MCDL: A REDUCED BUT EXTENSIBLE MULTIMEDIA CONTENTS DESCRIPTION LANGUAGE

Marco Furini
Computer Science Department
University of Piemonte Orientale
Via Bellini 25/G - 15100 Alessandria Italy
marco.furini@mfn.unipmn.it

ABSTRACT
Multimedia contents are more and more present in our life and their presentation usually rests on a text-based description that defines rules and properties of such contents and is produced by markup description languages (e.g., XML, MPEG7-DDL and SMIL). Although effective, the criticism to such approach regards the length of the produced description, which is considered too verbose. Since this may lead to performance problems for devices with limited resources, binary-based descriptions are being proposed. However, the simplicity of a text-based description cannot be underestimated and hence, the contribution of this paper is the proposal of a reduced, but extensible, markup Multimedia Contents Description Language (MCDL), which produces short text-based description of multimedia contents. MCDL is designed to organize and synchronize contents for the mobile music scenario and a comparison study with other languages shows that MCDL describes the same contents with a much shorter description.

KEY WORDS
Multimedia Contents Description, SMIL, MPEG7-DDL, Contents Synchronization

1 Introduction
Multimedia contents are entering the way we live, we work and we have fun. Web services, distance learning, online games, video streaming, on-line radio and emusic are only some examples of applications that are composed of multimedia objects.
The large availability of multimedia contents is creating the need of having efficient methods for managing these multimedia contents. Markup multimedia description languages are proposed to this aim: they produce a content-neutral text-based description in order to represent various modalities of the multimedia contents (video, images, audio, multilingual text) and to specify multiple abstraction levels (raw data, features, semantics, and meta-data).

The most popular description language is XML (eXtensible Markup Language) [1], which is a standard for preserving, transmitting and processing documents and complex data structures. XML uses text-based commands (named tags) to specify and to describe the contents; the resulting description is text-based, easy to interpret either by humans or by browsers and is quite similar to HTML (although in XML one can define his own tags, while HTML has only pre-defined tags).

The XML approach has been considered also in proposals related to the description of multimedia contents: SMIL (Synchronized Multimedia Integrated Language) [2] and MPEG7-DDL (MPEG7- Description Definition Language) [3] are the most popular languages designed to facilitate both the creation of multimedia presentations and the methods for searching, indexing and managing multimedia contents (audio, video, animations, graphics, texts, etc.).

Regardless of the used multimedia language, the resulting text-based description has some characteristics in common: i) it provides content authors with the ability to exchange their contents with other authors in an easy way; ii) it is easy to read by humans and machines; iii) it can be easily used by search engines to index and manage multimedia contents.

Although the benefits of having a text-based multimedia contents description are remarkable, this approach is criticized since the resulting description is usually too verbose [4, 5]. This may arise performance issues, as the parsing of such description may be complex, time-consuming and in general not suitable for scenarios where resource-limited devices are used (for instance, in a mobile scenario, devices may have memory and computational power constraints).

For this reason, instead of a text-based description, the production of a binary-based description is being considered. Two main problems arise with this approach: the need of a standard format and the impossibility for humans to directly read/edit the resulting description. Without a standard format, it is likely that a proliferation of different binary formats will happen, leading to interoperability problems. Without a text-based description, an additional overhead in handling the description will be introduced as a tool will be needed to convert a binary description into ASCII and back again.

It is worth noting that, currently, a binary-XML version is under development (but so far, a standard version
is not present) and the MPEG7 group has also introduced a binary-based description, named BiM (Binary Format for MPEG7 description). From this consideration, it seems that the direction toward binary-based description is clear. However, it is also worth mentioning that the MPEG4 group initially introduced a binary-based description (named BIFS, Binary For Scenes) in the MPEG4 standard, but, recently, they introduced the XMT (Extensible MPEG-4 Textual) format, a language that produces a textual syntax representation of multimedia contents. This highlights that the benefits of a text-based description are important and hence a language that produces short text-based description would be worth developing.

The contribution of this paper is the proposal of a multimedia contents description language designed to manage multimedia contents for the mobile music scenario. This scenario requires an organization of different multimedia contents (images, cd-cover, links to external resources, lyrics, karaoke-like service) and since it is mainly composed of resource-limited devices, it requires a short description of such contents.

Our proposed language is named MCDL (Multimedia Contents Description Language) and is provided with a small set of pre-defined tags that are designed to handle multimedia contents in the mobile music scenario. In particular, it describes multimedia contents and their temporal behavior in a short text-based description. The presence of few tags causes the parsing task to be an easy task with a negligible processing time and hence it is suitable for a resource-limited scenario. A comparison study with other multimedia contents description languages (e.g., with SMIL and MPEG7-DDL) is carried out and results show that the same multimedia contents can be described in a much shorter way with our MCDL. However, it is to point out that MCDL is meant to be an alternative language for describing contents inside the mobile music scenario and that it is not a replacement of other XML-based languages, which remain powerful and general contents description languages.

The reminder of this paper is organized as follows: in Section 2 we briefly present the characteristics of SMIL and MPEG-DDL; in Section 3 we present details of our MCDL proposal; Section 4 shows some fields of application of our MCDL. Conclusions are drawn in Section 5.

2 Background

In this section, we briefly present the characteristics of two of the most popular contents description languages: SMIL and MPEG7-DDL. Both languages produce a text-based description that can specify the spatial layout of different media elements (video, audio, graphics, text) as well as the temporal order in which these elements will be played out during the presentation. In addition, a multimedia description may also contain some references (e.g., links) to the media elements composing the presentation.
SMIL (Synchronized Multimedia Integration Language) [2] is an XML-based language that can be used to describe the spacial and temporal behavior of a multimedia presentation. It defines placement and time synchronization of different media objects (audio, video, still pictures, still text, text stream and animations).

A SMIL description consists of two parts: a header and a body. The header describes parameters like window size, multimedia element position, etc., while the body contains information about the multimedia objects, period of the object presentation, etc.

Table 1 shows a very simple example of a SMIL file that describes a karaoke application: the par tag defines a simple time grouping in which multiple elements can play back at the same time (in this case, one task is the play out of song.wma and the other is defined inside the seq tag). The seq tag defines a sequence of elements in which elements play one after the other. The tag text defines a media object and its properties. In this case the text You can call me a sinner is defined and its timing is described using the begin (it defines when the element becomes active) and dur (it specifies the time duration) attributes. With such description, when a SMIL media player executes the SMIL file, it plays out the audio file (in this case song.wma) and can display lyrics according to the specified timing descriptions.
2.2 MPEG-7 DDL

MPEG7-DDL (MPEG7 Description Definition Language) [3] has been designed to provide detailed formatting information and fine-grained descriptions of the structural, and low-level audio, visual and audiovisual features of multimedia contents. It provides a rich set of standardized tools to enable audiovisual descriptions and is based on the XML; the resulting description can be expressed in a textual form (i.e., human readable) or in a compressed binary form (BiM, Binary Format for MPEG-7, designed for storage or transmission).

Table 2 shows an example of MPEG7-DDL description for presenting an image and for providing a karaoke service. The contents description begins with `<mpeg7>` and ends with `</mpeg7>`. Between the two tags, several tags may be used to describe the multimedia contents. Images can be described through the tags `<image>...</image>`. The image is located through the `<MediaLocator>` and `<MediaUri>` tags.

To provide a karaoke service, timing properties have to be attached to lyrics. To this aim, the song has to be decomposed in audio segments (specified within the `<AudioSegment>...</AudioSegment>` tags). Each audio segment specifies its showtime and its duration (through the tags `<MediaTime>...</MediaTime>`) and the text related to such segment (through the tags `<TextAnnotation>` and `<FreeTextAnnotation>`).

From this basic description, it follows that an MPEG7-DDL description may be too verbose for describing some contents (like the karaoke service).

3 Multimedia Contents Description Language

In this section we present details of MCDL (Multimedia Contents Description Language), a markup description language that we propose to describe multimedia contents for the mobile music scenario. The main motivation behind our proposal is that the mobile music scenario is characterized by the presence of resource-limited devices and that in the near future emusic is likely to become a multimedia rich product with different media objects to render and to synchronize. In fact, in the near future, an emusic file will not be a simple audio track as it is today, but it will provide images, cover, links to websites, updated information, lyrics, karaoke, etc. All these multimedia objects have to be synchronized with the audio stream, and hence a multimedia contents description should be provided.

Due to the massive presence of resource-limited devices (e.g., cellphones) the description should be as short as possible in order to keep low the introduced overhead. As we previously mentioned, existing languages like SMIL and MPEG7-DDL can be used to this aim. However, as we show in Section 4, when using these languages, the resulting description is too verbose; the reason is that these languages may require several tags for describing
Table 2: The use of MPEG7-DDL.
very simple media contents like images, CD-cover and karaoke-like service. While the overhead in terms of time and memory usage can be tolerated in a classic computer/Internet scenario, it may arise problems in the mobile music scenario, where users are provided with a device with limited storage capacity and processing power and where bandwidth is usually paid upon consumption. An alternative to SMIL and MPEG7-DDL is the usage of binary formats, but this would introduce a different overhead, as a binary to text tool (and viceversa) should be provided to allow humans to read and to modify the described data.

For these reasons, we propose an alternative markup description language provided with a small set of tags that are designed to specifically describe the multimedia contents of the mobile music scenario. Since the language is designed ad-hoc, it produces a very short multimedia contents description. In summary, MCDL is provided with the following characteristics:

1. the language is designed to describe multimedia contents for the mobile music scenario; in particular, it describes and provides time synchronization to the following media objects: images, cover, links to websites, updated information, lyrics and karaoke;

2. the resulting multimedia contents description is text-based;

3. the set of tags is extensible for introducing additional multimedia contents description;

In other words, MCDL is designed to provide a very simple textual representation of the multimedia contents that can be associated with emusic (images, CD-cover, updated information and karaoke-like service) and the resulting description has to be much shorter than the one obtained from describing the same contents with SMIL and MPEG7-DDL.

### 3.1 Multimedia Contents Description

MCDL is a markup description language and hence it is provided with a set of tags in order to describe multimedia contents. Before going into details of the tags, here we present the general characteristics of MCDL.

MCDL is not case sensitive and uses tags with attributes and values to describe multimedia contents and timing capabilities: tags are enclosed between angle brackets and are in the form `<tag attribute=value>`, with the exception of tags that do not have attributes. The meaning of tags, attributes and values is the following:

- **Tag**: Defines the purpose of the description. The tag name comes after a left angle bracket. Most of the tags have attributes and some tags may consist of just the name. If the tag consists of a pair, the second tag (or closing tag) consists of just the tag name preceded by a slash (for instance, `</body>`). A closing
tag never has attributes. In Table 3 we show tags designed to describe the services proposed in this paper; it is worth noting that additional tags may be introduced to describe other contents, without any problem.

- **Attribute**: Each attribute defines one aspect of the tag; Table 4 shows the possible attributes: href locates a resource; unit and time are used to define the timing attribute of the tag; song is used for files that contains multiple songs; secure is used to inform the player whether the MCDL description is encrypted or not (TRUE or FALSE value). Attributes are followed by a equal sign (=) and by a value;

- **Value**: All values may be integers, names or boolean, depending on the attribute. Names must be enclosed in double quotation marks.

In the following, we describe how multimedia contents can be described using the tags provided in Table 3. In particular, we focus here on describing multimedia contents for the mobile music scenario: images, CD-cover, updated information, karaoke-like service and additional information.

**Time Granularity**: Since multimedia contents may need to be synchronized with the audio stream, it is
Table 4: Standard Tag Attributes.

<table>
<thead>
<tr>
<th>Tag Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>href=&quot;URL&quot;</td>
<td>Gives a fully qualified HTTP URL.</td>
</tr>
<tr>
<td>unit=value</td>
<td>Gives the value of the time unit for the audio-lyrics timing relation. In ms.</td>
</tr>
<tr>
<td>time=value</td>
<td>Gives the timing property of the associated tag.</td>
</tr>
<tr>
<td>song=&quot;title&quot;</td>
<td>Gives the title of the song if multiple songs are encoded in the same MP3 file</td>
</tr>
<tr>
<td>secure=value</td>
<td>Switches the player to security mode.</td>
</tr>
</tbody>
</table>

important to provide an abstraction of the time. In fact, some songs (e.g., rap songs) may require a finer time granularity than other songs (e.g., jazz or opera songs). MCDL allows specifying the time granularity through the <body> tag, where the attribute unit is used to define the length of a single time unit: it is defined in term of milliseconds (e.g., <body unit=100> means that a single time unit corresponds to 100 ms).

Audio Play Out: The emusic file to be played out can be described with the audio tag. The MCDL syntax is <audio href = "url">, where url is the attribute that defines the link to access the emusic file.

CD-Cover: The CD-cover is mainly composed of images and has a well defined format (front-cover, back-cover, internal-covers). To avoid an excessive increment of the file size, MCDL links the emusic file with an external resource that represents the CD-cover. In this way, users may access to the CD-cover only if they really want it. The MCDL syntax is <cover href = "url">, where url is the attribute that defines the link to access the CD-cover. For instance, <cover href ="http://somesite/cover.pdf"> links the emusic file with a resource (cover.pdf) located in somesite, while <cover href ="file://somedir/cover.pdf"> links the emusic file with an internal resource. Note that, MCDL does not pose any restriction on the type of resource that represents the CD-cover, as the rendering of the cover will be performed by the emusic player. Hence, the resource should be specified according with the emusic player capabilities.

Image: The emusic file is linked to an external image through the MCDL tag <img href = "url" time=value>, where url is the link to access a particular image and time represents the time at which the image should be displayed by the emusic player. For instance, <img href = "http://somesite/img.jpg" time=34> links the emusic file with a jpg file located in somesite and the emusic player should render this image at time 34. If the image is locally available the word http should be substituted with the word file. Also in this case, the image format depends on the emusic player characteristics and hence MCDL does not pose any restriction on the image type.
**Updated Information:** Although updated information (for instance, singer tour dates and singer information) may be directly stored into the emusic file, an external link should be used; The storing is avoided because these information require to be always updated, otherwise they will soon become obsolete. The MCDL syntax is `<www href="url">`, where url is the link to access the updated information. For instance, `<www href = "http://somesite/page.html">` links the emusic file with a webpage page.html located in somesite, where information are supposed to be updated (for instance, the singer website or an RSS feed).

**Karaoke:** Karaoke is a multimedia entertainment service that is receiving a large interest by providing users with on-screen lyrics information synchronized with audio play out. One of the main concerns in the design of a karaoke system is the adopted synchronization strategy [6, 7, 8, 9] that has to specify the timing property of the song lyrics.

MCDL uses the tag `<kar>` and the attribute `<time>` to describe the timing property. The MCDL syntax is `<kar time=value>text</kar>`, where value is the time at which the text has to be shown (text can be a syllable, a word or a group of word, depending on the karaoke granularity that the description wants to provide to the final user). For instance, “<kar time=49>You can call me a sinner</BR>” means that after 49 time units, the words You can call me a sinner are sung. The `<br>` tag is used as a new line. With these information, the emusic player can exactly know what the singer is singing. An example of MCDL timing description is given in Table 5.

**Lyrics:** No MCDL tags are designed to describe lyrics. Lyrics are the only text not included between angle brackets. Hence, to display lyrics, the emusic player has simply to skip all the MCDL tags. However, since multiple songs may be encoded in the same emusic file (an entire album may be available in a single file), it may be necessary to identify lyrics of the different songs. For this reason, the `<track song="title">` tag is introduced, which specifies the name of the song through the attribute song. For instance, if the tag `<track song="Hold It">` is present, it means that lyrics that follows the tag are related to the song Hold It.

**Additional Information:** While listening to a song, sometimes it would be interesting to have some additional information about the song or about the singer (for instance: where do the singer live? How many album did he/she release?). To this end, MCDL is provided with two particular tags. A first tag, named `add`, identifies the additional information area, which is concluded by the paired closing tag `/add`. Inside the `add` area it is possible to insert additional information (as plain text) and links to external documents. A `time` attribute inside the opening `add` tag is used as always to synchronize the additional information with the audio. To create a link
<body unit=100>
<img href="http://wsite/madonna.jpg">
<www href="http://wsite/madonna.htm">
<cover href="http://wsite/madonna.pdf">
<karm time=49> You can call me a sinner<br>
<karm time=84> You can call me a saint<br>
<karm time=126> Celebrate me for who I am<br>
<karm time=165> Dislike me for what I ain’t<br>
<karm time=226> Put me up on a pedestal<br>
<karm time=243> Or drag me down in the dirt<br>
<karm time=284> Sticks and stones will break my bones<br>
<karm time=316> But your words will never heard<br>
<karm time=362> I’ll be the garden<br>
<karm time=381> You be the snake<br>
<karm time=401> All of my fruit is yours to take<br>
<karm time=441> Better the devil that you know<br>
<karm time=480> Your love for me will grow<br>
...
</body>

Table 5: The use of MCDL to provide a karaoke service.

to another document, instead of using the www tag (it does not provide an anchor), we borrow the <A> tag from the HTML language. The syntax is: <a href="url">Text to be displayed</a>. The <a> tag is used to create an anchor to link from, the href attribute is used to address the document to link to, and the words between the open and close of the anchor tag will be displayed as a hyperlink.

Table 6 shows an example of the above. Suppose we want to provide additional information about the album Confessions on a dance floor, released by Madonna in 2005; Suppose also that we want these information to appear at a particular time (say at time 1310), then the add tag has timing attribute of value 1310 and information are provided as plain text between the add tag pair, while the starting point for browsing the web (here the Madonna web site) is provided by the A tag.

**Security:** Security is an important aspect in multimedia distribution; contents vendors are hesitant to release their contents if a protection mechanism is not provided to ensure the right usage of their contents. For this reason, both multimedia objects and the multimedia contents description may need to be protected against possible piracy.
Confessions on a Dance Floor is the tenth studio album by singer Madonna released in 2005. It debuted at number one on the Billboard 200, Madonna set a new record for Billboard magazine, becoming the tenth number one debut in a row in the year 2005.

Table 6: The use of the tags \(\textit{add}\) and \(\textit{A}\) to provide additional information.

While multimedia objects are usually protected with Digital Right Management (DRM) systems [10], which either wrap contents in a protective software layer or ensure that contents can only be examined by specific software, the multimedia description is usually provided without any protection. This may arise a security issue as lyrics is a type of content that should be protected against piracy or against a malicious user who can modify the stored information to provide wrong multimedia contents as well as wrong audio-lyrics timing synchronization (for instance, it can change the image location in order to link the song to an advertisement, or even offending, image).

To handle this security issue, MCDL is provided with the \(<\textit{encrypted}>..</\textit{encrypted}>\) tags. All the information provided between these tags are encrypted so that the multimedia content description is protected. Note that the encryption method and the keys to encrypt/decrypt the contents are not specified by MCDL, as they depends on the used DRM. The emusic player is notified that the multimedia contents description (or a part of it) is encrypted by the the usage of the \textit{secure} attribute, which has to be inserted inside the \(<\textit{body}>\) tag.

4 Fields of Application

MCDL has been designed to be employed into the mobile music scenario, as it produces short description of multimedia objects and their temporal behavior. Having a short description is necessary because this scenario is mainly composed of resource-limited devices. To analyze the performance of our MCDL, we select some media contents (a cover, one image and the karaoke service, where a timing property was given to each word) and we use three different multimedia description languages (our MCDL, SMIL and MPEG7-DDL). Table 7 shows the additional memory required by the resulting descriptions. Note that we consider different types of emusic song (in this case encoded with Advanced Audio Coding, AAC), from pop to rock, from alternative to dance. The benefits of using MCDL are remarkable as results show that the ratio MCDL:SMIL:MPEG7-DDL is equal to
<table>
<thead>
<tr>
<th>Song (singer)</th>
<th>Simple emusic audio track (bytes)</th>
<th>Contents description (bytes)</th>
<th>Increment (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like It or Not (Madonna)</td>
<td>4.320.396</td>
<td>8.646 (MCDL)</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.769 (SMIL)</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73.085 (DDL)</td>
<td>1.70</td>
</tr>
<tr>
<td>Cruel (T. Amos)</td>
<td>3.990.117</td>
<td>4.174 (MCDL)</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.180 (SMIL)</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.754 (DDL)</td>
<td>0.97</td>
</tr>
<tr>
<td>Daysleeper (R.E.M.)</td>
<td>3.544.538</td>
<td>4.209 (MCDL)</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.017 (SMIL)</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.749 (DDL)</td>
<td>1.04</td>
</tr>
<tr>
<td>American Idiot (Green Day)</td>
<td>2.834.937</td>
<td>4.339 (MCDL)</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.378 (SMIL)</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.109 (DDL)</td>
<td>1.38</td>
</tr>
</tbody>
</table>

Table 7: All songs are AAC encoded with 128kbps. The karaoke timing description is provided for any single word.
Hence, MCDL produces shorter descriptions than SMIL and MPEG7-DDL.

The mobile music scenario is not the only field of application of our MCDL. For instance, the audiobook industry may attach multimedia contents to the digital audiobook version, so that a simple audiobook file would become a media rich product and hence an audiobook could be not only listened to but also watched. The written text may be shown on a display, synchronized with the narrator’s voice (a sort of karaoke effect); images may appear timely on the screen, integrating the narrator’s descriptions; explanations may appear on the display for improved understanding; links to websites would allow a deeper, more informed reading. Similar benefits would be enjoyed also by the learning discipline as the simultaneous written and audio stimuli can help to memorize new words more quickly, as pointed out by several researches (this is mainly due to the simultaneous use of the left part of the brain, used to understand the written language, and to the right part of the brain, used to understand the audio part). The learning of a foreign language will be easier.

The synchronization of audio and images may bring benefits also to the movie industry, as the narration of a screenplay could be coupled with a story board, so that actors may get the immediate picture of what the director intends to do. For instance, a director could send a simple audio file to his/her cast to provide them with a vision of what he/she wants to do.

In summary, the usage of MCDL would bring benefits to all the fields where audio and small devices are involved and to all the users who owns a portable audio player (commuters, students, young people, etc.).

5 Conclusions

In this paper we proposed MCDL, a markup description language designed to produce short multimedia contents description of multimedia objects and their temporal behavior. MCDL has been provided with a small set of tags, specifically designed to describe multimedia contents for the mobile music scenario.

The language provides a text-based description and its performance has been evaluated in a scenario where the same multimedia contents have been described through three different languages (our MCDL, SMIL and MPEG7-DDL). Results showed that MCDL can describe the same contents in a shorter description. In particular, the description is reduced of a factor of 3 with respect to the SMIL language and of a factor of 8 with respect to the MPEG7-DDL.

It is to point out that MCDL has been designed to be an alternative language for a particular scenario (the mobile music scenario); it is not a general purpose description language as SMIL and MPEG7-DDL. However, the usage of MCDL offers great benefits to scenarios where audio and small devices are involved. Since it is likely that this type of scenario will see a large expansion in the next few year, a reduced, but extensible, markup
description language will be worth using.

References


