**ISO/IEC JTC 1/SC 29/WG 03 N0639**

**ISO/IEC JTC 1/SC 29/WG 03  
MPEG Systems   
Convenorship: KATS (Korea, Republic of)**

**Document type:** Output Document

**Title:** Technologies under Consideration for OMAF

**Status:** Approved

**Date of document:** 2022-07-26

**Source:** ISO/IEC JTC 1/SC 29/WG 03

**No. of pages:** 5 (with cover page)

**Email of Convenor:** young.L@samsung.com

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg3>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 03 MPEG SYSTEMS**

**ISO/IEC JTC 1/SC 29/WG 03 N0639**

**July 2022, Virtual**

|  |  |
| --- | --- |
| **Title** | **Technologies under Consideration for OMAF** |
| **Source** | **WG 03, MPEG Systems** |
| **Status** | **Approved** |
| **Serial Number** | **21768** |

# Introduction

This document includes technologies under consideration for ISO/ IEC 23090-2 (OMAF)

This includes the following:

1. support server-side dynamic adaptation (m60221, m59581)

# Support Server-side Dynamic Adaptation

This contribution proposes an updated OMAF architecture which is described in 4.2 overall architecture of OMAF specification to deal with above SSDA(m60221). And It includes additional function to send the viewport information to server-side processing(m59581).

The proposed text is marked in yellow with removed text in ~~red strike through~~ on top of the current OMAF spec

## Overall architecture (OMAF clause 4.2) (m60221)

Figure X shows a typical content flow process for an omnidirectional media application.

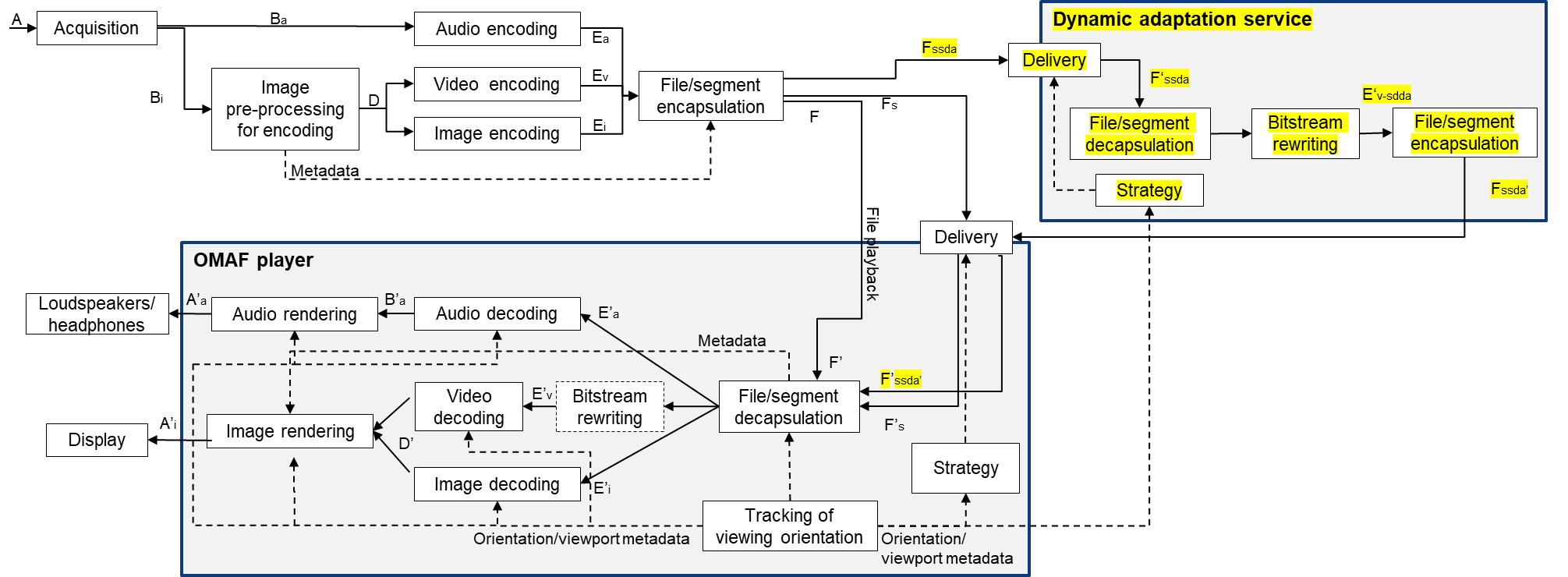


Figure X — Content flow process for omnidirectional media with projected video

The following interfaces are specified in this document:

* E'a, E'v, E'i: audio bitstream, video bitstream, coded image(s), respectively; see Clause10.
* E'v-sdda : video bitstream, see Clause 10.
* F/F': media file; see Clause7. Moreover, media profiles specified in Clause 10 include the specification of the track formats for F/F', which may contain constraints on the elementary streams contained within the samples of the tracks.
* Clause 8 specifies the delivery related interfaces for DASH delivery.
* Clause 9 specifies the delivery related interfaces for MMT delivery.

The other interfaces in Figure X are not specified in this document.

NOTE While the syntax and semantics of the bitstreams Ea, Ev, and Ei are the same as those for E'a, E'v, E'i, respectively, the input interface to the file/segment encapsulation module is not specified.

NOTE While the syntax and semantics of the bitstreams Ev, is the same as for E'v-sdda the input interface to the file/segment encapsulation module is not specified.

A real-world audio-visual scene (A) is captured by audio sensors as well as a set of cameras or a camera device with multiple lenses and sensors. The acquisition results in a set of digital image/video (Bi) and audio (Ba) signals. The cameras/lenses typically cover all directions around the centre point of the camera set or camera device, thus the name of 360-degree video.

Audio may be captured using many different microphone configurations and stored as several different content formats, including channel-based signals, static or dynamic (i.e. moving through the 3D scene) object signals, and scene-based signals (e.g. Higher Order Ambisonics). The channel-based signals typically conform to one of the loudspeaker layouts defined in ISO/IEC 23091-3. In an omnidirectional media application, the loudspeaker layout signals of the rendered immersive audio program are binauralized for presentation via headphones.

For audio, no stitching process is needed, since the captured signals are inherently immersive and omnidirectional.

This document specifies the following types of omnidirectional video and images, which differ in the architecture in the image pre-preprocessing for encoding and in the image rendering processing blocks:

* Projected omnidirectional video/images:
* Image pre-processing for encoding: The images (Bi) of the same time instance are stitched, possibly rotated, and projected onto a 2D picture coordinates using a mathematically specified projection format. Optionally, the resulting projected pictures may be further mapped region-wise onto a packed picture. Either projected pictures or packed pictures are subject to video or image encoding.
* Image rendering: Either regions of the decoded packed pictures (if region-wise packing has been applied) or the entire projected picture (otherwise) is mapped onto a rendering mesh suitable for the projection format in use.
* Fisheye omnidirectional video/images:
* Image pre-processing for encoding: Circular images (Bi) captured by fisheye lenses are arranged onto a 2D picture, which is then input to video or image encoding.
* Image rendering: The decoded circular images are stitched using the signalled fisheye-specific parameters.
* Mesh omnidirectional video:
* Image pre-processing for encoding: A 3D mesh consisting of mesh elements is generated, where mesh elements can be either parallelograms or regions of a sphere surface. The images (Bi) of the same time instance are stitched, possibly rotated, and projected onto the 3D mesh. Mesh elements are mapped onto rectangular regions of one or more 2D pictures, which are input to video encoding.
* Image rendering: Rectangular regions of the decoded 2D picture(s) are mapped to the 3D mesh, which is used directly as the rendering mesh.

Further details of the architecture for projected, fisheye, and mesh omnidirectional video/images are provided in subclauses4.3, 4.4, and4.5, respectively.

The pre-processed pictures (D) are encoded as coded images (Ei) or a coded video bitstream (Ev). The captured audio (Ba) is encoded as an audio bitstream (Ea). The coded images, video or audio are then composed into a media file for file playback (F) or a sequence of an initialization segment and media segments for streaming (Fs, Fssda, and Fssda’), according to a particular media container file format. In this document, the media container file format is the ISO Base Media File Format specified in ISO/IEC 14496-12. The file encapsulator also includes metadata into the file or the segments.

The segments Fs are delivered using a delivery mechanism to a player or using a server-side dynamic adaptation mechanism to a player.

In the server-side adaptation mechanism, the received segments (F'ssda) and extracts the coded bitstream and parses metadata. Viewport-dependent video is carried in multiple tracks, which is merged in the bitstream rewriting into a single video bitstream (E'v-sdda). This single video bitstream(E'v-sdda) is composed into a sequence of an initialization segment and media segments for streaming(Fssda’)’

The file that the file encapsulator outputs (F) is identical to the file that the file decapsulator inputs (F'). A file decapsulator processes the file (F') or the received segments (F's and F’ssda’) and extracts the coded bitstreams (E'a, E'v, or E'i) and parses the metadata. Viewport-dependent video may be carried in multiple tracks, which may be merged in the bitstream rewriting into a single video bitstream E'v prior to decoding. The audio, video or images are then decoded into decoded signals (B'a for audio, and D' for images/video). In the image rendering block, the decoded pictures (D') are projected onto the screen of a head-mounted display or any other display device based on the metadata parsed from the file. Likewise, decoded audio (B'a) is rendered, e.g. through headphones, according to the current viewing orientation. The current viewing orientation is determined by the viewing orientation tracking functionality. When a head-mounted display is in use, the viewing orientation tracking can involve head tracking and possibly also eye tracking. When sphere-relative overlays are in use, the viewing orientation tracking functionality can include or be complemented by viewing position tracking and rendering of overlays with background visual media can take both the viewing position and the viewing orientation into account. Besides being used by the renderer to render the appropriate part of decoded video and audio signals, the current viewing orientation may also be used by the video and audio decoders for decoding optimization. In viewport-dependent delivery, the current viewing orientation is also passed to the strategy module in OMAF player or Dynamic adaptation service, which determines the video tracks to be received based on the viewing orientation.

The process described above is applicable to both live and on-demand use cases.

## additional function to send the viewport information to server-side processing (m59581))

The proposal text is below:

x.x.x Signalling of server-side viewport adaptation information

A **EssentialProperty** element with a @schemeIdUri attribute equal to "urn:mpeg:mpegI:omaf:2022:serverSideAdaptation" is referred to as a server-side viewport adaptation (SSVA) descriptor.

When SSVA descriptor is exist in AdaptationSet or Representation, the query parameter which indicate the SphereRegion is added to the segment file URL. The values of indication of Sphere Region are centre\_azimuth, centre\_elevation, centre\_tilt, azimuth\_range, and elevation\_range.

Descriptions of these values are same as the values of cc.coverageInfo.

ex. http://hoge.com/vvc.mp4?centre\_azimuth=9830400&centre\_elevation=3932160&centre\_tilt=0& azimuth\_range=3932160&elevation\_range=3932160

----