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/M35091
October 2014, Strasbourg, France

Source	**Telecom ParisTech, GPAC Licensing**
Status	**For consideration at the 110th MPEG Meeting**
Title	**Clarification on carriage of TTML in ISOBMFF**
Author	Romain Bouqueau, Cyril Concolato, Jean Le Feuvre

1 Introduction

As part of the implementation of the support for TTML in MP4 in GPAC, we have discovered some problems regarding the inference of the size to be set in the 'tkhd' ISOBMFF box according to MPEG-4 part 30. This contribution proposes to clarify MPEG-4 part 30.

2 Analysis

MPEG-4 Part 30 has the following 3 problems.

2.1 Video information not available at packaging time

The carriage of timed text assumes that a video is available in the same file as the subtitle track. This is not always the case, for instance when using unmultiplexed representations in DASH. This is not strictly necessary at packaging time. The actual size is only needed at playback time and is either explicit in the subtitle stream (e.g. using px) or resolved at playback time (because it either does not provide an explicit size or uses percentages).

2.2 Using "Video" is too specific

In section 4.1 of MPEG-4 part 30, it is written:
Unless specified by an embedding environment (e.g. an HTML page), the track header box information (i.e. width, height) shall be used to size the subtitle or timed text track content with respect to the video; otherwise, it may be ignored by the embedding environment.

MPEG-4 part 30 makes the assumption there is always a "video". This is not always the case. The visual element might not be a video. For instance it might be a graphical animation or a sound track for which a timed text track is provided. Therefore the phrasing should be more general and use the term "visual presentation size" as used in the Track Header Box.

2.3 TTML attributes not always sufficient to determine track size

TTML, as a self "embedding environment", proposes a logical path to infer the value of 'width' and 'height'.

In TTML [http://www.w3.org/TR/ttaf1-dfxp/], the value of 'width' and 'height' are not easy to determine and are inferred differently depending on the scenario as follows (in section 7.7.1):

> "If the `tts:extent` attribute is specified on the `tt` element, then it must adhere to 8.2.7 tts:extent, in which case it specifies the spatial extent of the *Root Container Region* in which content regions are located and presented."

So, when the tts:extent attribute is specified (pixels units are mandatory in this case), the values to be put in tkhd.width/height are clear.

> "If no `tts:extent` attribute is specified, then the spatial extent of the *Root Container Region* is considered to be determined by the *Document Processing Context*. The origin of the *Root Container Region* is determined by the *Document Processing Context*."

So, when the tts:extent attribute is not specified, the TTML specification refers to the Root Container Region as defined by the Document Processing Context, defined as follows:

> **"Document Processing Context** The implied context or environment internal to a *Content Processor* in which document processing occurs, and in which out-of-band protocols or specifications may define certain behavioral defaults, such as the establishment or creation of a *Synthetic Document Syncbase*."

Does an MP4 player correspond to the "implied context or environment"? The definition of Synthetic Document Syncbase below does not provide further information with regards to tts:extent.

> **"Synthetic Document Syncbase** A document level syncbase [SMIL 2.1], § 10.7.1, synthesized or otherwise established by the *Document Processing Context* in accordance with the *Related Media Object* or other processing criteria."

However, a note provides additional information (but as a note):

> "Note: In the absence of other requirements, and if a Related Media Object exists, then it is recommended that the Document Processing Context determine that:
>
> - if no tts:extent is specified on the root tt element, the extent of the Root Container Region be established as equal to the extent of the Related Media Object Region; and
> - the origin of the Root Container Region be established so that this region is centered in the **Related Media Object Region**."

Then, it seems clear that if a "Related Media Object" exists, we could use its size. The definition of this term is present in the SMIL 2.1 specification:

> **"Related Media Object** A (possibly null) media object associated with or otherwise related to a Document Instance. For example, an aggregate audio/video media object for which a Document Instance provides caption or subtitle information, and with which that Document Instance is associated."

Conclusion:

It seems therefore clear that if an ISOBMFF file contains both a visual track and an associated TTML track, the Root Container Region is the visual element region and therefore that the tkhd.width/height of the TTML track shall be equal to those of the visual element.

What happens when two video tracks are present in the file? How is the association between the tracks made?

To the contrary, it is unclear when the Related Media Object is not present, i.e. when no visual track is present in the ISOBMFF file.

Additional element:

Furthermore, the IMSC specification (http://www.w3.org/TR/2014/WD-ttml-imsc1-20140930/) uses the ittp:aspectRatio attribute defined as follows:

> "If `ittp:aspectRatio` is present, the root container SHALL be mapped to a rectangular area within the related video object
>
> i. the ratio of the width to the height of the rectangular area is equal to `ittp:aspectRatio`,
> ii. the center of the rectangular area is collocated with the center of the related video object frame,
> iii. the rectangular area (including its boundary) is entirely within the related video object frame (including its boundary), and
> iv. the rectangular area has a height or width equal to that of the related video object frame."

So, this attribute should be taken into account when computing the tkhd.width/height.

3 Proposal

The followed clarifications are proposed for section 4.1 of MPEG-4 part 30:

3.1 Default width and height

Add the following sentence as the 3rd sentence in 4.1:

"The width and height of the subtitle or timed text track may be set to 0 when no size information is provided in the track content".

As a consequence, the following signaling would be possible:

	The timed text content does not have an explicit size and is applicable with any size of video	The timed text document has an explicit size w, h
applicable to	- TTML with no tts:extent - WebVTT - SVG tracks with no width/height or with 100%/100%	- TTML with tts:extent="640px 480px" - SVG with width='640' height='640'
Track Width/Height	0/0 or any couple of values matching the TTML ittp:aspectRatio (if any), or the SVG viewBox aspect ratio (if preserveAspectRatio)	640/480
DASH Representation	no @width, no @height if 0/0, or the given tkhd values	width="640" height="480"

3.2 Track Reference

Define the following track reference

"subt - used on timed text and subtitle tracks to indicate to which track the timed text/subtitle track applies "

3.3 Video vs. visual object

In section 4.1, replace:

" the track header box information (i.e. width, height) shall be used to size the subtitle or timed text track content with respect to the video"

with

" the track header box information (i.e. width, height) shall be used to size the subtitle or timed text track content with respect to the associated visual object, if any"

Replace:

"The width and height of the subtitle or timed text track should be appropriate for the width and height of the video track (as declared in the track header) it is intended to overlay, even if the video is not stored in an ISOBMFF file or stored as a track in a different ISOBMFF file."

with

" The width and height of the subtitle or timed text track should be appropriate for the width and height of the visual track (as declared in the track header) it is intended to overlay, even if that track is not stored in an ISOBMFF file or stored as a track in a different ISOBMFF file."

Replace:

"Note – timed text and subtitle tracks are normally stacked in front of the video."

with

"Note – timed text and subtitle tracks are normally stacked in front of the associated visual object."

In 4.2, replace:

" The `timescale` field in the media header box should be set appropriately to achieve the desired timing accuracy; it is recommended to be set to the value of the `timescale` field in the corresponding video track’s media header box."

with

" The `timescale` field in the media header box should be set appropriately to achieve the desired timing accuracy; it is recommended to be set to the value of the `timescale` field in the media header box of the corresponding track, potentially identified by means of a "subt" track reference."

4 Conclusion

We recommend MPEG to adopt the proposed text in a corrigendum to 14496-30 and to define the new track reference in an amendment to Part 12.

5 Acknowledgments

This work has been partially funded by the EU ICT-KIC and by EBU.