MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Digital Item Adaptation: Overview of Standardization and Research Activities

Anthony Vetro and Christian Timmerer

TR2005-056   December 2005

Abstract

MPEG-21 Digital Item Adaptation (DIA) has recently been finalized as part of the MPEG-21 Multimedia Framework. DIA specifies metadata for assisting the adaptation of Digital Items according to contraints on the storage, transmission and consumption, thereby enabling various types of quality of service management. This paper provides an overview of DIA, describes its multimedia applications, and reports on some of the ongoing activities in MPEG on extending DIA for use in rights governed environments.

IEEE Transactions on Multimedia, June 2005

Digital Item Adaptation: Overview of Standardization and Research Activities

Anthony Vetro, *Senior Member, IEEE,* and Christian Timmerer

***Abstract*—MPEG-21 Digital Item Adaptation (DIA) has recently been finalized as part of the MPEG-21 Multimedia Framework. DIA specifies metadata for assisting the adaptation of Digital Items according to constraints on the storage, transmission and consumption, thereby enabling various types of quality of service management. This paper provides an overview of DIA, describes its use in multimedia applications, and reports on some of the ongoing activities in MPEG on extending DIA for use in rights governed environments.**

***Index Terms*—Adaptation, Digital Item, MPEG, multimedia, quality of service, universal multimedia access.**

I. Introduction

UNIVERSAL multimedia access (UMA) has become the driving concept behind a significant amount of research and standardization activity [1]. It essentially refers to the ability for any type of terminals to access and consume a rich set of multimedia content. Ideally, this is achieved seamlessly over dynamic and heterogeneous networks and devices, independent of location or time, and taking into account a wide variety of possible user preferences.

Toward this goal of universal accessibility, techniques for scalable coding and transcoding were developed. An early application of such techniques was in the distribution of television programming from the studio over bandwidth-limited broadcast networks. Shortly after, the challenges in transmitting video over wireless networks and the Internet emerged. While scalable coding and transcoding are quite different approaches, they both offer a means to adapt various coding parameters of the media, such as the bit-rate and spatial/temporal resolution.

An alternative to real-time transcoding or scalable coding is creating content in multiple formats and at multiple bit-rates and making an appropriate selection at delivery time [2]. This solution is not compute-intensive, but highly storage-intensive. This approach could be sufficient for cases in which the types of receiving terminals are limited, but for the general case, where no limitations exist on the terminal or access network, real-time adaptation is necessary.

Technology that provides resources satisfying the usage environment is indeed a fundamental component of a complete delivery system, but still more is needed to achieve the stated goal. With the dawn of multimedia content description, namely MPEG-7 [3], several tools have been developed to further support universal accessibility. It should be noted that in the context of this paper, tools mainly refer to description tools, which are synonymous with metadata. In MPEG-7, there exist tools for describing the summarization of media resources, tools that provide transcoding hints about the media resources, and tools that indicate the available variations of a given media resource; all of which have been standardized in support of UMA [4]. Needless to say, it is also critical to know the media's coding format of the source content, which is also specified by MPEG-7. A good review of how these tools are applied toward the UMA problem is given in [5].

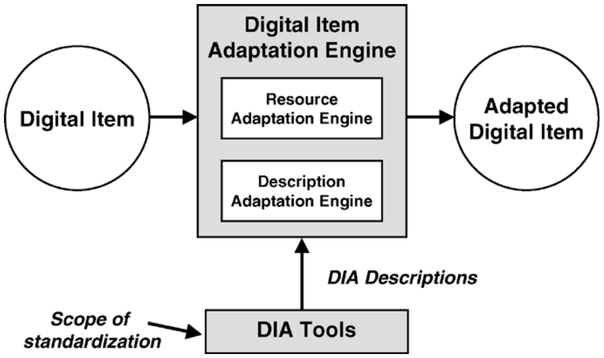
Digital Item
Digital Item Adaptation Engine
Resource Adaptation Engine
Description Adaptation Engine
Adapted Digital Item
DIA Descriptions
Scope of standardization
DIA Tools

Fig. 1. Concept of MPEG-21 DIA.

Recognizing that there are still some missing elements to this picture, MPEG-21 has gone forward to standardize tools that attempt to fill those gaps. In particular, Part 7 of the standard [6], referred to as Digital Item Adaptation (DIA), specifies tools to assist with the adaptation of Digital Items. Digital Items are defined as structured digital objects, including standard representation, identification and metadata, and are the fundamental unit of distribution and transaction within the MPEG-21 framework. A high-level illustration of the DIA is given in Fig. 1. As shown in this figure, only tools used to guide the adaptation engine are specified by the standard. The adaptation engines themselves are left open to various implementations.

The purpose of this paper is not to provide an in-depth coverage of all the various tools and their specification. Our emphasis will be on reviewing key aspects of each tool and providing some insights regarding their application, use and relation to one another. This paper will also discuss ongoing standardization activities and open issues, as well as avenues for further research. We refer readers interested in further details of

Manuscript received April 29, 2004; revised September 28, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. J. R. Smith.

A. Vetro is with Mitsubishi Electric Research Labs, Cambridge MA 02139 USA (e-mail: avetro@merl.com).

C. Timmerer is with the Department of Information Technology (ITEC), Klagenfurt University, 9020 Klagenfurt, Austria (e-mail: christian.timmerer@itec.uni-klu.ac.at).

Digital Object Identifier 10.1109/TMM.2005.846795

the standard to [6] and [7], as well as the other DIA-specific papers appearing in this special issue [9], [12], [14], [17].

It should be noted that the W3C Consortium is also engaged in efforts to bridge this gap between multimedia content and devices. In particular, the Device Independence Working Group has recently completed a first version of the Composite Capability/Preference Profiles (CC/PP), which specifies a structure and vocabularies for device capabilities and user preferences [8].

The remainder of this paper is organized as follows. In the next sections, brief overviews of a number of tools are given. Section II provides an overview of usage environment description tools. Section III introduces a new language for high-level descriptions of the bitstream syntax. Section IV presents tools that describe the relationship between quality of service (QoS) constraints, feasible adaptation operations satisfying these constraints, and associated media resource qualities that result from the adaptation. Section V covers tools that enable low-complexity adaptation of metadata. Section VI describes tools for session mobility and Section VII describes information that could be used for the configuration of an adaptation engine. Ongoing standardization efforts, future research topics and open issues, are presented in Section VIII.

II. Usage Environment Descriptions

The usage environment description includes the description of terminal capabilities and network characteristics, as well as User[1] characteristics and characteristics of the natural environment. Such descriptions provide a fundamental input to any adaptation engine. The parameters that have been specified by the standard are reviewed below. Some example applications are also discussed.

A. Terminal Capabilities

The notion of a terminal in DIA is rather generic and represents all devices regardless of their location in the delivery chain. So, in addition to the typical consumer electronic devices such as mobile phones, televisions, audio players and computers, a terminal could also be a server, proxy or any intermediate network node. With this definition, terminal capabilities must be described in terms of both receiving and transmitting capabilities. Such a description is used to satisfy consumption and processing constraints of a particular terminal. The major categories are outlined below.

1) Codec Capabilities: Encoding and decoding capabilities specify the format a particular terminal is capable of encoding or decoding, e.g., an MPEG profile/level. Given the variety of different 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 DIA would provide a means to describe such limits, e.g., the maximum bit-rate that a decoder could handle.

2) Input-Output Characteristics: Display capabilities, audio output capabilities and User interaction inputs are the key items considered under this category. Describing the capabilities of a display is obviously very important as certain limitation 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: There are a variety of properties specified under this umbrella, but due to space limitations, we only cover power and storage characteristics, and CPU benchmark measures. The power characteristics tool is intended to provide 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 a Digital Item is consumed, e.g., whether it needs to be streamed or could be stored locally. To gauge computational performance, DIA has adopted a benchmark-based description, 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.

B. Network Characteristics

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 [9].

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.

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

[1] In the MPEG-21 framework, a User (capitalized) is any entity that interacts in the MPEG-21 environment or makes use of a Digital Item. Such Users include individuals, consumers, communities, organizations, software agents, etc.

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.

C. User Characteristics

The descriptions concerning the characteristics of a User as specified by DIA may be used for a number of purposes, including the adaptive selection or personalization of content.

1) User Info: General information about a User is specified in DIA by importing the Agent DS specified by MPEG-7 [3]. The Agent DS describes general characteristics of a User such as name and contact information, where a User can be a person, a group of persons, or an organization.

2) Usage Preference and History: As with User Info, corresponding tools specified by MPEG-7 define the usage preference and usage history tools in DIA. The usage preference tool is a container of various descriptions that directly describe the preferences of a User. The usage history tool describes the history of actions on Digital Items by a User, e.g., recording a video program, playing back a piece of music; as such, it describes the preferences of a User indirectly.

3) Presentation Preferences: This class of tools defines a set of preferences related to the means by which Digital Items and their associated resources are presented or rendered for the User. Within this category, DIA has specified a rather rich set of tools that include preferences related to audio-visual rendering, to the format or modality a User prefers to receive, to the priority of the presentation, as well as the preferences that direct the focus of a Users attention with respect to audio-visual and textual media.

4) Accessibility Characteristics: These tools provide descriptions that would enable one to adapt content according to certain auditory or visual impairments of the User. For audio, an audiogram is specified for the left and right ear, which specifies the hearing thresholds for a person at various frequencies in the respective ears. For visual related impairments color vision deficiencies are specified, i.e., the type and degree of the deficiency. For example, given that a User has a severe green-deficiency, an image or chart containing green colors or shades may be adapted accordingly so that the User can distinguish certain markings. Such descriptions would also be very useful to simply determine the modality of a media resource to be consumed, but may also be used for more sophisticated adaptations, such as the adaptation of color in an image [10].

5) Location Characteristics: There are two tools standardized by DIA that target location-based characteristics of a User: mobility characteristics and destination. The first of these tools aims to provide a concise description of the movement of a User over time. In particular, directivity, location update intervals and erraticity are specified. Directivity is defined to be the amount of angular change in the direction of the movement of a User compared to the previous measurement. The location update interval defines the time interval between two consecutive location updates of a particular User. Erraticity defines the degree of randomness in a User's movement. Together, these descriptions can be used to classify Users, e.g., as pedestrians, highway vehicles, etc. Destination is a tool to indicate, as the name implies, the destination of a User, i.e., an intended location at a future time. The destination itself may be specified very precisely, e.g., with geographic coordinates, or more conceptually using a specified list of terms. In conjunction with the mobility characteristics, this tool could also be used for adaptive location-aware services.

D. Natural Environment Characteristics

The natural environment pertains to the physical environmental conditions around a User such as lighting condition or noise level, or a circumstance such as the time and location that Digital Items are consumed or processed.

In DIA, the *Location* description tool refers to the location of usage, while *Time* refers to the time of usage. Both are specified by MPEG-7 description tools, namely the Place DS for Location and the Time DS for Time [3]. Tools that describe location and time are referenced by both the mobility characteristics and destination tools and may also be used to support adaptive location-based services.

With respect to the visual environment, illumination characteristics that may affect the perceived display of visual information are specified. It has been observed that the overall illumination around a display device affects the perceived color of images on the display device and contributes to the distortion or variation of perceived color [11]. By compensating the estimated distortion, actual distortion caused by the overall illumination can be decreased or removed.

For audio, the description of the noise levels and a noise frequency spectrum are specified. The noise level is represented as a sound pressure level in decibels, while the noise spectrum are the noise levels for 33 frequency bands of 1/3 octave covering the range of human audible bandwidth. An adaptation engine would enhance the perceived quality of the adapted audio signal by modifying the frequency attenuation of the original audio signal according to the noise characteristics. Interested readers may refer to [12] for further details.

III. Bitstream Syntax Descriptions

A bitstream is defined as a structured sequence of binary symbols. DIA uses XML to describe the high-level structure of a bitstream, i.e., how it is organized in packets or layers of data. The resulting XML document is called bitstream syntax description (BSD). In most cases, the BSD does not describe the bitstream on a bit-per-bit basis and may describe the bitstream at different syntactic layers, e.g., finer or coarser levels of detail such as frames or scenes for a video resource, depending on the application.

Due to the large diversity of competing or complementary scalable coding formats, a device needs to maintain the corresponding number of software/hardware modules in order to facilitate the manipulation of bitstreams based on all these formats. Thus, to solve this limitation and leverage the use of scalable multimedia formats, DIA defines a generic framework based on XML for bitstream adaptation [13], which can be also applied in constrained and streaming environments [14].

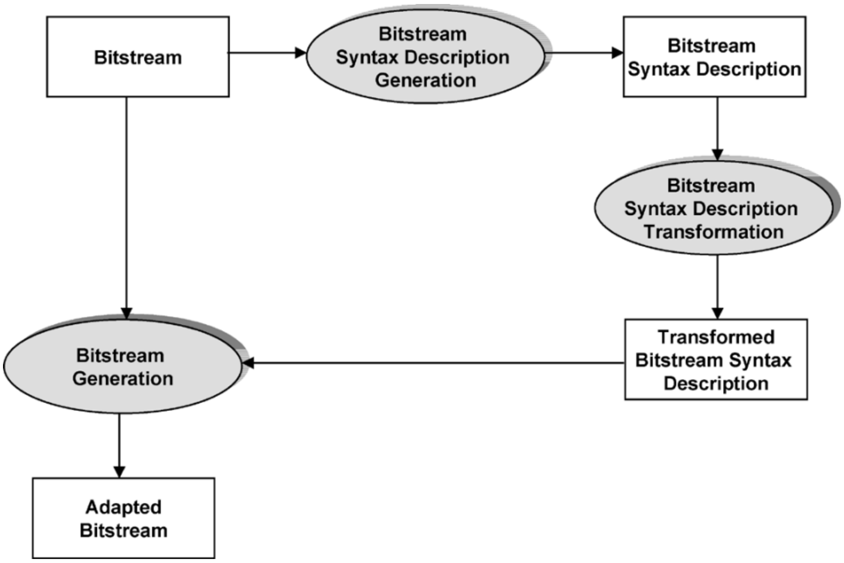
Bitstream
Bitstream Syntax Description Generation
Bitstream Syntax Description
Bitstream Syntax Description Transformation
Transformed Bitstream Syntax Description
Bitstream Generation
Adapted Bitstream

Fig. 2. BSD adaptation architecture.

The actual bitstream adaptation is performed by transforming the BSD corresponding to an input bitstream, and then generating the adapted version of the bitstream from the transformed BSD. This process is illustrated in Fig. 2.

The behavior of both the BSD generation and the bitstream generation processes are normatively specified by the standard. However, the BSD transformation process is not normative and can be performed, for instance, by using standard XML transformation languages such as the Extensible Stylesheet Language Transformations (XSLT) [15] or the Streaming Transformations for XML (STX) [16].

This transformation mainly comprises the removal of elements as well as minor editing operations, such as modifying the value of an element. To achieve this, the BSD may include pointers to specific segments of the bitstream, as well as the actual values of certain syntax elements of the bitstream. During the BSD transformation, certain segments may be removed and certain values may be modified. Note that this transformation could also include the addition of new elements to a BSD ensuring compliance with the coding format or increasing the quality of the bitstream.

For BSD and bitstream generation, a new language named *Bitstream Syntax Description Language (BSDL)* has been specified in DIA. BSDL is built on top of XML Schema and defines further restrictions and extensions for the multimedia domain which can be used to write *Bitstream Syntax Schemas (BS Schemas)*. These schemas convey the information required by a generic processor to generate a bitstream from its description and vice-versa for a particular coding format. Note that DIA does not standardize any BS Schemas, even not for well-known coding formats such as JPEG2000 or MPEG-4, since they may be different for various application requirements.

In addition to BSDL, the standard also specifies a *generic Bitstream Syntax Schema (gBS Schema)* written in BSDL, which facilitates generic and coding-format independent BSD's. Consequently, these BSD's are referred to as *generic Bitstream Syntax Descriptions (gBSDs)* and are applicable to any coding format. In addition to its generality, a gBSD provides semantically meaningful marking of syntactical elements described by use of a "marker" handle. This allows one to include application- or domain-specific information in the gBSD, e.g., marking violent scenes within a gBSD describing an action movie. Furthermore, the description of a bitstream can be constructed in a hierarchical fashion that allows grouping of bitstream elements for efficient, hierarchical adaptations. The flexible addressing scheme supports various application requirements and random access into the bitstream. Finally, it enables distributed adaptations in terms of multi-step adaptations where several adaptation steps are concatenated along the delivery path from the provider to the consumer.

To ease the referencing of the information assets required for a BSD-based adaptation, i.e., the bitstream, its (g)BSD, and corresponding BSD transformation style sheets with appropriate parameters, the *BSDLink* tool is specified within DIA. Additionally, it provides means for steering this kind of bitstream adaptation by various MPEG description tools. One example of a steering description is the AdaptationQoS description as described in the next section.

IV. Terminal and Network QoS

The goal of terminal and network QoS is to select optimal parameter settings for media resource adaptation engines that satisfy constraints imposed by terminals and/or networks while maximizing QoS. To facilitate this, the *AdaptationQoS* tool specifies the relationship between constraints, feasible adaptation operations satisfying these constraints, and associated utilities (qualities). In the following, the main features of this tool are highlighted; further details are given in a companion paper in this special issue [17].

Three types of modules are specified by the standard to define this relationship, which allows an optimal choice for representing data according to the actual application requirements, e.g., feasible set of adaptation operators, usage environment constraints, or parameters for adaptation engines. The representation associated with each of these modules is

TABLE I
AVAILABLE REPRESENTATIONS TO EXPRESS RELATIONS IN ADAPTATIONQOS

Module	Representation	Comments
Look-Up Table	MxN matrix, in which elements of matrix may be defined as string, integer, floating-point values, or NA value	Suitable for dense and discrete data; generally applicable for most adaptation requirements, but may incur additional overhead when data is sparse
Utility Function	N-dimensional vector, where elements of vector may be defined as string, integer, floating-point values, or NA value	Suitable for sparse and discrete data; more efficient than look-up tables when many matrix coefficients are assigned an NA value
Stack Function	Expressed using the Reverse Polish Notation (RPN), where an expression is written as a sequence or stack of operators and arguments [18]	Suitable for functional and continuous data, i.e., when a continuous approximation of the actual mapping by an equation is desirable

summarized in Table I. Although each of the modules operates differently, a generic interface to the modules is provided through Input/Output Pins (IOPins). Each IOPin is a uniquely identifiable variable that is globally declared and referenced from within a module. An IOPin can be interpreted as input, output, or both, which also allows the interconnection of different modules. The values of the IOPins can be either directly declared as continuous or discrete within the IOPin or as an external parameter specified by its semantics. Furthermore, IOPin values can be further constrained by using the *Universal Constraints Description (UCD)* tool, also specified within DIA.

The UCD allows further constraining the *usage* and *usage environment* of a Digital Item by means of limitation and optimization constraints. Both types of constraints are formulated using the stack function syntax as mentioned above. For instance, the usage of a Digital Item containing an image resource can be constrained in such a way that it should not be adapted to smaller than 20% of the receiving terminal's display resolution – this may be a restriction set by the content provider. However, the dimension of the image should be maximized if adaptation is required. On the other hand, the usage environment may define a similar constraint where the image should be at least smaller than 75% of the rendering terminal's display resolution due to specific application requirements, e.g., the rendering application should not run in full-screen mode to conserve resources.

V. METADATA ADAPTABILITY

Metadata is widely used to associate additional information to various types and collections of multimedia content. This allows for search and retrieval, as well as content navigation. We note that the metadata itself may be viewed as a primary resource to be visually rendered and browsed, e.g., as a menu.

With regards to metadata adaptation, there are several important concerns. First, as the content is adapted, the associated metadata must also change accordingly. Second, if the metadata is transmitted and consumed, it may need to be scaled in order to meet terminal and network constraints. Third, given a very rich and detailed description of a piece of content, filtering is often necessary to obtain only the necessary or interesting parts of the metadata for a User. Finally, for the case when there exist multiple sources of metadata for the same resource, an efficient means of integrating these different sources of metadata into a single description could be very useful. DIA specifies tools to assist with all of the above issues. In the following, we cover the tools that assist with the scaling, filtering and integration of metadata.

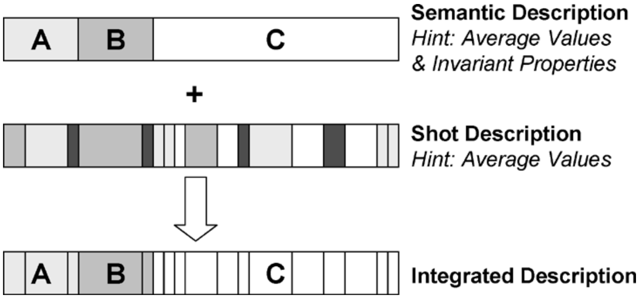

Fig. 3. Illustration of metadata integration in which semantic and syntactic/shot descriptions are efficiently combined using metadata adaptation hints.

A. Filtering and Scaling

Given a large description instance that should be sent and processed on a mobile device, there may be several difficulties that filtering and/or scaling the metadata would help to overcome [19]. For one, the processing power and memory on a mobile device is likely to be limited, so sending unnecessary or irrelevant metadata is wasteful, not only to the processing and memory resources, but the bandwidth that is needed to send the metadata as well. Also, if the metadata is to be displayed in some form, it is better to reduce the depth of the description instance beforehand and filter out elements that are not needed. To support this application, such attributes as the size of the metadata, total number of elements, the depth and number of occurrences of certain elements, are specified. This information could be used to reduce the parsing complexity of the actual metadata instance.

B. Integration

In contrast to filtering and scaling, the integration of metadata is considered when one resource is described by multiple descriptions [20]. In this case, redundant information should be discarded and complementary information should be integrated.

Consider the case in which two descriptions of the same movie have been produced as illustrated in Fig. 3. One description includes low-level features of each shot that have been automatically generated, while the other description contains semantic descriptions for specific segments of the program that have been manually generated. Hint information is provided with each description; the particulars of this hint information and its use for metadata integration are discussed further below.

As a first step toward integrating descriptions, it would be helpful to know whether or not the descriptions share a similar spatio-temporal structure or decomposition. This may be ascertained through hint information that gives the *average values* for segment durations (in time) within each description. In the example of Fig. 3, we assume that the temporal boundaries of the two descriptions perfectly align with each another; if they did not, some additional processing to align the boundaries would be needed. The next step is to know whether certain description values can be propagated to sub-segments. This functionality is achieved through the *invariant properties* hint information. In our example, we assume that the semantic descriptions are invariant and can be propagated to subsegments according to the shot-level decomposition. As a result, the integrated description as shown in the figure can be created, which includes both shot-level and semantic-level descriptions.

VI. Session Mobility

In a world where peer-to-peer networking and redistribution of content among Users is becoming commonplace and secure, the means by which Digital Items are transferred from one device to another device is an important consideration. This section discusses some key principles in preserving a User's current state of interaction with a Digital Item.

The Digital Item Declaration (DID) [21] specifies a uniform and flexible abstraction and interoperable schema for declaring the composition and structure of Digital Items. As a basic functionality, we can declare a Digital Item by specifying its resources, metadata, and their interrelationships. Further, the DID can be configured by *Choices* and *Selections* that are part of the declaration. We refer to this instantiation of Choices/ Selections as the *configuration-state* of a Digital Item.

In DIA, session mobility refers to the transfer of configuration-state information that pertains to the consumption of a Digital Item on one device to a second device. This enables the Digital Item to be consumed on the second device in an adapted way. During this transfer, *application-state* information, which pertains to information specific to the application currently rendering the Digital Item, may also be transferred.

To make the session mobility concepts more concrete, consider the following example of an electronic music album and the illustration in Fig. 4. The different songs of this album are stored at a content server and expressed as individual Selections within a particular Choice of the DID. There may also be other Choices within the DID that configure the platform, acceptable media formats, and so on, for a given User. In this example, assume that a first User is listening to the second song on device A, and this User would like to share this particular song with a second User on device B. Assuming that they have the right to do this, the configuration-state of the DID on the first device, i.e., the Selection that pertains to the second song, would be captured and transferred to the second device. Since the platform and acceptable media formats may be different on the second device, potential Choices in the original DID concerning those Selections would still need to be made there. Supplemental application-state information may also be transferred such as timing information related to songs or layout information if images and video are involved.

As we see from the above example, the User's current state of interaction with a Digital Item is completely described by both the configuration-state and application-state information.

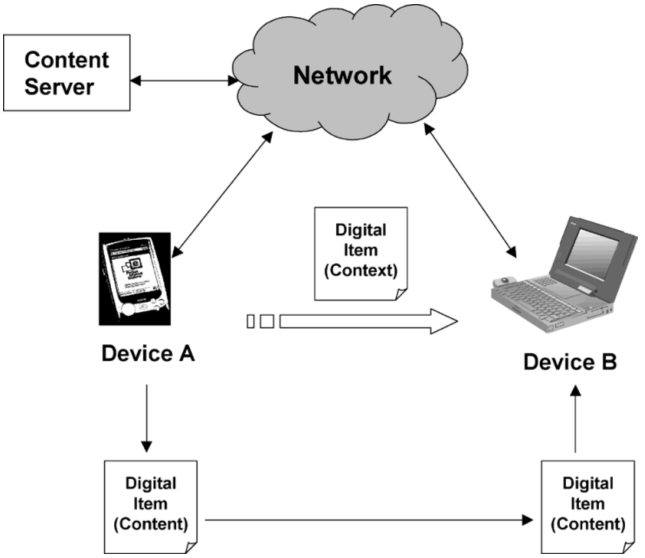

Fig. 4. Illustration of session mobility concept in which session of Device A is transferred with appropriate contextual information.

VII. DIA Configuration

The DIA Configuration tools are used to help guide the adaptation process considering the intentions of a DID author. There are essentially two types of tools that have been specified. The first is used to suggest the means by which Choice/Selections in a DID should be processed, e.g., displayed to Users or configured automatically according to DIA descriptions, while the second is used to suggest particular DIA descriptions that should be used for resource adaptation.

As an example, consider a DID with two Choices: one pertaining to the video quality and another pertaining to the media format. The Descriptor of each Choice in the DID may include a DIA tag that indicates whether the Choice should be manually selected by the User or automatically configured according to relevant metadata. Assume that the user selects the video quality and the media format is configured automatically. Given that the media format should be configured automatically, the DIA Configuration tool may further describe the particular DIA description that would be required to perform the configuration; in this case, the supported media formats on the terminal.

VIII. Additional Activities, Research Topics and Open Issues

This section describes ongoing standardization activity related to DIA, as well as several emerging research topics and open issues.

A. Permissible and Secure Adaptation

Two important aspects of enabling permissible and secure adaptation are discussed here. The first is related to the description of adaptation methods and parameters, which could be used to define rights expressions to govern adaptations in an interoperable way. The second aspect assumes that permission to perform a particular adaptation has been granted, but a secure adaptation of possibly encrypted content still needs to be performed.

Conversions descriptions can be used to identify suggested or permitted conversions for particular resources, or to describe terminal capabilities in terms of its supported conversions. In an amendment of DIA, a framework has been specified that facilitates the description of conversion related information, including the adaptation operation, parameters of the adaptation, and change conditions [22]. As an example, consider the cropping of an image as the adaptation operation. In this case, the x-y offset, and the width and height of the cropped region completely describe the operation. A constraint on the parameters of the operation may also be imposed, e.g., both width and height must be less than or equal to 256. The UCD tool introduced in Section IV and described further in [17] would be used to specify such a constraint.

In order to express the rights associated with an adaptation, a license conformant to the Rights Expression Language (REL) is needed [23]. This license would indicate permissible changes as detailed by the conversion descriptions. To govern these rights in an interoperable manner, the existing terms of the Rights Data Dictionary (RDD) must be specialized . This means that existing terms that are related to the specific adaptation must be identified, and then a new term is created as a specialization of those existing terms. The specification provides examples on how this is done, but it does not standardize any particular adaptation operations or parameters. In this way, only the format and the semantics of the conversions and its parameters are defined, such that the various industries that will employ this standard could create new types of conversions and parameters. To achieve interoperability these newly created conversions and parameters need to be made accessible, e.g., by means of a registration authority.

It should be noted that the REL and RDD already provide tools to permit playing, modifying, and adapting; however, only with coarse control. The above mentioned amendment of DIA essentially enables finer-grained control over the changes that can occur when playing, modifying, or adapting Digital Items and their component resources. In this way, a flexible copy protection system that allows only certain changes to the content could be designed. It should be emphasized that although changes to the content are governed, the standard does not specify the actual implementation of the adaptation engine.

In the above, a means to grant permission to perform a particular adaptation is discussed. However, performing the actual adaptation of an encrypted bitstream in a secure manner is another issue. It is assumed that breaches in security by decrypting and re-encrypting within the network should be avoided. This problem has been addressed in [24] with a secure scalable streaming format that combines scalable coding techniques with a progressive encryption technique. However, handling this for nonscalable video and streams encrypted with traditional encryption techniques is still an open issue.

B. Adaptation in Constrained and Streaming Environments

In Section III we described a generic framework for multimedia content adaptation by transforming its BSD and using a generic processor to retrieve the adapted version of the bitstream. DIA does not standardize the transformation process of the BSD or which transformation language should be used. One recommended solution is the widely used XSLT. In practice, however, one limitation of XSLT is that it requires the full input documents in memory in order to transform the BSD. This is a burden for constrained devices with memory restrictions and in streaming environments when the complete BSD cannot be available at once. Two approaches are discussed in [14] to overcome these shortcomings. The first approach is based on STX and the filtering of SAX events. The second re-uses the fragmenting concepts introduced in MPEG-7 Systems [25], where one large BSD is divided into smaller processing units and treated like a complete BSD enabling a legacy XML transformation processor to manipulate them.

C. Transport, Negotiation and Exchange of Descriptions

An important practical issue for DIA is how the transport, negotiation and exchange of DIA descriptions will be handled. The use case scenarios vary greatly from a simple pull scenario in which a mobile device requests content from a remote server, to more complex and hybrid scenarios in which certain DIA descriptions are transported with the content, but then additional DIA descriptions are required at a later stage to perform the adaptation.

It is noted that an amendment to the MPEG-2 Systems specification that defines a means to transport metadata, e.g., MPEG-7 descriptions, may be used for the transport of DIA descriptions [26]. We also note that a binary format of the DID [21], which may bundle DIA descriptions, is being specified in Part 16 of the MPEG-21 standard, which is largely based on the binary format of MPEG-7 [25]. Such a format could be used to overcome limitations in network bandwidth in certain application spaces and also provide a common exchange format. Of course, other standardization bodies or forums interested in using DIA descriptions may also define specific mechanisms for the intended purpose and application space. We imagine that DIA descriptions, UED in particular, may also be carried as part of a request for the resource, e.g., through HTTP or RTSP.

D. Semantic Clues for Adaptation

Adding semantic meaningful information to multimedia content has been an important issue for several years. A means to associate such information with multimedia content in an interoperable way could be achieved with the Semantic DS in MPEG-7 [3]. However, only the description format is specified, while the means to extract and consume such information is left open.

There are several groups working on automatic extraction of semantic information from multimedia content. In [27], a framework based on statistical machine learning has been proposed to extract semantic concepts from low level features. Therefore, semantic concepts, e.g., "outdoors" or "face", are modeled using support vector machine (SVM) classification, which are used for the detection process of the unobserved content, i.e., content that is not part of the training set.

Once extracted, these semantic concepts, i.e., the labels, could be included into the gBSD's *marker* attribute as introduced in Section III, which enables semantically meaningful adaptation at the bitstream level. Both the extraction of semantic information and the use of such information for adaptation are open research topics.

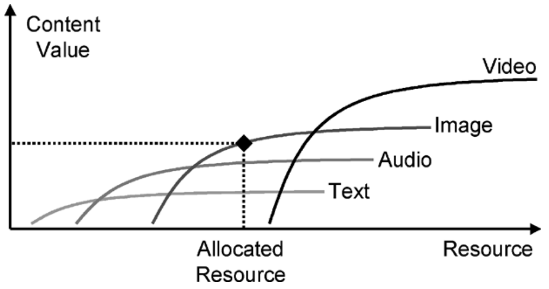

Fig. 5. Modality curves plotted in resource-value space.

E. Modality Conversion

In cases where the scaling of content is not sufficient to meet terminal or network constraints, the conversion from one modality to another could provide a better alternative. For instance, rather than sending a low-quality video sequence over a narrow-band connection, it may be better to send selected key frames. This is an example of a simple video-to-image conversion. Of course, many more complex scenarios and conversions across modalities could be imagined.

Tools that support modality conversion in DIA are rather limited at the moment. Usage environment description tools could be used to signal that a particular terminal is not capable of decoding a particular format or rendering a specific modality. Also, User characteristics that indicate the preference toward certain modalities or formats have been specified. However, the use of modality conversion to satisfy a richer set of terminal and network constraints within the AdaptationQoS framework described in Section IV has not yet been fully explored.

In [28], a resource-value framework for converting or selecting different content formats was presented as a generalization of the classic rate-distortion theory. In this context, the notion of resource is analogous to rate, but includes a broader set of terminal and network constraints, while content value is analogous to distortion, but applies across different content scales and modalities. Given a resource constraint, the version of the content providing the maximum content value would be selected. This framework was further explored in [29] in which the points of intersection between the different modality curves were studied. Fig. 5 illustrates an example of resource-value curves for different modalities. Given these curves, the optimal modality for an allocated resource could be determined.

Generally speaking, one of the main difficulties with modality conversion is estimating and comparing the content values across different modalities. While various methods of estimating content value have been explored in the above works that consider general signal-level trends in the adapted content, a complete understanding of all the factors that contribute to the content value, including subjective preferences, remains as an open research problem [30].

F. Maximum User Experience

The previous subsection mainly addressed the difficulty in determining content values for different scales and modalities to satisfy terminal and network constraints, where the objective is still very much aimed at enabling multimedia access. However, as observed in [31], the end point of multimedia consumption is the user, not the terminal. Hence, the ultimate aim is to maximize the user's experience. It is noted that human factors, such as mood, play a significant role in this and are not easily quantified or predicted.

Personalization of content to a particular user is certainly one way to increase the user's experience [32]. As mentioned in Section VIII.D, techniques for annotating and modeling semantics using statistical methods have shown promise for automatic labeling of video content [27]. These labels could then be used for semantic-level transcoding or personalization. For instance, a summary of a news program could be generated according to a user's interest, or an educational video could be customized according to a user's learning progress. Additionally, in [33], a visual attention model is proposed for the adaptation of images that considers not only attributes that attract a users perception of objects in the image, but also factors such as the minimum perceptible size. Based on this model, a better scaling and/or cropping of the image for small display devices could be achieved.

Indeed, there are many factors to be accounted for in an adaptation framework that aims to maximize user experience. User preferences are certainly a dominant factor, but are dynamic in that they are very likely to vary with the content, task, and usage environment [30]. While there are still many open research issues that need further study, it is clear that adaptive delivery to users can indeed be achieved to some extent with tools that are available today, and that this will certainly be improved upon with a better understanding of human factors that pertain to the consumption of multimedia.

Acknowledgment

The authors thank all of those who have been involved in bringing this part of the MPEG-21 standard to completion, in particular S. Devillers, who served as a co-editor.

References

[1] A. Vetro, C. Christopoulos, and T. Ebrahami, Eds., *IEEE Signal Process. Mag.—Special Issue on Universal Multimedia Access*, vol. 20, no. 2, Mar. 2003.

[2] J. R. Smith, R. Mohan, and C.-S. Li, "Scalable multimedia delivery for pervasive computing," *Proc. ACM Multimedia*, pp. 131–140, 1999.

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

[4] *ISO/IEC 15938-5:2003, Information Technology - Multimedia Content Description Interface - Part 5: Multimedia Description Schemes*.

[5] P. van Beek, J. R. Smith, T. Ebrahimi, T. Suzuki, and J. Askelof, "Metadata-driven multimedia access," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 40–52, Mar. 2003.

[6] *ISO/IEC 21000-7:2004, Information Technology - Multimedia Framework - Part 7: Digital Item Adaptation.*.

[7] A. Vetro, C. Timmerer, and S. Devillers, "Digital Item Adaptation," in *The MPEG-21 Book*. Hoboken, NJ: Wiley, to be published.

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

[9] M. van der Schaar and Y. Andreopoulos, "Rate-distortion-complexity modeling for network and receiver aware adaptation," *IEEE Trans Multimedia*, vol. 7, no. 3, pp. 471–480, Jun. 2005.

[10] J. Song, S. Yang, C. Kim, J. Nam, J. W. Hong, and Y. M. Ro, "Digital item adaptation for color vision variations," in *Proc. SPIE Conf. Human Vision Electronic Imaging VIII*, Santa Clara, CA, Jan. 2003.

[11] Y. Huh and D. S. Park, "Illumination Environment Description for Adaptation of Visual Contents,", Fairfax, VA, ISO/IEC JTC1/SC29/WG11 m8341, May 2001.

[12] B. Feiten, I. Wolf, E. Oh, J. Seo, and H.-K. Kim, "Audio adaptation according to usage environment and perceptual quality metrics," *IEEE Trans Multimedia*, vol. 7, no. 3, pp. 446–453, Jun. 2005.

[13] G. Panis *et al.*, "Bitstream syntax description: A tool for multimedia resource adaptation within MPEG-21," *EURASIP Signal Processing: Image Commun. J.*, vol. 18, no. 8, pp. 721–747, Sep. 2003.

[14] S. Devillers, C. Timmerer, J. Heuer, and H. Hellwagner, "Bitstream syntax description-based adaptation in streaming and constrained environments," *IEEE Trans Multimedia*, vol. 7, no. 3, pp. 463–470, Jun. 2005.

[15] W3C Consortium. (1999) "XSL Transformations (XSLT) Version 1.0", W3C Recommendation. [Online]. Available: http://www.w3.org/TR/xslt

[16] P. Cimprich. (2003) Streaming Transformations for XML (STX) Version 1.0. Working Draft. [Online]. Available: http://stx.sourceforge.net/documents/

[17] 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.

[18] C. L. Hamblin, "Computer languages," *Australian J. Sci.*, vol. 20, pp. 135–139, 1957.

[19] H. Nishikawa *et al.*, "Metadata centric content distribution based on MPEG-21 digital item adaptation," in *Proc. 5th Asia-Pacific Symp. Information Telecom. Technol.*, Noumea, New Caledonia, Nov. 2003.

[20] N. Adami, M. Corvaglia, and R. Leonardi, "Comparing the quality of multiple descriptions of multimedia documents," in *Proc. Multimedia Signal Processing Workshop*, Virgin Islands, Dec. 2002.

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

[22] *ISO/IEC 21000-7:2004/FPDAM.1, Information Technology - Multimedia Framework - Digital Item Adaptation: Conversions and Permissions*, Jan. 2005.

[23] 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.

[24] S. Wee and J. Apostolopoulos, "Secure scalable streaming and secure transcoding with JPEG 2000," in *Proc. Int. Conf. Image Processing*, Barcelona, Spain, Sep. 2003.

[25] *ISO/IEC 15938-1:2003, Information Technology - Multimedia Content Description Interface - Part 1: Systems.* .

[26] *ISO/IEC 13818-1:2000/Amd.1:2003, Carriage of metadata over ITU-T Rec. H.222 | ISO/IEC 13818-1 streams.*.

[27] C.-Y. Lin, B. L. Tseng, M. Naphade, A. Natsev, and J. R. Smith, "VideoAL: A novel end-to-end MPEG-7 semantic video automatic labeling system," in *Proc. IEEE Int. Conf Image Processing*, Barcelona, Spain, Sep. 2003.

[28] R. Mohan, J. R. Smith, and C.-S. Li, "Adapting multimedia internet content for universal access," *IEEE Trans. Multimedia*, vol. 1, pp. 104–114, Mar. 1999.

[29] T. C. Thang, Y. J. Jung, J. W. Lee, and Y. M. Ro, "Modality conversion for universal multimedia services," in *Proc. 5th Int. Workshop Image Analysis Multimedia Interactive Services*, Lisboa, Portugal, Apr. 2004.

[30] S. F. Chang and A. Vetro, "Video adaptation: Concepts, technologies and open issues," *Proc. IEEE—Special Issue on Advances in Video Coding and Delivery*, vol. 93, no. 1, pp. 148–158, Jan. 2005 .

[31] F. Pereira and I. Burnett, "Universal multimedia experiences for tomorrow," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 63–73, Mar. 2003.

[32] B. L. Tseng, C.-Y. Lin, and J. R. Smith, "Using MPEG-7 and MPEG-21 for personalizing video," *IEEE Multimedia*, vol. 11, no. 1, pp. 42–53, Jan.-Mar. 2004.

[33] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *ACM Multimedia Syst. J.*, vol. 9, no. 4, pp. 353–364, 2003.

Anthony Vetro (S'92–M'96–SM'04) 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 Senior Team Leader. His current research interests are related to the encoding and transport of multimedia content, with emphasis on video transcoding, rate-distortion modeling and optimal bit allocation. He has published more than 90 papers in these areas and holds a number of U.S. patents. Since 1997, he has been an active participant in MPEG, contributing to the development of the MPEG-4 and MPEG-7 standards. Most recently, he served as editor for Part 7 of MPEG-21, Digital Item Adaptation.

Dr. Vetro has been a member of the Technical Program Committee for the IEEE International Conference on Consumer Electronics since 1998 and has served the conference in various capacities. He has been a member of the Publications Committee of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS since 2002 and elected to the AdCom of the IEEE Consumer Electronics Society from 2001 to 2003. He is an Area Chair for ICME 2005, Special Session Chair for VCIP 2005, and Chair of the SPIE Multimedia Systems and Applications conference in 2004–2005. He was a member of the Editorial Board for the *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* from 2001 to 2004. He is a member of the Technical Committee on Visual Signal Processing and Communications of the IEEE Circuits and Systems Society. He served as Guest Editor (with C. Christopoulos and T. Ebrahimi) for the special issue on Universal Multimedia Access of *IEEE Signal Processing Magazine*. He has also received several awards for his work on transcoding, including the 2003 IEEE Circuits and Systems CSVT Transactions Best Paper Award.

Christian Timmerer received the Dipl.-Ing. degree in applied informatics from the Department of Information Technology (ITEC), University of Klagenfurt, Klagenfurt, Austria, where he is currently pursuing the Ph.D. degree in multimedia adaptation.

He joined the University of Klagenfurt in 1999 and is currently a University Assistant and Chairs the IT Administration Group of the Department of Information Technology. He has been working on coding-format agnostic resource adaptation within the MPEG-21 Multimedia Framework. Other research interests include the transport of multimedia content, multimedia adaptation in constrained and streaming environments, and distributed multimedia adaptation. He has been participating in the work of ISO/MPEG for several years, notably as the Deputy Head of the Austrian delegation, Coordinator of several core experiments, Co-Chair of several *ad-hoc* groups, and as editor for Part 7 and 8 of MPEG-21, Digital Item Adaptation and Reference Software.