines examples of audio descriptors for a particular audio sample. Once the MPEG-7 descriptions are available, search engines can be employed to search, filter, or browse multimedia material by comparing the individual low-level features of each image, video, or sound asset based on suitable similarity measures [17]. For most low-level descriptors, the MPEG-7 standard only partly describes how to extract these features. Extraction for most parts of

TABLE 2 Examples of MPEG-7 Visual and Audio Low-Level Descriptors



| Visual Descriptors | Audio Descriptors |
|---|---|
| *Color Spaces* | *Audio Waveform Descriptor, Audio Power* |
| *Scalable Color* | *Audio Spectrum Envelope* |
| *Dominant Color* | *Audio Spectrum Centroid* |
| *Color Layout* | *Audio Spectrum Spread* |
| *Color Structure* | *Audio Spectrum Flatness* |
| *GoF/GoP Color* | *Audio Fundamental Frequency* |
| *Homogenous Texture* | *Audio Harmonicity* |
| *Non-Homogenous Texture (Edge Histogram)* | *Timbral Temporal* |
| *3D Shape Descriptor - Shape Spectrum* | *Timbral Spectral* |
| *Region-Based Shape - Angular Radial Transform* | *Spectral Basis* |
| *Contour-Based Shape - Curvature Scale-Space Representation* | |
| *2D/3D Shape* | |
| *Motion Activity* | |
| *Camera Motion* | |
| *Warping Parameters* | |
| *Motion Trajectory* | |
| *Face Detection and Recognition* | |

the MPEG-7 standard is thus not normative. At the other end of the processing chain, the means by which MPEG-7 descriptions are further processed, i.e., for search and filtering of content, is also not specified by MPEG-7. In particular, the similarity matching technique between images, video, and/or sound is left to the individual applications. This approach provides maximum flexibility to applications for both extraction and retrieval, as well as space to innovate within the constraints of an interoperable metadata exchange format. In practice, the interoperability is provided by the schema definition, as the search engines know for what and where to look.

Practical search engine implementations may need to match content based on a weighted combination of descriptors, i.e., color and texture, or even between sound and video, and maybe also including common text-based queries. For this purpose, the MPEG-7 DSs facilitate a binding among combinations of descriptors. DSs also offer a rich set of metadata that might pertain to the higher level semantics of the multimedia content or other attributes

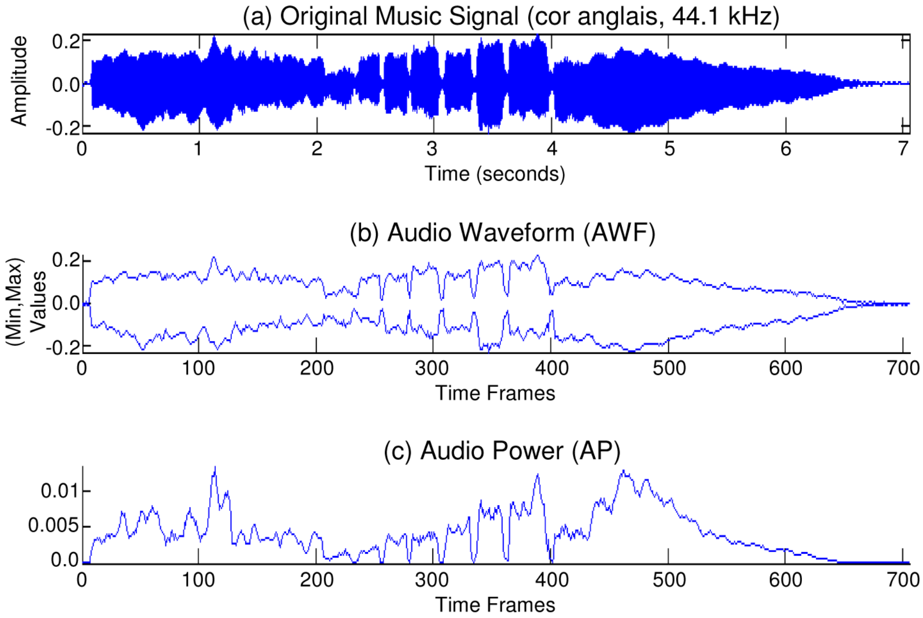


**Fig. 3.** ***MPEG-7 audio descriptors extracted from music signal. (a) Original music signal (cor anglais, 44.4 kHz). (b) Audio waveform (AWF). (c) Audio power (AP).***

such as production information, ratings, keywords, links to associated material, and so on. Such metadata is invaluable for multimedia retrieval. In addition to the vast number of DSs that could be used for retrieval purposes, there are also many DSs that could be applied for adaptive delivery of multimedia contents. For instance, the description of summaries, which provides a compact representation, or an abstraction, of the multimedia content, can be used for adaptive delivery of content in a variety of cases in which limitations exist on the capabilities of a terminal or even an end-user's time. Also, different versions of the multimedia content could be described by MPEG-7, which may ultimately provide a better match to the target device [18].

Ds and DSs are defined using the MPEG-7 Description Definition Language (DDL), which is based on the XML Schema Language. The DDL defines the syntactic rules to express and combine Ds and DSs. It allows users to create their own Ds and DSs and provides the means to express spatial, temporal, structural, and conceptual relationships between the elements of a DS and between DSs. It also provides the means to define a rich model for links and references between one or more descriptions and the data that it describes. The resulting descriptions can be expressed in either a textual form (i.e., human readable XML for editing, searching, filtering) or a compressed binary form (i.e., for storage or transmission). The Binary Format for MPEG-7 (BiM) defines a generic framework to facilitate the carriage and processing of MPEG-7 descriptions. It enables the streaming and the compression of any XML documents. BiM coders and decoders can handle any XML.

As stated in [10], unlike other approaches, such as the Resource Description Framework (RDF), ontology-based modeling, or the Web Ontology Language (OWL), the MPEG-7 DDL does not support the definition of semantic relations (although this was initially foreseen). In practice, the semantics of relations between the syntactic constructs are often only defined in Part 5 of the MPEG-7 standard, Multimedia Description Scheme (MDS), and hence lack the formal semantics of the semantic Web languages; this may have an impact in terms of semantic interoperability. It is also noted in [10] that MPEG-7 uses the DDL to define a normative schema that not only provides the necessary syntax, but also facilitates the description of the semantics of a single multimedia object or collections of objects in the form of a multimedia unit. These schema, however, are part of the MDS, not the description language. This may be one of the technical difficulties, among others, in adopting MPEG-7 for commercial purposes. In addition to this, the MPEG-7 standard was also completed in a difficult period of time for information technologies (end 2001-early 2002), so one might also speculate that there may have been some market and technology licensing difficulties at play as well.

Overall, MPEG-7 is a very powerful standard that has made an impact in a number of applications. Most notably, BiM has been widely embraced by organizations such as the 3rd Generation Partnership Project (3GPP) for mobile applications and the Association of Radio Industries and Businesses (ARIB) in Japan for efficient transmission of broadcasting metadata. As for the MPEG-7 description schemes, an MPEG-7 profile including a subset of MPEG-7 DSs has been adopted by the TV-Anytime Forum (see the following). The adoption of the MPEG-7 low-level descriptors is still slow and at present seen only in niche applications. In part, it might be that the benefits of an interoperable framework for multimedia descriptions have yet to be realized. Also, for many applications, it is important to automatically extract semantic information from the media, and the analysis tools either have limited capabilities or are not widely available. We will revisit some of these issues in Section VI when Multimedia Application Formats are discussed and address some further needs in Section VII.

### B. TV-Anytime Standard

TV-Anytime is an open standard for metadata describing TV and radio programs that is designed to support Personal Video Recorders (PVRs), program guides such as the ones outlined in Section II-B, and related technologies [7], [14]. The prime goal is to allow access to content from a wide variety of sources. The specifications are designed to exploit local persistent storage in consumer electronics platforms; they are network independent with regard to the means for content delivery to consumer electronics equipment, including various delivery mechanisms—e.g., Advanced Television Systems Committee (ATSC), Digital Video Broadcasting (DVB), Direct Broadcast Satellite (DBS), and others—as well as the Internet and enhanced TV.

A set of open specifications was developed that enable the interoperable searching, selection, acquisition, and management of content independent of the means of delivery. It addresses unidirectional broadcasts that are associated with bidirectional ancillary information and metadata services. This is made possible using the tools proposed by TV-Anytime, covering:

1) content and user-related description metadata;
2) content identification and location;
3) access to metadata services and associated security mechanisms.

The TV-Anytime Metadata Specification contains a TV-Anytime Usage History Thesaurus, a TV-Anytime Genre Dictionary, and the TV-Anytime Description Schemes, many of which reference MPEG-7 tools. In the context of TV-Anytime, metadata means "descriptive data about content," such as program title and synopsis, as well as information about user preferences and history. User preference information, such as favorite actors or TV shows, is included within the scope of TV-Anytime metadata to allow software agents to select content on the consumer's behalf. The collected usage history provides a list of the actions carried out by the user for an observation period, which can subsequently be used by automatic analysis methods to generate user preferences. Usage scenarios include tracking and monitoring the content viewed by individual members of a household and building a personalized TV guide by tracking user viewing habits. It is further possible to provide, similar to MPEG-7, segmentation metadata which supports the ability to define, access, and manipulate temporal intervals (i.e., segments) within an audio-visual (AV) stream. By associating metadata with segments and segment groups, it is possible to restructure and re-purpose an input multimedia stream to generate alternative consumption and navigation modes. This is useful, for example, to construct a summary of the content with highlights or a set of bookmarks that point to "topic headings" within the stream.

A normative TV-Anytime Content Referencing Specification was issued to allow acquisition of a specific instance of a specific item of content. This ability is needed to refer to content (in example, a series of programs) independent of its location, whether that location is on a particular broadcast channel on some date and time or on a file server connected to Internet. It should also be noted that the TV-Anytime Content Referencing Identifier (CRID) syntax has been specified by the IETF recently as RFC 4078, which will help to propagate this content referencing scheme to many Internet connected devices.

The TV-Anytime System Description Specification allows an application to "show" the system behavior of a TV-Anytime broadcast system with an interaction channel used for consumer response. It focuses on the use of the TV-Anytime content reference specification in combination with the TV-Anytime metadata specification in a system context.

ARIB (in Japan), ATSC and CEA (in the U.S.), DVB (in Europe), and others are working on the adoption of TV-Anytime in their respective environments. This process is being supported through cross membership of the respective groups. Liaisons have also been established with the EBU, MPEG, and Pro-MPEG to continue the ongoing harmonization effort on user profiling.

### C. MPEG-21 Digital Item Declaration and Identification

MPEG-21 aims at defining the technology needed to support users to exchange, access, consume, trade, and otherwise manipulate digital media in an efficient, transparent, and interoperable way [11]. This open framework is based on two essential concepts: the definition of a fundamental unit of distribution and transaction [the Digital Item (DI)] and the concept of Users interacting with Digital Items. MPEG-21 identifies and defines the mechanisms and elements needed to support the multimedia delivery chain as well as the relationships between and the operations supported by them. Within the parts of MPEG-21, these elements are elaborated by defining the syntax and semantics of their characteristics, such as interfaces to the elements.

A DI as defined in MPEG-21 [19] is essentially a versatile "virtual container" for metadata, structure, and content (called resources in MPEG-21 terms). Based on this central notion of a Digital Item, MPEG-21 standardized a rich delivery framework that shifted away from the specification of bitstream syntax and semantics, decoder behavior (as in MPEG-1, MPEG-2, and MPEG-4), and multimedia description tools (as in MPEG-7) towards a higher level framework that supports multimedia interaction and dynamic content.

It may be useful to contrast a DI with a familiar HTML web page. The primary distinction with an HTML Web page is that the purpose of the underlying structure of a DI is aimed purely at declaring its constituent parts. In contrast, the HTML web page structure aims at marking-up text and resource content for presentation purposes. DIs are not required to contain information (i.e., provided by an author) on how the resources and metadata should be presented. As such, DIs may operate in application areas where there are agreed upon rules for presentation or may contain presentation descriptions as resources. Another important functionality of the DI is its ability to configure itself, i.e., taking into account usage environment, terminal conditions, and network conditions. This supports transparent and augmented use of DIs across a wide range of networks and end-user devices.

The means to declare the structure of a DI is provided by the Digital Item Declaration (DID) specification. The DID expresses and identifies the resources (e.g., MPEG-4 files) and metadata (e.g., MPEG-7 or Dublin Core descriptions) which the authors consider to be constituents of the DI; the DID facilitates a structuring of the resources with the metadata for the whole DI. In addition, the DID binds together single entities and groups of resources and metadata also allowing the metadata to be connected to certain fragments in a media source.

The MPEG-21 Digital Item Identification (DII) framework addresses a key issue in multimedia communications, namely the ability of a device to uniquely identify various parts of digital objects and items. This is particularly important metadata in domains where digital information is being copied and manipulated. The role of an Identifier in MPEG-21 is to identify resources and the intellectual property related to the Digital Items (and parts thereof). Identifiers may also be used to link Digital Items with related information such as descriptive metadata and to identify different types of Digital Items. The key to MPEG-21 usage is the fact that MPEG-21 can integrate existing identifiers used in a particular application domain. Hence, the DII part in MPEG-21 concentrates on how to integrate existing schemes into the MPEG-21 framework.

## IV. MULTIMEDIA METADATA STANDARDS: DESCRIBING RIGHTS AND PROTECTION

Rights and protection information are becoming increasingly important for the management of multimedia data, most notably due to emerging business models that distribute multimedia over the Internet or via broadband networks. As noted in Table 1, content management metadata generally deals with the expression of rights, protection metadata, and governance. This section will address some of the most relevant metadata standards that target the interoperable description of content rights and protection features, notably those developed in MPEG. In Section VII-B, the role of standardized content management metadata in the context of interoperable DRM systems is discussed.

Other interesting developments in this area come from: 1) Organization for the Advancement of Structured Information Standards (OASIS) which, e.g., developed "a digital rights language that supports a wide variety of business models and has an architecture that provides the flexibility to address the needs of the diverse communities that have recognized the need for a rights language" [20]; 2) TV-Anytime [7] which established a means of securely enabling consumer content usage while providing standardized interfaces to legacy conditional access and content protection systems; 3) International Digital Publishing Forum (IDPF), previously Open eBook Forum, which, e.g., developed "epub"—a file extension of an XML format for reflowable digital books and publications that allows publishers to produce and send a single digital publication file through distribution and offers consumers interoperability between software/hardware for unencrypted reflowable digital books and other publications [21]; 4) Digital Media Project (DMP) which targets the "development, deployment and use of digital media that respect the rights of creators and rights holders to exploit their works, the wish of end users to fully enjoy the benefits of digital media and the interests of various value-chain players to provide products and services" [22], and Open Mobile Alliance (OMA) which specifies "standardized DRM solutions for content services across mobile networks, but in a network and content agnostic manner, which can then be used for any content and in a wide variety of environments, services and devices" [23].

### A. MPEG-21 Rights Expression Language

The MPEG-21 Rights Expression Language (REL) [11], [24] is an XML-based, machine-readable language providing a method for specifying rights and conditions associated with the distribution and use of assets like content, e.g., "Rob permits Alan to play a particular movie for one week if he pays $10." The REL provides a standard, precise, flexible, extensible, and rich way to express grants of rights. It is agnostic to types of assets, platforms, and media, and expressive enough to support applications that can go beyond digital rights management (DRM), e.g., privacy protection, playing the role of a generic authorization language. With this purpose in mind, the REL:

1) defines syntax and semantics of a machine interpretable language that can be used to specify rights unambiguously that apply throughout the content's life cycle;
2) provides an authorization model to determine if an authorization or access control request can be granted the right to perform an action on a resource according to REL expressions;
3) supports many business models in the end-to-end distribution value chain.

The REL has no intention of replacing legal rights and does not specify how and when rights should be created, communicated, audited, and enforced. It is based on eXtensible rights Markup Language (XrML) 2.0 [25] which was selected by MPEG for its expressiveness and unambiguity over Open Digital Rights Language (ODRL) [26].

The REL data model is based on the notion of grant, see Fig. 4. A license is a container of grants, issuers, and some other related information and it is the central (metadata) entity in the model. A license conveys that an issuer authorizes rights in the forms of grants (one or more). A grant implies giving to a claimant something that could be withheld. The MPEG-21 REL models a grant as "Principal+Right+Resource+Condition" with each of these terms defined as follows.

1) *Principal* refers to the identification of a party (User) to whom a *Right* is being granted.
2) *Right* relates to the usage of the Digital Item (content, metadata, and associated structure) that one User provides to another, and thus the action that a *principal* can be granted to exercise against some *resource* (content) under some *condition*.
3) *Resource* regards the identification of the 'object' (DI or part thereof) to which a *principal* can be granted a *right*.
4) *Condition* specifies the terms, conditions, and obligations under which rights can be exercised.

In summary, a grant specifies that a *Principal has a Right over a Resource under certain Condition*. With this model, very flexible rights metadata expressions can be generated allowing the machine-processable exchange of rights, e.g., using agents. The adoption of common rights metadata along the content life-cycle is important not only for interoperability reasons but also because rights metadata will naturally be changed and manipulated as content moves along the multimedia chain.

Closely coupled with the REL, the MPEG-21 Rights Data Dictionary (RDD) is a dictionary of key *terms* which are required to describe rights of Users and provides a

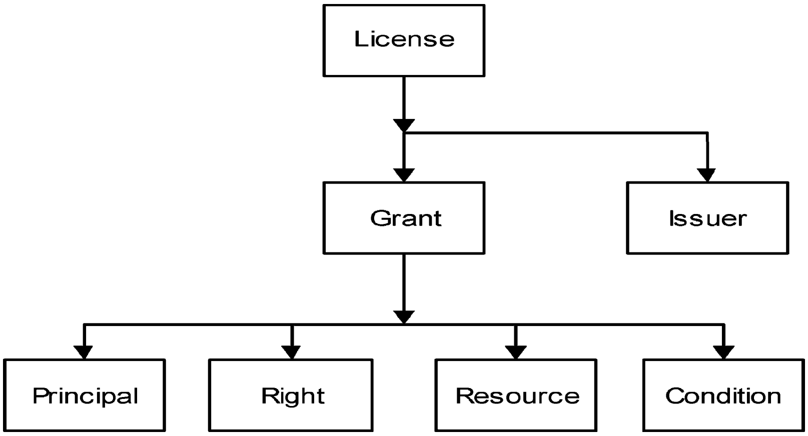


**Fig. 4.** ***MPEG-21 REL data model.***

mechanism for the extension to further terms through a registration authority. The RDD includes:

1) a 2000 terms data dictionary based on a logical structure, event-based data model—the Context Model—which is used to construct a natural language ontology for terms in rights management;
2) methodology for continuing extensibility;
3) management mechanism by ISO Registration Authority.

The RDD supports interoperability of meaning (semantics) for the MPEG community. There are 14 RDD "ActTypes" Terms (verbs) defined to support the REL specification; these verbs are the baseline actions that can be used in REL grants. Together, the REL and the RDD modularly provide metadata tools at different layers which are essential to build a harmonized and integrated complete metadata framework.

### B. MPEG-21 Intellectual Property Management and Protection Components

The MPEG-21 intellectual property management and protection (IPMP) components standard [11] provides a way (notably metadata) to include protected and governed content in an MPEG-21 DI which is the basic multimedia content entity in the MPEG-21 framework. Its objectives are to:

1) provide a mechanism for Users to protect a DI and its declaration using a specified protection scheme;
2) provide metadata to express governance, e.g., rights, over a specific part of a DI hierarchy to be governed while maintaining the transactability and schematic validity as a DI;
3) allow DIs to be used in conjunction with many DRM schemes;
4) offer a degree of interoperability between schemes by acting as "rich" metadata containers which can describe content and its availability in one of several DRM formats.

While the IPMP Components standard does not specify a full DRM system, and thus, on its own, cannot make IPMP systems interoperable, it does allow terminals to be able to understand how to process protected content, and, with the appropriate rights (and business agreements), it will permit interworking between systems. The MPEG-21 IPMP Components standard includes two main tools.

1) *IPMP Digital Item Declaration Language (IPMP-DIDL)*—Specifies a way to include protected content in an MPEG-21 DID document, see Fig. 5. In particular, a schema is specified that provides a means to attach protection metadata to a specific part of the DI hierarchy. This facilitates the representation of a protected DI structure within a DID document by encapsulating protected DIDL elements and linking appropriate

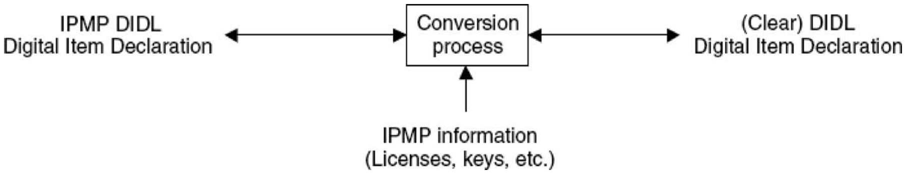


**Fig. 5.** ***Relation between DIDL and IPMP-DIDL.***

*IPMP Information* to them, thereby allowing for encryption and other forms of protection over the DIDL hierarchy while maintaining the DI as an exchangeable container. The description of IPMP governance and tools is required to satisfy IPMP conditions for a DI or its parts to be accessed.

2) *IPMP Information*—Provides information about the protection of elements in a DI (the metadata itself), expressing governance in a flexible and extensible manner. It defines structures for expressing information relating to the protection of content, including tools, mechanisms and licenses by:
   - *IPMP General Info Descriptor*—Contains general information about IPMP tools and rights expressions relating to a complete DID;
   - *IPMP Info Descriptor*—Contains the description of the specific IPMP governance and tools applied to a certain part of a DI hierarchy protected with IPMP DIDL, that is, the specific tools applied, keys, a license specific to that content, and so on.

While managing governance and thus rights is a difficult issue, especially in the standardization arena due to the dominance of large companies and industry with conflicting interests involved, governance metadata is central for effective machine-readable exchange of multimedia content. Content management metadata is clearly an area where standardization has to move forward while carefully considering the impacts of increased interoperability in business models and trust management.

## V. MULTIMEDIA METADATA STANDARDS: DESCRIBING CONTEXT

While content descriptions certainly provide useful information for the retrieval and delivery of content, a description of the usage environment or context is also necessary so that the original source content could be matched with its final destination (see Fig. 6). Generally speaking, the usage environment covers a wide range of factors that might affect the optimal means by which content is ultimately consumed. This section focuses on the most important context factors, including terminal capabilities, network characteristics, user characteristics, and natural environment characteristics.

The most relevant metadata standards in this space include the MPEG-21 Digital Item Adaptation (DIA) standard [27], [28], which specifies a rich set of tools based on XML to enable multimedia adaptation, and the composite capability/preference profiles (CC/PP) developed by the Device Independence Working Group of the W3C Consortium, which specifies a structure and vocabularies for device capabilities and user preferences based on the resource description framework specification (RDF) from W3C [29]. CC/PP is also the basis for the User Agent Profile (UAProf) specification [30] of the OMA, which specifies hardware and software characteristics of the device as well as information about the network to which the device is connected and application/user preferences.

Since the primary aim of this section is to underscore the general utility of context metadata for multimedia retrieval and delivery, a detailed and exhaustive comparison of the different standards in this space is not given.

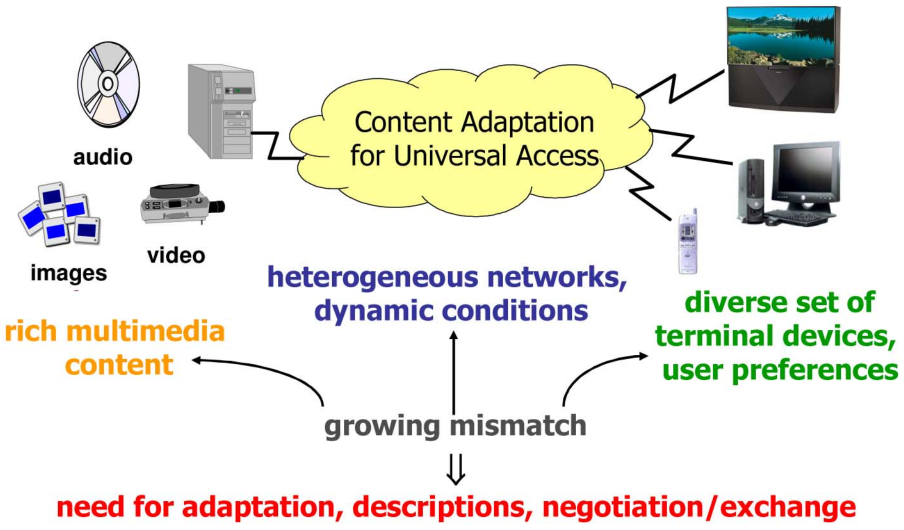


**Fig. 6.** ***Concept of universal multimedia access, related to a growing mismatch between multimedia content sources and terminal/network capabilities.***

Instead, a select set of metadata fields in the categories that are most relevant to multimedia retrieval and delivery is covered, with a focus on elements that have been standardized as part of MPEG-21 DIA. It should be noted that while DIA and CC/PP differ in their solutions (e.g., XML-based versus RDF-based, coverage of certain device attributes, etc.), the general approach is essentially the same. That is, both specify an explicit description of the context that is negotiated and exchanged. In contrast to this approach, the Digital Living Network Alliance (DLNA) has considered an alternative approach, which is to specify context profiles with mandatory support for certain media formats. This approach is currently favored in the home networking industry and will also be discussed as part of this section.

### A. MPEG-21 Digital Item Adaptation

The MPEG-21 DIA specification provides a very rich set of metadata tools to guide the multimedia adaptation process and facilitate access to a diverse usage environment, where this usage environment comprises terminal capabilities, network characteristics, user characteristics and natural environment characteristics. In addition to these usage environment description tools, DIA also specifies a number of tools that enable one to formulate explicit limitation and optimization constraints. In this way, additional guidance could be provided to an adaptation engine in a standardized way so that a more satisfactory adaptation could be determined and/or to limit the space of feasible adaptations so that the required effort to search for an optimal solution is reduced. The standard also specifies a means to describe the relationship between the above constraints, the feasible adaptation operations satisfying these constraints and associated utilities that result from a particular adaptation. Further details on such an adaptation framework may be found in [31] and [32]. In the following, a brief overview of selected usage environment description tools and a discussion of their relevance for multimedia retrieval and delivery is provided.

Terminal capabilities are characterized in terms of both receiving and transmitting capabilities. Such a description is used to satisfy consumption and processing constraints of a particular terminal. Important attributes include codec capabilities, input–output characteristics, and other device properties, such as CPU characteristics. These various description categories include the following.

1) *Codec Capabilities*—Specify the format a particular terminal is capable of encoding or decoding, e.g., an MPEG-X profile@level. Given the variety of content representation formats that are available today, it is not only necessary to be aware of the formats that a terminal is capable of handling, but it is sometimes important to also know the limits of specific parameters that affect the operation of the codec. In MPEG standards, the level definition often specifies such limits. However, it is possible that some devices are designed with further constraints or that no specification of a particular limit even exists. Therefore, the codec parameters as defined by MPEG-21 DIA provide a means to describe such limits, e.g., the maximum bit rate that a decoder could handle.
2) *Input–Output Capabilities*—Include a description of display characteristics, audio output capabilities, and various properties of several types of input devices. Describing the capabilities of a display is obviously very important as certain limitations that impact the visual presentation of information must be taken into consideration, such as the resolution, the color capabilities, and rendering format. The same is true for audio output devices, where descriptions of frequency range, power output, signal-to-noise ratio, and the number of output channels, are described. Finally, user interaction inputs define the means by which a user can interact with a terminal. With such information, an adaptation engine could modify the means by which a user would interact with resources. For instance, knowing whether a terminal has the ability to input information through a keypad or microphone may affect the interface that is presented to the user.
3) *Device Properties*—Characterize power-related attributes of a device, as well as storage, data IO characteristics, and CPU benchmarks. A description of the power characteristics provides information pertaining to the consumption, battery capacity remaining, and battery time remaining. With such attributes, a sending device may adapt its transmission strategy in an effort to maximize the battery lifetime. Storage characteristics are defined by the input and output transfer rates, the size of the storage, and an indication of whether the device can be written to or not. Such attributes may influence the way that content is retrieved, e.g., whether it needs to be streamed or could be stored locally. To gauge computational performance, a benchmark-based description might also be useful, where the CPU performance is described as the number of integer or floating-point operations per second. With such a measure, the capability of a device to handle a certain type of media, or media encoded at a certain quality, could be inferred.

Two main categories are considered in the description of network characteristics: capabilities and conditions. The capabilities define static attributes of a network, while the conditions describe dynamic behavior. These descriptions primarily enable multimedia adaptation for improved transmission efficiency.

1) *Network Capabilities*—Include attributes that describe the maximum capacity of a network and the

minimum guaranteed bandwidth that a network can provide. Also specified are attributes that indicate if the network can provide in-sequence packet delivery and how the network deals with erroneous packets, i.e., does it forward, correct, or discard them.

2) *Network Conditions*—Specify attributes that describe the available bandwidth, error, and delay. The error is specified in terms of packet loss rate and bit error rate. Several types of delay are considered, including one-way and two-way packet delay, as well as delay variation. Available bandwidth includes attributes that describe the minimum, maximum, and average available bandwidth of a network. Since these conditions are dynamic, time stamp information is also needed. Consequently, the start time and duration of all measurements pertaining to network conditions are also specified. However, the end points of these measurements are left open to the application performing the measurements.

User characteristics play an important role in the way that content might be filtered or customized. In the context of multimedia retrieval and delivery, user preferences might suggest a preferred format for different classes of devices. MPEG-7 has standardized a collection of metadata related to user preferences. The basic data types have also been adopted by other standards including TV-Anytime [14] and MPEG-21 DIA [28]. It should be noted that the standards do not specify how such user data is collected, but one could imagine that if they are not provided directly to, e.g., a service provider or device, that they could be collected in an automated and transparent means. As this data is often considered private, the distribution and maintenance of such data needs to be carefully handled, and content management metadata and the associated protection tools as described in Section IV have an important role to play here.

1) *User Interaction*—Describes preferences of users pertaining to the consumption of the content, as well as usage history. The MPEG-7 content descriptions can be matched to the preference descriptions in order to select and personalize content for more efficient and effective access, presentation, and consumption.

2) *User Preference*—Describes preferences for different types of content and modes of browsing, including context dependency in terms of time and place. It is also possible to indicate a weight that corresponds to the relative importance of different preferences, the privacy characteristics of the preferences, and whether preferences are subject to update, such as by an agent that automatically learns through interaction with the user.

3) *Usage History*—Describes the history of actions carried out by a user of a multimedia system. The usage history descriptions can be exchanged between consumers, their agents, content providers, and devices, and may in turn be used to determine the user's preferences with regard to content.

Finally, there are several description tools that describe aspects related to the natural environment. The main purpose of these tools is to enable multimedia adaptation according to particular location, time, or audio–visual environment.

1) *Location and Time*—Refers to the location and time of usage of content, respectively. Both description tools make use of MPEG-7, in particular the Place DS and Time DS. Besides being stand-alone tools, both tools are utilized in the specification of user characteristics as well.

2) *Audio–Visual Environment*—Describes audio–visual attributes that can be measured from the natural environment and affect the way content is delivered and/or consumed by a user in this environment. For audio, the description of the noise levels and a noise frequency spectrum is specified, while illumination characteristics that may affect the perceived display of visual information are specified for the visual environment.

## B. DLNA Media Format Profiles

In contrast to metadata standards that explicitly describe contextual information, DLNA has released design guidelines to achieve interoperability among home devices, which among other things deals with media format interoperability [33]. The approach that has been taken is to define a set of *mandatory* media format profiles that all devices within a device class or category are required to support. According to the guidelines, a profile defines the combination of AV compression formats, media-specific attributes and parameters, as well as a system level format and any other information that would sufficiently describe the encoded content. The intention of such a model is to achieve a baseline format for home network interoperability. Optional media formats are also specified to allow for broader support of other popular media formats.

The latest volume of media format profiles as defined by DLNA specifies the detailed guidelines to enable interoperability between DLNA devices in the digital home [34]. An example profile would define the AV media formats as well as the encapsulation or system layer format. For instance, a profile ID of `AVC_MP4_BL_L2_CIF30_AAC` indicates that the video coding format is compliant to H.264/AVC Baseline (BL) Profile at Level 2. The picture resolution is CIF (352 × 288) and the maximum frame rate is 30 Hz. The audio format is AAC with a maximum bit rate specified by DLNA as 128 kb/s. The encapsulation for this DLNA profile is designated as MP4 meaning the MPEG-4 file format is used. As one could imagine, given all the possible AV media and encapsulation formats, the total number of profiles is potentially quite large; therefore, the guidelines specify only the most practical and useful combinations for given regions of the world. Each device class is then designated sets of mandatory and optional formats based on the media format profiles that have been defined. As new media or transport formats become available, the DLNA guidelines would need to be extended accordingly to define new profiles.

In order to support interoperability between devices of different classes, it is expected that some "translation" between the required media format profiles of different device classes would be needed. Additionally, DLNA specifies rules about conversion between optional and mandatory formats to ensure that content can be enjoyed on all compliant devices. Interoperable DLNA devices have been demonstrated in major trade shows and are expected to penetrate the market in the next year or two.

While DLNA is expected to be successful in the home networking and consumer electronics market and OMA devices are currently utilizing the UAProf specification in the mobile domain, there do exist greater needs for context metadata as devices begin to connect with media outside their current domains. Of course, this is unnecessary if all devices could conform to the same specifications and guidelines and proprietary solutions did not exist. Unfortunately, this is far from today's case; it is still very difficult to network consumer electronics and mobile devices, for instance, and allow content playback between different service domains. Context metadata is essentially a bridge between these different worlds and will become more essential as the demand for media connectivity between devices grows. While this may take a number of years for the industry to resolve, it is likely that pockets of interoperability between different domains begin to emerge before a complete and open solution is achieved.

## VI. MULTIMEDIA METADATA STANDARDS: NEW DIMENSION

The colossal evolution of the digital multimedia landscape in the past 10–15 years has brought the complexity of multimedia applications from simple, almost stand alone, tools to sophisticated sets of tools that provide a number of intimately related functionalities. As a consequence, today's users need well integrated packages of tools—super-formats—that provide complete solutions, for example, coding formats combined with content metadata and IPMP tools, rather than fragmented tools for which the ideal combination and integration has to be found. This trend is especially meaningful for metadata and metadata standards since they are a hard sell on their own, if not somehow combined or linked with the data they describe. This motivation was also behind the development of the Advanced Authoring Format (AAF) and Material eXchange Format (MXF) [35] developed by various organizations such as SMPTE and EBU. For example, the MXF was developed to create a universal format to exchange media files with associated data and metadata between otherwise incompatible systems; it was designed to work across networks with servers, workstations, and other digital media devices. In this context, MPEG also felt the need to develop standard packages of technologies providing the industry with solutions better fitting the user needs, the so-called multimedia application formats (MAFs).

Before proceeding with a more detailed description of MAFs, it is worth pointing out that the definition of technology packages for particular application domains is especially important for metadata since specific ontological elements, and thus semantics, may be added to enrich the ontology. Such extensions in the semantic dimension should be possible without changes to a baseline specification, which could be, for example, more focused on the syntactical dimension. It is even possible to think about technology packages across (metadata) standards developed by different standardization bodies, e.g., MPEG and W3C, to bring an even larger scale of interoperability. While there may be other complications, having a clean ontological mapping between different standards would certainly be a major advance.

### A. MPEG Multimedia Application Formats

Until recently, MPEG standards basically defined tools addressing well identified user requirements; these tools were typically clustered depending on their functionalities and related content types, e.g., MPEG-X Systems, MPEG-X Video/Visual, MPEG-X Audio. The combination of MPEG tools to build multimedia applications was left to the companies developing the applications and products and, eventually, to the users. Most of the time, industry consortia were created (outside MPEG) with the target to define the combination and integration of the tools adequate for a specific application domain. The farthest MPEG typically went in terms of specific application domains was through the definition of profiles and levels addressing specific classes of applications, e.g., systems only, video/visual only, and audio only profiles and levels; never combinations of them [36].

With the increasing complexity, sophistication, and deployment speed of multimedia applications, MPEG finally recognized that the past approach of letting adopters of MPEG standards define the best combination of tools to use, within and across MPEG standards, was not only preventing interoperability but also impeding the usage of MPEG tools, notably metadata tools. This was especially critical for products developed by smaller companies, which without having the resources of big companies, e.g., to participate in the standardization process itself, could hardly define the best combination of MPEG tools for their business.

To address this *status quo* and provide the users a new layer of standards and thus interoperability, MPEG decided to define a new type of standard, know as MPEG-A, which specifies the so-called MAFs stipulating a combination of already tested and verified tools taken from the entire MPEG standards body and providing an appropriate global technical solution for a class of applications. A given MAF uses tools and profiles from selected MPEG standards (or parts of standards) and combines them into a single standard. Ideally, a MAF specification consists of references to existing profiles within MPEG standards and does not specify by itself new technology. MAFs offer another standardization model where MPEG provides the users with packaged solutions, including coding and metadata formats combinations, without waiting for industry consortia to define these combinations with all the disadvantages this implies. To better meet this goal, the MPEG-A standard specifies not only the multimedia application format itself but it also provides the related software implementation. The software demonstrates how MAFs are used and offers vendors an easier start for developing multimedia products. MPEG's ultimate objective for MAFs is to stimulate the increased use of MPEG technology through additional interoperability of different media at the application level.

This new MPEG standardization dimension seems particularly relevant for the future deployment of the MPEG-7 and MPEG-21 standards, notably to overcome some of the issues mentioned at the end of Section III-A. The specification of standard super-formats where metadata tools appear in combination with other multimedia tools, notably for coding and protection, playing a central role in the functionalities provided, may be an important step to stimulate the adoption of the main MPEG metadata standard.

### B. Music and Photo Player MAF Cases

Among the various MPEG MAFs already defined, there are a couple which can very well demonstrate the benefits of this new standardization approach: the Music and the Photo Player MAFs. While the Music Player MAF targets interoperability for digital music libraries in which each music asset is defined as a combination of audio, metadata, and images (for example, the cover image associated with the relevant music recording), the Photo Player MAF has a similar target for digital photo libraries. For both scenarios, a standard file format is required to allow for easy management and organization of digital content for exchange, browsing, retrieval, categorization, etc. This type of packaging standard is believed to be very important to accelerate the large adoption of metadata standards which is still small.

The Music Player MAF will give users the capability of handling audio data, metadata, and images for individual pieces of music as well as for entire collections of music, such as complete albums or playlists. Similarly, the Photo Player MAF allows the users to wrap a collection of JPEG photos with associated metadata into a single file that can be easily exchanged among users and between diverse digital devices, such as digital cameras, PDAs, camera phones, personal computers, and portable media players; users can also define subcollections supporting different ways to organize and play the content, for example based on people present, events, and places (categories may be hierarchical and can be defined by the users).

The Music Player MAF defines three file formats—song, album, and playlist—which allow containing a single music track (MP3) with associated metadata and a single (JPEG) image, and create complete album and playlist files based on song files. While the song file is based on the MPEG-4 file format, the album and playlist files are based on the MPEG-21 file format and the MPEG-21 DID [11]. The MPEG-4 file format supports the storage of metadata associated to a data track; the associated metadata describing the audio track, like artist or song name, is expressed in MPEG-7 MDS tools [12]. Since the MP3 bitstream files can contain associated metadata, typically ID3 tags [1], a specific ID3-MPEG-7 MDS mapping has been defined for this purpose in the MAF specification.

In terms of metadata, the Photo Player MAF goes beyond the Music Player MAF since it wraps (in MPEG-4 files) a compact set of MPEG-7 MDS and Visual metadata as well as Exif [37] metadata (mapped into MPEG-7) since this metadata is commonly used in digital cameras and thus available with JPEG images. Two different subsets of MPEG-7 metadata exist in the Photo Player MAF: collection-level and item-level metadata. Each subset shall be binarized according to the MPEG-7 BiM format [12], using a corresponding simplified version of the MPEG-7 schema. Photo-player devices can be implemented with either a fixed binary encoder/decoder, according to syntax defined in the MAF, or can include a full BiM encoder/decoder, which infers the binary syntax from the XML Schema. The collection-level descriptive metadata gathers information about the collection such as the creator, creation time, last update, and the name of the collection. The item-level descriptive metadata gathers information about each item in the collection; both MPEG-7 MDS and MPEG-7 (low-level) visual descriptors can be used, notably the dominant color, scalable color, color layout, color structure, edge histogram, and homogeneous texture descriptors; the inclusion of low-level visual descriptors allows performing advanced content-based search and retrieval such as query by example, photo categorization, and situation-based clustering. This is the first time a standard package format includes low-level metadata which is a rather big step for metadata standards.

To increase the value of these MAFs and address more business models, the Music Player MAF has already specified a specific solution to address the important music industry requirement of content protection. In the Music Player MAF, protection is provided in two "flavors" which derive from different understandings of interoperability in the context of protected music. In one flavor, the music tracks, images, and metadata are wrapped in a MPEG-4 file protected with fixed encryption (AES 128), without key management components, while in the other flavor music tracks, images and metadata are wrapped in a MPEG-21 file protected with a flexible tool selection and key management components. Protection is an issue currently under consideration for the Photo Player MAF; it is expected that the protection tools for this MAF will be aligned with those in the Music Player MAF.

## VII. METADATA STANDARDS: FUTURE NEEDS AND CHALLENGES

This paper has reviewed a number of metadata standards related to content description, context description, as well as the description of rights and protection. Standardized application formats that attempt to combine various tools, notably metadata tools, into a single interoperable format for particular classes of applications have also been discussed. Considering the aims of this effort, which is mainly to demonstrate the utility of metadata standards with traditional AV playback components, it is worth examining the future needs for metadata standards. In the following, a brief outline of such needs and future challenges in the context of the interoperable metadata scenarios discussed in Section II, and the various metadata standards described in Sections III–VI, is provided.

### A. Ease of Use

Internet users have become very accustomed to searching and browsing on the web from a PC. Similar functions are already being supported on mobile devices, television sets, automotive systems, etc. However, current non-PC search and retrieval systems are rather inefficient in performing this task. There is a number of reasons for this ranging from a poor user interface to unnatural methods of input. It should be clear that the user interface and method of interaction to perform search and retrieval functions efficiently needs to change significantly with each environment. This implies a tighter coupling between the metadata used to facilitate search and retrieval functions with the interface and mode of interaction. In the following, we examine a couple of different ways that future metadata standards could be tailored to improve the ease of use in diverse usage environments.

As an example, consider multimedia search and retrieval functions in an automobile, e.g., while on a long roadtrip or in an unfamiliar city. The tasks at hand are to locate nearby seafood restaurants, browse the menus of these restaurants, access any available reviews, possibly make a reservation at the desired place, and finally get directions to the target location. While these functions would be routine and easily executed from a PC, the automobile environment is a far more challenging setting since the modes of input and consumption are very limited, especially for drivers. Assuming these functions are supported by an advanced form of car navigation units, it should be clear that text-based entry would not be optimal towards achieving the desired goals. In contrast, a speech-based interface would be much more accommodating, provided that the robustness of speech recognition engines is sufficiently high. One standard that might support such a direction is VoiceXML [38], which is a markup language for creating voice user interfaces that use automatic speech recognition and text-to-speech synthesis. VoiceXML is essentially a standard dialog design language that developers could use to build conversational applications. One drawback of this approach is that the dialog tends to be more rigid and structured and does not allow for a more free-form Google-like search process of existing web content. To enable a more flexible and loosely structured process, future metadata standards might be tailored to operate on the basis of speech primitives, such as phonemes. In this way, speech-based queries could be better matched with speech-based metadata that pertains to the content, which in our example would be restaurant names, menu items, information about ratings, etc. Similar challenges for search and retrieval of multimedia information as described in this example also exist in the mobile and television environments, e.g., see the work of Wittenburg, *et al.* [39] on applying speech-based query to EPG search on the television. While the solutions may be different depending on the context and type of media to be consumed, it is necessary to identify common requirements for future metadata standards that enable richer forms of multimodal input, such as voice and gestures, to be used for multimedia information retrieval as illustrated in Fig. 7. In 2002, the W3C launched an activity on multimodal interaction, which aims to allow users to dynamically select the most appropriate mode of interaction for their current needs and improve ease-of-use. Interested readers are referred to the documents and reports that have been produced by this group, which include various use cases and requirements, as well as the specification of a multimodal architecture [40].

Another very important dimension to the ease-of-use problem is managing the complexity of all the functionality on a given device. In a mobile device for example, search and retrieval could be one of 20 different features and the function might be very difficult to access and use through the complex and embedded menu structure on most mobile phones. As televisions begin to connect to the home network and gain access to content stored on external hard-drives and the Internet, we will also begin to see problems in managing the different tasks and sources of media. To alleviate some of these burdens, a new standard is being developed by CEA that aims to

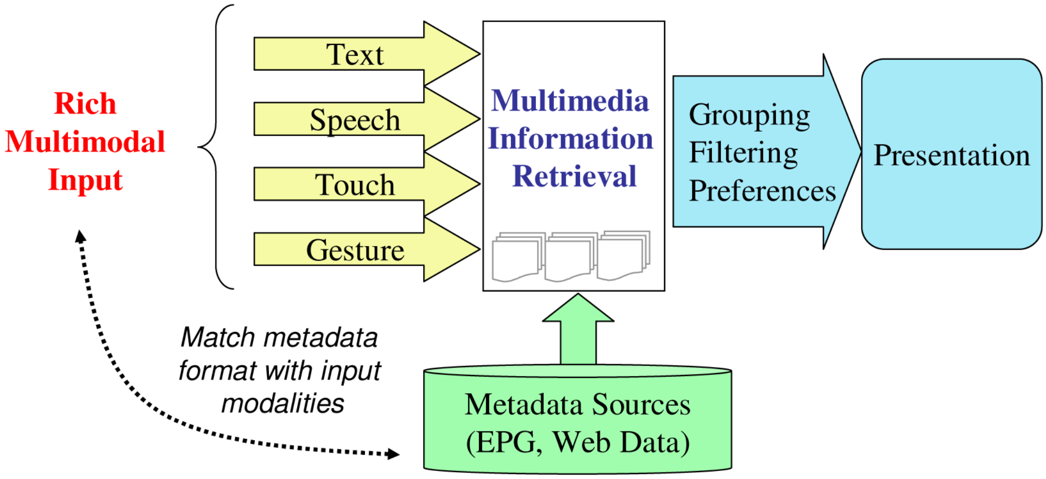

Rich Multimodal Input
Text
Speech
Touch
Gesture
Multimedia Information Retrieval
Grouping Filtering Preferences
Presentation
Match metadata format with input modalities
Metadata Sources (EPG, Web Data)

**Fig. 7.** *Evolution of metadata formats to match with different input modalities.*

specify a representation of task models for common tasks on a particular device and how they are performed [41]. In a task-based interface, the user and system interact primarily in terms of high-level goals, which the system decomposes into primitive actions that are directly supported by the one or more devices involved. Task models could help the search and retrieval process by representing goals more abstractly, i.e., rather than leaving the user to determine how to do a particular search, task-based interfaces would allow the user to interact at the level of what they actually want to do, e.g., search for a particular sports highlight. Task models would also support more natural language interaction since the task models could offer a vocabulary of goals for users to talk with and provide the system with information for how to achieve them, e.g., helping to narrow a search according to specific parameters of interest. Filling the semantic gap between low-level and high-level descriptions has remained one of the key challenges in multimedia retrieval, and it is worth considering whether metadata standards that help to model such relations would reap certain benefits.

## B. Transparency

Interoperability of multimedia retrieval and delivery must be achieved without any notable burden on users. For this to be realized, all processes from search and query to rights management, delivery, and consumption, must work globally and be automated to some extent. While metadata standards described in this paper are one important step towards achieving this goal, there do remain some open problems for seamless and transparent interaction with content.

We first consider the problem of media format interoperability, which is very important given the growing number of media formats and explosion of multimedia devices with varying capabilities. Assume a desired piece of content is located via a search portal, which provides access to contents in distributed locations. If the multimedia that has been located cannot be delivered over the current network or played back on the current device, there is a need for adaptation. If the search portal is equipped with the appropriate codecs, it might be able to pull the content from the source server and perform the conversion in real-time without the user even knowing that such a conversion is being executed. However, if the search portal is not able to perform the conversion, an alternative would be to invoke a service discovery process, where an adaptation service that is capable of transcoding the content according to the provided specifications could be found. This scenario could be enabled using tools specified by OWL-based Web Service Ontology (OWL-S) [42] and MPEG-21 DIA conversion capabilities descriptions [28]. A related problem is determining what types of adaptations are permissible. This is a challenging problem since various attributes of the media/coding format as well as the adaptation process need to be specified in an interoperable manner. The MPEG-21 DIA specification in conjunction with the MPEG-21 REL and RDD specifications provide a solution to this problem as well; finally, governance of these permissions may be performed through the MPEG-21 IPMP Components metadata.

One major problem related to transparency is the lack of interoperability between DRM systems. As one example, consider a cable television system that uses conditional access with certain usage rules. Once the content has been transported through the cable network and stored in the cable set-top box, there is currently no way to access this content and translate the usage rules to another content management and copy protection system. This lack of interoperability prevents wider access to content that the user may be entitled to. Similar cases exist when users subscribe to a particular multimedia service that provides content via the Internet but have limited playback and portability due to noninteroperable DRM between different devices. There are clearly needs for standards in this domain so that a transparent experience and increased portability of contents that are within the user's rights could be enabled. Fig. 8 illustrates the notion of an interoperable DRM layer that essentially performs a translation of the usage rules, which are expressed as metadata, from one system to another. As an example, this would enable content delivered by a cable operator such as Comcast to be retrieved and played on an Apple iPod or Microsoft device.

### C. Application-Oriented Metadata

Metadata for the description of multimedia content tends to be somewhat generic. This is done to enable a wide range of applications with the same metadata format. To be more useful for particular applications, some level of customization seems necessary while still maintaining interoperability. To better understand the implications that this might have on future metadata standards, we consider two distinct applications in the following: surveillance and sports.

In most surveillance applications, retrieval of specific events are likely to require more than what is currently offered by MPEG-7 technology. For instance, a particular application might have the need to identify specific surveillance related events such as "back door alarm triggered" with a given time stamp and key frames around that time. Other applications might require the results of a motion trajectory analysis to be recorded as a set of classes that is useful and well understood in the given application context. In the sports application, there exist a similar set of useful extensions. For instance, recording the formation of the defense in a football game for a particular play is essential for coaches in postgame analysis. Recruiters might be interested in retrieving all the scenes where a designated player identified by the team and jersey number is near the penalty box. Finally, broadcasters have wide range of annotation needs that enable them to retrieve content clips easily at a later time to become part of news segments or half-time analysis reports.

While MPEG-7 or other metadata standards could provide a useful set of descriptions for the two application domains discussed above, it should be clear that many of these requirements are not directly satisfied by existing standards or profiles. For instance, the particular structure of the metadata might not be in an ideal format, certain key elements might be missing, there might be a great deal of unnecessary elements and dependencies associated with a particular metadata tool, etc. At the same time, it would not make sense to define such specific metadata tools for each application and every imagined purpose. One solution would be to profile a useful subset of tools offered by existing standards such as MPEG-7 and then consider extensions of the profiled schema according to specific

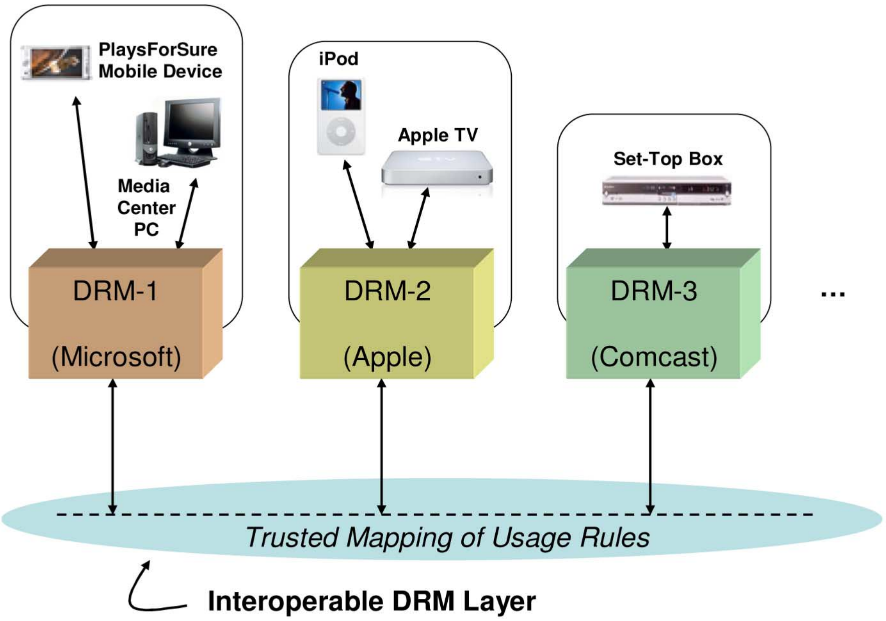


**Fig. 8.** ***Multimedia retrieval and delivery framework with interoperable DRM layer that performs mapping of usage rules (expressed as metadata) in a trusted environment and based on well-defined guidelines and specifications.***

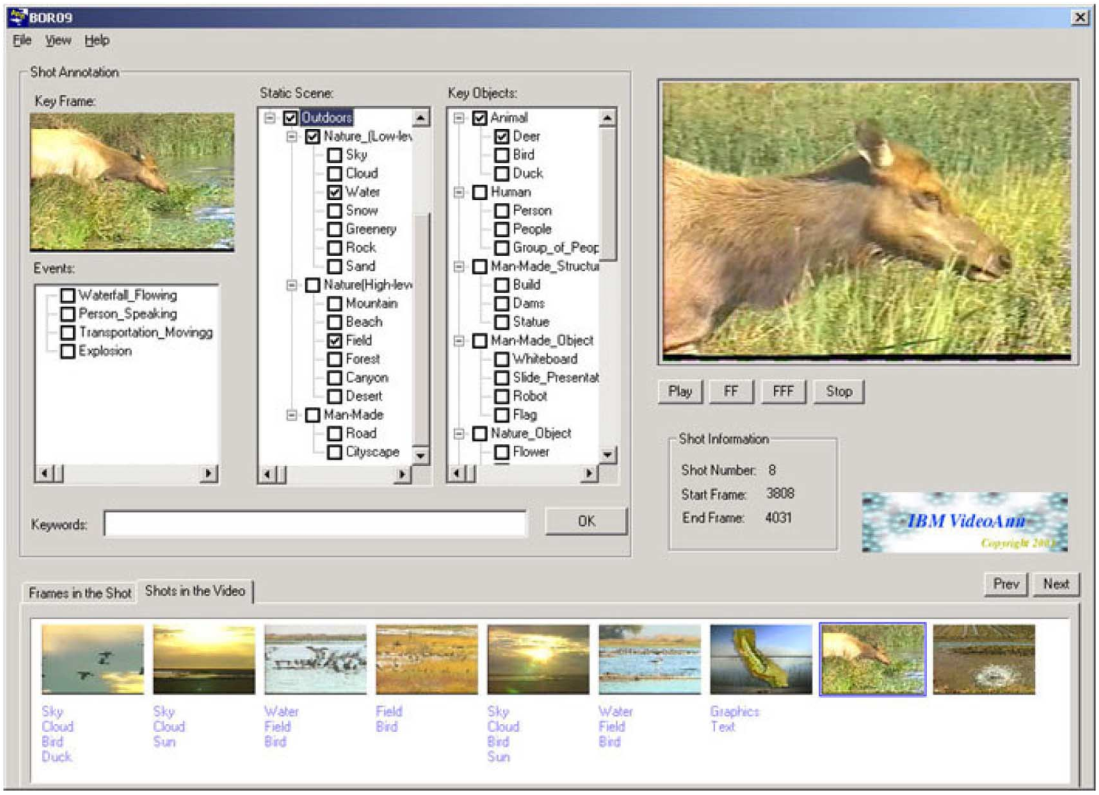
application requirements; these extensions may be defined by MPEG or outside fora that are focused on specific applications and have domain experts in those areas. Since the current MPEG-7 standard offers a very rich and powerful set of description tools, many of which provide a very useful framework for a number of applications, further profiling work is needed based on specific industry needs; this may also involve defining new classification schemes with specific application-oriented controlled terms.

### D. Multimedia Authoring

There is currently an abundance of software available for capturing multimedia, compressing it in a particular format and editing. However, a complete authoring package including interoperable multimedia content descriptions and encapsulation of media and metadata into a standardized file format is needed to make it easier for content producers to generate complete multimedia packages with rich metadata.

One example of this is the annotation tool developed by IBM researchers [43], which accepts MPEG video as input and produces MPEG-7 descriptions for each shot in the video sequence. The annotations include static scene descriptions, key object descriptions, event descriptions, and other lexicon sets. The annotated descriptions are associated with each video shot and are stored as MPEG-7 descriptions in an output XML file. A screen shot of this tool is shown in Fig. 9. Another metadata editing tool has recently been released by NHK in Japan [44], [45]. This tool aims to provide a common platform for generating content-based metadata as well as a means for editing and integrating new feature extraction techniques. The proposed metadata production framework is compatible with the MPEG-7 standard.

Generally speaking, it is necessary for such annotation mechanisms that support interoperable metadata formats to be included as an integral part of multimedia publishing software. In this way, content creators and users would have a natural and streamlined way to associate interoperable metadata with the content being produced, and audio-visual payloads with associated metadata are encapsulated in a single file format, such as the MP4 or MPEG-21 file format, or transport stream, such as the MPEG-2 Transport Stream. It should be noted that combining metadata and media formats is also supported quite naturally by the new MAFs being developed by MPEG. An automated means of multimedia analysis may also be included as part of the software package to enable richer content descriptions without knowledge of the underlying descriptors and description schemes [46]. This



Events:
- □ Waterfall_Flowing
- □ Person_Speaking
- □ Transportation_Movingg
- □ Explosion

Static Scene:
- ☑ Outdoors
  - ☑ Nature_(Low-lev
    - □ Sky
    - □ Cloud
    - ☑ Water
    - □ Snow
    - □ Greenery
    - □ Rock
    - □ Sand
  - □ Nature(High-lev
    - □ Mountain
    - □ Beach
    - ☑ Field
    - □ Forest
    - □ Canyon
    - □ Desert
  - □ Man-Made
    - □ Road
    - □ Cityscape

Key Objects:
- ☑ Animal
  - ☑ Deer
  - □ Bird
  - □ Duck
- □ Human
  - □ Person
  - □ People
  - □ Group_of_Peop
- □ Man-Made_Structu
  - □ Build
  - □ Dams
  - □ Statue
- □ Man-Made_Object
  - □ Whiteboard
  - □ Slide_Presentat
  - □ Robot
  - □ Flag
- □ Nature_Object
  - □ Flower

Keywords:

Shot Information
Shot Number: 8
Start Frame: 3808
End Frame: 4031

Sky
Cloud
Bird
Duck

Sky
Cloud
Sun

Water
Field
Bird

Field
Bird

Sky
Cloud
Bird
Sun

Water
Field
Bird

Graphics
Text

**Fig. 9.** ***Screenshot of video annotation tool (courtesy of IBM [43]).***

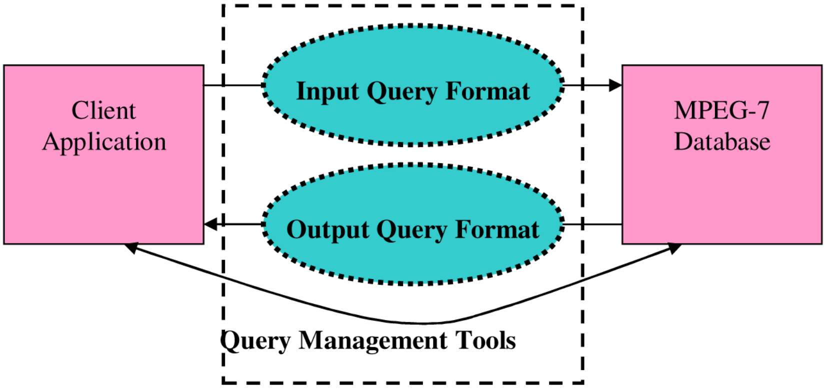


**Fig. 10.** ***MPQF normative elements (in dashed box) [48].***

coupling of both low-level and high-level descriptors provides a mapping of low-level descriptors to semantic concepts and has the potential to substantially improve the retrieval of multimedia information.

### E. Standard Querying

In order to help servers to better cope with the various user queries, an interoperable query format is needed. In addition to this need to standardize the input query format, constraints on the output may also be imposed to limit the response to only that information that the user is interested in or could handle. A prime example of a query format for generic XML documents is the XQuery specification [47], which specifies a language to provide several kinds of expressions that may be constructed from keywords, symbols, and operands; these expressions serve as the query format and are used to guide the retrieval process. Such a need for multimedia data was also felt by MPEG who after developing the MPEG-7 standard found there was a lack of interoperability in terms of the capability to query specific multimedia features of MPEG-7 enabled databases, devices, and applications in an interoperable way. While the current MPEG-7 standard provides tools to describe multimedia content, the interface to support queries in a MPEG-7 database was not defined; since these standard interfaces are not defined, each MPEG-7 database offers its own query interface, which prevents clients from experiencing aggregated services from various MPEG-7 databases. Without a well-defined standardized input query format, users may not be able to access multiple databases easily because they have to tailor their request to match the constraints of each database. Moreover, without a well-defined standardized capability of describing the output query format, which may be also part of the input query, users cannot control or specify the format of the result sets from the various databases.

With these limitations in mind, MPEG decided recently to develop the so-called MPEG Query Format (MPQF) framework which intends to provide a standard format for the requests sent to the server and the response sent from the server and additional tools for query management capability. The final goal is to provide the industry with a unified/standardized way to accept and respond to user requests for multimedia contents searches. The MPQF standard will not specify the behavior of the server because the specific behavior of the server will differ from implementation to implementation. However, the clients using MPQF do expect the servers to understand the received query given in MPQF and to provide the requested data in the requested MPQF output format.

To provide such capabilities, the MPQF framework will include three major normative elements (see Fig. 10) [48].

1) *Input Query Format*—Defines the combination of syntax and semantics for the interface between clients and servers, through which the client provides search criteria and associated data as well as the syntax and semantics of the interface, through which the clients want the server to return the result data.
2) *Output Query Format*—Provides an interface for the response from the server to the client. The greatest part of the response format is defined by the Input Query Format which is sent from the client to the server.
3) *Query Management Tools*—Tools to support the functionality that is required to manage the query transaction between the clients and the servers. The query management tools do not include tools that are supported by network protocols and are intended to be network and media agnostic.

It is expected the MPQF will integrate well with MPEG-7 and be based on XML-related technology. The MPQF specification is planned to be finalized by early 2008.

## VIII. CONCLUSION

Since multimedia data is now available everywhere in growing amounts, metadata is increasingly important for the efficient and effective retrieval, filtering, and management of this content. Moreover, many application scenarios ask for metadata interoperability which raises the need for metadata standards. This paper reviewed the *status quo* in terms of metadata standards to understand the level of maturity of the technology and of its deployment. After studying future needs, trends, and challenges, this paper attempts to project what the next developments will be in terms of metadata standardization. The major conclusion of this paper is that further harmonization among metadata standards is needed and that a modular development approach that targets complementary and application-specific extensions is needed. Authoring and querying in transparent, easy, and more powerful ways are still major issues to be addressed. ■

## Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments which helped to significantly improve the paper.

## REFERENCES

[1] ID3 Home Page. [Online]. Available: http://www.id3.org/

[2] J. R. Smith and P. Schirling, "Metadata standards roundup," *IEEE Multimedia*, vol. 13, no. 2, pp. 84–88, Apr.–Jun. 2006.

[3] The Dublin Core Metadata Initiative. [Online]. Available: http://dublincore.org/

[4] Soc. Motion Picture and Television Engineers. [Online]. Available: http://www.smpte.org/home

[5] European Broadcasting Union. [Online]. Available: http://www.ebu.ch/

[6] Moving Picture Experts Group. [Online]. Available: http://www.chiariglione.org/mpeg

[7] TV-Anytime Forum. [Online]. Available: http://www.tv-anytime.org/

[8] World Wide Web Consortium. [Online]. Available: http://www.w3.org/

[9] International Press Telecommunications Council. [Online]. Available: http://www.iptc.org/pages/index.php

[10] F. Nack, J. van Ossenbruggen, and L. Hardman, "That obscure object of desire: Multimedia metadata on the web, part—2," *IEEE Multimedia*, vol. 12, no. 1, pp. 54–63, Jan.–Mar. 2005.

[11] I. Burnett, F. Pereira, R. Van de Walle, and R. Koenen, Eds., *The MPEG-21 Book*. New York: Wiley, 2006.

[12] B. S. Manjunath, P. Salembier, and T. Sikora, Eds., *Introduction to MPEG-7: Multimedia Content Description Language*. New York: Wiley, 2002.

[13] J. van Ossenbruggen, F. Nack, and L. Hardman, "That obscure object of desire: Multimedia metadata on the web, part—1," *IEEE Multimedia*, vol. 11, no. 4, pp. 38 –48, Oct.–Dec. 2004.

[14] Broadcast and On-line Services: Search, Select, and Rightful Use of Content on Personal Storage Systems ("TV-Anytime"); Part 3: Metadata; Sub-part 1: Phase 1—Metadata schemas, ETSI TS 102 822-3-1, Jan. 2006, version 1.3.1.

[15] CEA-2033, "Open EPG—A specification for electronic program guide data interchange, " 2007.

[16] W3C Multimedia Semantics Incubator Group. [Online]. Available: http://www.w3.org/2005/Incubator/mmsem/

[17] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. New York: Wiley, 2005.

[18] R. Mohan, J. R. Smith, and C. S. Li, "Adapting multimedia internet content for universal access," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 104–114, Mar. 1999.

[19] I. S. Burnett, S. J. Davis, and G. M. Dury, "MPEG-21 digital item declaration and identification—Principles and compression," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 400–407, Jun. 2005.

[20] Organization for the Advancement of Structured Information Standards. [Online]. Available: http://www.oasis-open.org/home/index.php

[21] International Digital Publishing Forum. [Online]. Available: http://www.idpf.org/

[22] Digital Media Project. [Online]. Available: http://www.chiariglione.org/project/

[23] W. Buhse and J. van der Meer, "The open mobile alliance digital rights management [standards in a nutshell]," *IEEE Signal Processing Mag.*, vol. 24, no. 1, pp. 140–143, Jan. 2007.

[24] X. Wang, T. DeMartini, B. Wragg, M. Paramasivam, and C. Barlas, "The MPEG-21 rights expression language and rights data dictionary," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 408–417, Jun. 2005.

[25] XrML-eXtensible rights Markup Language. [Online]. Available: http://www.xrml.org/

[26] Open Digital Rights Language Initiative. [Online]. Available: http://odrl.net/

[27] A. Vetro and C. Timmerer, "Digital item adaptation: Overview of standardization and research activities," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 418–426, Jun. 2005.

[28] Information Technology—Multimedia Framework—Part 7: Digital Item Adaptation, ISO/IEC 21000-7:2007, 2nd ed.

[29] W3C Recommendation, Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0, Jan. 2004.

[30] OMA-UAProf-V2.0, Open Mobile Alliance. [Online]. Available: http://www.openmobilealliance.org/release_program/uap_v2_0.html

[31] S. F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," *Proc. IEEE*, vol. 93, no. 1, pp. 148–158, Jan. 2005.

[32] D. Mukherjee, E. Delfosse, J. G. Kim, and Y. Wang, "Optimal adaptation decision-taking for terminal and network quality-of-service," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 454–462, Jun. 2005.

[33] Digital Living Network Alliance, Home Networked Device Interoperability Guidelines Version: v1.5, vol. 1, Architecture and Protocols, 2006.

[34] Digital Living Network Alliance, Home Networked Device Interoperability Guidelines Version: v1.5, vol. 2, Media Format Profiles, 2006.

[35] MXF Implementation Site. [Online]. Available: http://www.mxf.info/

[36] K. Diepold, F. Pereira, and W. Chang, "MPEG-A: Multimedia application formats," *IEEE Multimedia*, vol. 12, no. 4, pp. 34–41, Oct.–Dec. 2005.

[37] Exif. [Online]. Available: http://www.exif.org/

[38] W3C Recommendation, *Voice Extensible Markup Language (VoiceXML) Version 2.0*, Mar. 2004. [Online]. Available: http://www.w3.org/TR/voicexml20/

[39] K. Wittenburg, T. Lanning, D. Schwenke, H. Shubin, and A. Vetro, "The prospects for unrestricted speech input for TV content search," in *Proc. ACM Advanced Visual Interfaces (AVI)*, Venice, Italy, May 2006.

[40] W3C Multimodal Interaction Activity. [Online]. Available: http://www.w3.org/2002/mmi/

[41] CEA-2018, Task Model Representation, 2007.

[42] D. Martin *et al.*, Ed., OWL-S: Semantic Markup for Web Services OWL-S. [Online]. Available: http://www.daml.org/services/owl-s/1.1/overview/

[43] IBM MPEG-7 Annotation Tool. [Online]. Available: http://www.alphaworks.ibm.com/tech/videoannex

[44] NHK Metadata Production Framework. [Online]. Available: http://www.nhk.or.jp/strl/mpf/

[45] M. Sano, Y. Kawai, H. Sumiyoshi, and N. Yagi, "Metadata production framework and metadata editor," in *Proc. ACM Multimedia Conf.*, Santa Barbara, CA, 2006.

[46] IBM Multimedia Analysis and Retrieval System. [Online]. Available: http://www.alphaworks.ibm.com/tech/imars

[47] W3C Recommendation, *XQuery 1.0: An XML Query Language*, Jan. 2007.

[48] MPEG Requirements, *MPEG-7 Query Format Requirements*, Doc. ISO/IEC MPEG N8509, Hangzhou MPEG Meeting, Oct. 2006. [Online]. Available: http://www.chiariglione.org/mpeg/working_documents.htm

## ABOUT THE AUTHORS

**Fernando Pereira** (Fellow, IEEE) is currently a Professor in the Electrical and Computer Engineering Department, Instituto Superior Técnico, Lisbon, Portugal. He is responsible for the participation of IST in many national and international research projects. He acts often as project evaluator and auditor for various organizations. He has been participating in the work of ISO/MPEG for many years, notably as the head of the Portuguese delegation, Chairman of the MPEG Requirements Group, and chairing many *ad hoc* groups related to the MPEG-4 and MPEG-7 standards. His areas of interest are video analysis, processing, coding and description, and interactive multimedia services. He has published more than 200 papers in technical journals.

Dr. Pereira is an Area Editor of the *Signal Processing: Image Communication Journal* and is or has been an Associate Editor of IEEE Transactions of Circuits and Systems for Video Technology, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, and *IEEE Signal Processing Magazine*. He is a member of the IEEE Signal Processing Society Image and Multiple Dimensional Signal Processing Technical Committee and of the IEEE Signal Processing Society Multimedia Signal Processing Technical Committee. He was an IEEE Distinguished Lecturer in 2005. He has been a member of the scientific and program committees of many international conferences.

**Anthony Vetro** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Polytechnic University, Brooklyn, NY.

He joined Mitsubishi Electric Research Labs, Cambridge, MA, in 1996, where he is currently a Group Manager with responsibility for research and development in the area of multimedia and information coding. He has published more than 100 papers and has been an active member of the MPEG and JVT standardization committees for several years, where he is currently serving as an Editor for the Multiview Video Coding Amendment of H.264/AVC.

Dr. Vetro serves on the program committee for various conferences and has held several editorial positions. He is currently an Associate Editor for *IEEE Signal Processing Magazine* and Chair of the Technical Committee on Multimedia Signal Processing of the IEEE Signal Processing Society. He is also a member of the Technical Committees on Visual Signal Processing and Communications and Multimedia Systems and Applications of the IEEE Circuits and Systems Society. He recently served as Conference Chair for ICCE 2006 and Tutorials Chair for ICME 2006 and has been a member of the Publications Committee of the IEEE Transactions on Consumer Electronics since 2002. He was also a Guest Editor for the special section on MPEG-21 for IEEE Transactions on Multimedia, June 2005, and for the special issue on Universal Multimedia Access for *IEEE Signal Processing Magazine*, March 2003. He has received several awards for his work on transcoding, including the 2003 IEEE Circuits and Systems CSVT Transactions Best Paper Award.

**Thomas Sikora** (Senior Member, IEEE) received the Dipl.-Ing. degree and Dr.-Ing. degree in electrical engineering from Bremen University, Bremen, Germany, in 1985 and 1989, respectively.

Currently, he is a Professor and Director of the Communication Systems Department, Technical University of Berlin, Berlin, Germany. In 1990, he joined Siemens, Ltd., and Monash University, Melbourne, Australia, as a Project Leader responsible for video compression research activities in the Australian Universal Broadband Video Codec consortium. He became a Member of the Research Staff of the Heinrich-Hertz-Institute (HHI), Berlin, in 1994, and directed the Interactive Media Department at HHI between 1997 and 2001. He has been involved in international ITU and ISO standardization activities as well as in several European research activities for a number of years. He acted as the Chairman of the ISO-MPEG Video Group between 1995 and 2001, responsible for the development and standardization of the MPEG-4 and MPEG-7 video coding algorithms. He also served as the Chairman of the European COST 211ter video compression research group. He frequently works as an industry Consultant on issues related to interactive digital audio and video. He is an appointed member of the advisory and supervisory board of a number of German companies and international research organizations. He has published two books on MPEG-7 and more than 200 refereed journal and conference papers in the field of image, video, and audio processing, and he has been an invited plenary speaker at a number of international conferences.

Dr. Sikora is a recipient of the 1996 German ITG award (German Society for Information Technology). He was the Editor-in-Chief of the IEEE Transactions on Circuits and Systems for Video Technology until 2006. He is an Associate Editor of the *EURASIP Signal Processing Journal* and an Advisory Editor for the *EURASIP Signal Processing: Image Communication Journal*. From 1996 to 2000, he was on the Editorial Board the *IEEE Signal Processing Magazine*. He is a member of ITG.