**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2014/N14728**

**July 2014, Sapporo, Japan**

|  |  |
| --- | --- |
| **Title** | **White Paper on Carriage of Timed Text and Other Visual Overlays** |
| **Author** | Cyril Concolato, Michael Dolan |

With the growing importance of HTML 5 video applications and the definition of new formats for subtitling such as WebVTT [1] and TTML [2], the need is dire to be able to package those subtitles, and additional synchronized graphics overlays such as SVG [3], in file formats that help delivering and streaming them, in particular in the context of adaptive streaming technologies such as MPEG-DASH.

The amendment to the ISO Base Media File Format, ISO/IEC 14496-12:2012/Amd.2:2014[[1]](#footnote-1), defines new tools to carry any timed text stream such as subtitles or synchronized textual graphics in ISO-based media files such as MP4 files. The MPEG-4 Part 30 standard, formally known as ISO/IEC 14496-30:2014, further specifies the carriage of WebVTT and TTML text streams. This paper presents the key concepts of these standards, in particular the timing model. This paper does not address older forms of timed text that can be carried, such as that described in ISO/IEC 14496-17.

# MPEG-4 Part 12 and carriage of subtitles

Part 12 defines the syntax and semantics of an MP4 file. An MP4 file is logically made of tracks. An MP4 track is a logical structure organized into samples and sample descriptions. Samples carry information that is valid from a given time and for a given a duration. Samples carry data that is continuous (no gap in time between samples) and non-overlapping (the end of a sample is the start of the next sample). This has good properties, and in particular allows random access into the track. A sample description carries information that is valid for the duration of several samples, typically for the whole track.

The amendment to Part 12 for Timed Text defines new text track types for a broad range of timed text formats. In particular, two track types have been defined: the ‘text’ type for track content that results in text rendering only; and the ‘subt’ type for track content that may result in text and graphics rendering. In principle, any text-based timed format can be carried now in ISO-based media files.

# MPEG-4 Part 30

## Introduction

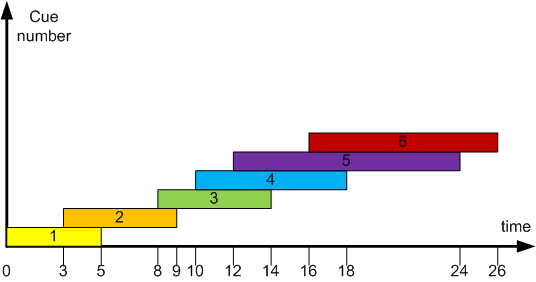
Part 30 provides specific guidance for two popular timed text format technologies defined by W3C – Timed Text Markup Language (TTML) and Web Video Text Tracks (WebVTT) enabling use of those formats in context such as MPEG-DASH or HTML5 Media Source Extensions. It defines also common aspects related to layout, such as where the subtitle region is compared to the video region.

One particular aspect of the carriage of both WebVTT and TTML is the carriage of timing information in the sample data. In both cases each sample contains the elements to be presented during the sample validity. In the case of WebVTT, the time values in each sample are relative to the sample, but the time values used in TTML are relative to the start of the track. Each time base approach has boundary issues for editing and decoding of fragmented files. However, in some applications, such non-live and non-mezzanine (i.e. not edited) it is common to encode the entire track in a single fragment (and thus in the case of TTML also a single XML document). This is due to the relative size of text tracks compared to video and audio. This single fragment in a track encoding practice devolves to both timebase approaches being equivalent.

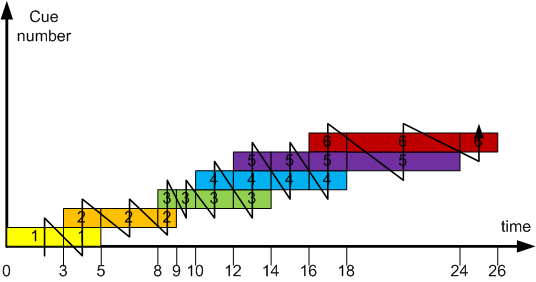
## Carriage of WebVTT

WebVTT content is carried in MP4 files using ‘tracks’, of type ‘text’. WebVTT header and metadata are carried in the sample description while WebVTT cues are in samples. Tracks of this type can be identified by the MIME type codecs sub-parameter “wvtt”.

One particularity of WebVTT is that cues may overlap. To enable carriage of overlapping WebVTT cues in MP4 tracks, WebVTT cues need to be split into non-overlapping cues and gathered into samples, as explained below. MP4 Parsers will typically do the reverse operation so that the carriage in MP4 is transparent to the application. More generally, the carriage has been designed such that the WebVTT content after import/export in an MP4 is identical, including comments and text content that is not valid according to the syntax but processable by a conformant WebVTT parser.



As an example, if you take the cues as depicted in the figure above, the cues will be split and organized into samples as depicted in the figure below. Cue 1 is split into 2 cues (1a and 1b, the boundary being at time 3). Cue 1a becomes sample 1. Cue 2 is split into 3 cues (2a, 2b and 2c: boundaries being at time 5, and time 8). Cue 1b and Cue 2a form Sample 2. Cue 2b forms Sample 3. And so on.



Additionally, special care has been taken to avoid cue timing information duplication. In particular, the cue start times and end times are stored at the sample level, not at the cue level, enabling editing of the track, with the same tool as a video track, including when cues contain internal timestamps.

In terms of physical bytes, WebVTT data (header, cues, …) is wrapped into ISO structures called boxes. There is a box for header and metadata, for cue ids, cue settings, cue payloads, and in-between cue text (such as comments).

WebVTT, as defined by the W3C, may be also be used to carry synchronized metadata such as JSON content or any other text content. Carriage of such metadata in MP4 files is also enabled with the defined carriage.

## Carriage of TTML

TTML content is carried in MP4 files using ‘tracks’ of type ‘subt’. A TTML sample carries an complete XML document, and may also contain or reference additional resources such as images or fonts, potentially shared between documents/samples. Tracks of this type can be identified by checking the MIME type codecs sub-parameter “stpp”. Because TTML can contain multiple extension namespaces, including ones defined externally to W3C, additional document details can be learned by checking the namespace and XML schema location information provided in the sample description.

For TTML “00:00:00” does not mean the start of the sample in which the document is carried.

So in the example below, the document describes rendering between times 00:31:00 and 00:32:00 but the sample may have a start time before the time 00:31:00 and may last after 00:32:00. Authors should be careful, that elements associated with times outside the sample validity will not be rendered.

<tt>   
 <body>  
 <div>

<p begin=“00:31:00” end=“00:32:00”>31-32 minutes</p>

</div>

</body>

</tt>

Many formats such as EBU-TT-D, SMPTE-TT and CFF-TT derive from TTML and as such can also be carried in ISO-based media files.

# References

[1] “WebVTT: The Web Video Text Tracks Format”, W3C, <http://dev.w3.org/html5/webvtt/>

[2] “Timed Text Markup Language 1 (TTML1) (Second Edition)”, W3C Recommendation 24 September 2013, <http://www.w3.org/TR/ttaf1-dfxp/>

[3] “Scalable Vector Graphics (SVG) 1.1 (Second Edition)”, W3C Recommendation 16 August 2011, <http://www.w3.org/TR/SVG/>

1. Soon to be published as 14496-12:2014, 4th Edition [↑](#footnote-ref-1)