



# MPEG-I Scene Description

---

Authors: Imed Bouazizi, Lukasz Kondrad, Yago Sanchez de la Fuente, Emmanuel Thomas, Emmanouil Potetsianakis, Mary-Luc Champel, Gurdeep Bhullar, Philippe Guillotel, and Thomas Stockhammer

White Paper published November 4, 2022

## Abstract

The Moving Pictures Expert Group (MPEG), a collection of several working groups of ISO/IEC JTC1 SC29, has been working on technologies and standards for immersive media under the umbrella of the MPEG Immersive project (MPEG-I). ISO/IEC JTC1 SC29 WG03, much better known as MPEG Systems, recognized the need for an interoperable and distributable scene description solution, as a key element to foster the emergence of immersive media services and to enable the delivery of its immersive content in the consumer market. As part of the MPEG-I project, the MPEG Working Group 3 started investigating architectures for immersive media and possible solutions for a scene description distribution format in 2017, which resulted in the ISO/IEC 23090-14 standard. The first edition of the standard was technically completed in summer 2022 and will be published in early 2023.

In this paper, we provide an introduction to the standard, its building blocks, and pointers to additional resources and related standards. We conclude the paper with the outlook and future plans of the standard for the second edition.

## Introduction

Immersive media is becoming increasingly prevalent and is already showing early signs of influence on the way we work and entertain ourselves. Immersion is achieved by introducing the depth dimension in media modalities (visual and auditory) traditionally digitally expressed in a 2D fashion. The trend of transition from 2D to 3D media was initially started by Virtual Reality (VR), mainly driven by the availability of affordable VR Head-Mounted Displays (HMDs). However, unique immersive experiences can be delivered using Augmented Reality (AR), which is becoming popular over time, supported by releases of consumer devices, like see-through HMDs and glasses.

Immersive media applications offer an experience where the user is immersed into virtual or hybrid environments. The user is able to experience the content in 3D and enjoy more degrees of freedom compared to traditional 2D content. Platforms providing immersive media also often give the user the ability to interact with the content and/or with other users in shared virtual spaces.

The need for a solution to enable cross-platform exchange and interaction in 3D environments became evident, and a number of fora and Standards Developing Organizations (SDOs) started

to define required technology. MPEG Systems provides the definition of a scene description framework in ISO/IEC 23090-14 that serves as entry point format to describe to rich 3D scenes, enabling immersion, fusion with the real world, and rich interactivity, while providing real-time media delivery. Figure 1 shows initial prototypes examples of immersive applications, powered by the MPEG-I scene description.



**Figure 1 Demos and Prototypes powered by MPEG-I Scene Description**

ISO/IEC 23090-14 provides a set of vendor-extensions under the `MPEG_` prefix to Khronos glTF<sup>1</sup>, (also available as ISO/IEC 12113), as well as extensions to the MPEG-defined file format, also known as ISO/IEC 14496-12 ISO-BMFF. These extensions enable description and delivery of timed immersive media into glTF-based immersive scenes. Furthermore, the standard defines an architecture together with an application programming interface (API) that allows the application to separate the access to the immersive timed media content from the rendering of this media. The separation and the definition of this API allow developers to implement a wide range of optimization techniques, such as the adaptation of the retrieved media to the network conditions, partial retrieval, access at different levels of detail, and adjustment of the content quality.

## Architecture

Due to its complexity, immersive media cannot be processed using traditional 2D client architectures for example as known from TV lean-back based experiences. Immersive media involves interaction and being significantly more present in the content and, hence it comes with a set of new of technical and architectural requirements. Figure 2Figure 2 depicts the architecture framework defined in ISO/IEC 23090-14 that addresses the aforementioned requirements.

A 2D media player is replaced by a Presentation Engine, which is able to perform multi-modal 3D rendering of a scene that may be composed of audio, visual, and haptics media. The Presentation Engine builds the core of immersive applications. To keep the focus of the Presentation Engine on high-fidelity rendering, the MPEG-I Scene Description standard

---

<sup>1</sup> Khronos Group, glTF : <https://www.khronos.org/glTF/>

introduces and specifies a separate component, the Media Access Function (MAF). The MAF is responsible for the media access, e.g. download, streaming or connecting to other peers, as well as for decryption, decoding and post-processing related functions. While the Presentation Engine is the core interface with the user perception, it delegates the handling of the media access and retrieval to the MAF. The MAF constructs suitable media pipelines to transform the accessed compressed media from a delivery format into so-called *buffer formats* that can be directly rendered by the Presentation Engine. For example, a media pipeline could perform fetching, decoding, decryption, and post-processing of the referenced media. The raw decoded media data is then provided in buffers to the Presentation Engine that will then render the buffers to a 3D experience.

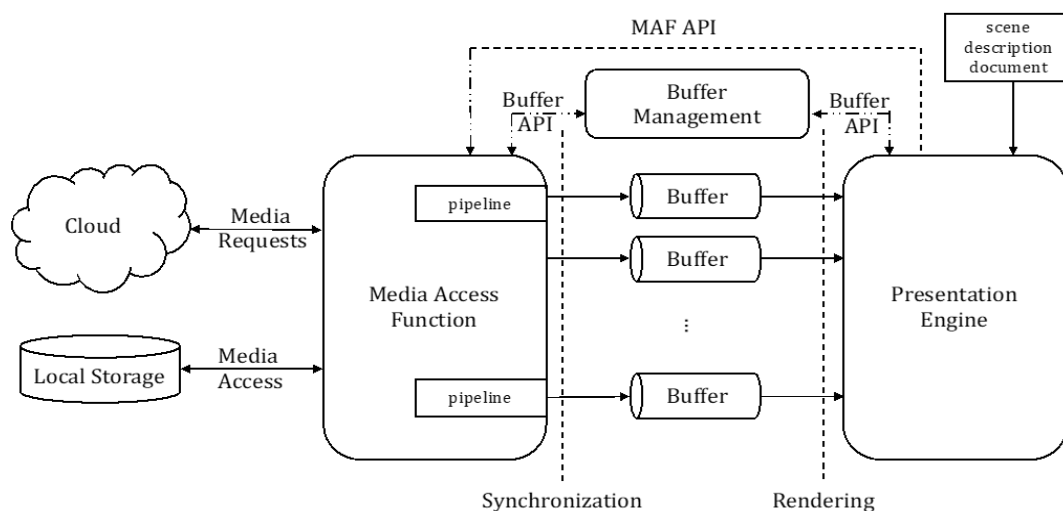


Figure 2: Architecture framework

Among others, the ISO/IEC 23090-14 standard describes different interoperable buffer formats as well as delivery formats for different media types and delivery scenarios. The MAF uses information such as the MIME type and codec parameters to identify support of the media reconstruction and assemble the proper media pipeline. The concepts of the media pipeline, the handling of media in the presentation engine, and the description in MPEG-I Scene description is closely aligned with way how HTML-5 was extended to support real-time media by the introduction of the `<video>` element, as well as the Media Source Extensions (MSE) for media streaming.

## glTF2.0 Extensions

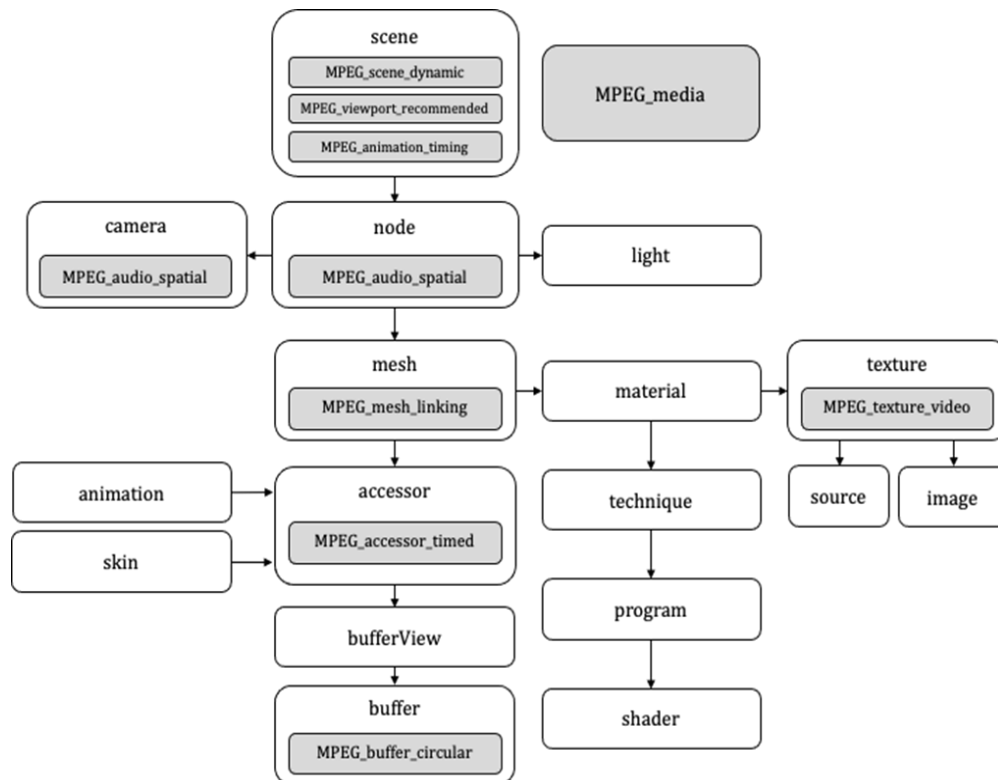
ISO/IEC 23090-14 defines a number of vendor-extensions to Khronos glTF2.0 using the registered prefix `MPEG_` that address the requirements that the MPEG System group had initially identified as essential for the distribution of real-time immersive media content. The extensions can be separated into two groups. The first group is formed with extensions that provide the base description of media used in the scene and its formats, which can be expected by the Presentation Engine:

- `MPEG_media`: an extension listing external media that are referenced in a scene description by other elements. The extension allows creators to describe content in alternative formats, thus offering the ability to select the most appropriate format to access based on the application capabilities.
- `MPEG_accessor_timed`: an extension indicating to an application that the media described by the accessor element is timed. Additionally, the extension may signal to the application presence of dynamic fields that change on a frame-by-frame basis.
- `MPEG_buffer_circular`: an extension indicating that the buffer is used for passing timed media, with concurrent read/write access to the buffer. The extension also links the underlying buffer to a media described by the `MPEG_media` extension. The buffer extension is used to pace the access to the buffer, compensating for any differences between the media flow through the media pipeline in the MAF and the rendering process in the Presentation Engine. The extension also indicates that this extended buffer does not have the original glTF 2.0 restriction of containing only static data.

The above extensions define the fundamental mechanisms used by the scene description, while the second group of extensions builds on top of them and includes the following:

- `MPEG_texture_video`: an extension that provides a mechanism to build materials with dynamic textures by channeling the texture data through buffers described by `MPEG_accessor_timed` and `MPEG_buffer_circular` extensions.
- `MPEG_audio_spatial`: an extension that offers support for spatial audio in a 3D scene by introducing audio sources and an audio listener, where the audio sources reference timed accessors.
- `MPEG_scene_dynamic`: an extension that adds the capability of updating a 3D scene by pointing to an update stream. A sample of the update stream can be a new scene document or a patch document that alters the composition of the scene at runtime. Example usage of this extension is a shared space, where users may join or leave, thus adding and removing their avatar representations to/from the scene.
- `MPEG_viewport_recommended`: an extension that provides a stream of sparse samples that indicate the author recommended pose for viewing the scene. This might especially be useful in cases where the scene is consumed on a non-immersive device, such as 2D display device without interactions.
- `MPEG_mesh_linking`: an extension to link two meshes and provide mapping information. This extension allows, for instance, applying animations to dynamic meshes with changing topologies by defining it as a dependent mesh on a shadow mesh with a static topology. The transformations and required metadata are defined for the shadow mesh and are transferred to the dynamic mesh by using the mapping information defined within the extension.
- `MPEG_animation_timing`: an extension to control animation timelines.

The set of the extensions and their placement in the node hierarchy is depicted in Figure 3.



**Figure 3: MPEG Extensions to glTF2.0 defined in MPEG-I Scene Description first edition**

These extensions have been registered at Khronos and submitted to Khronos to be added to glTF2.0. Khronos's 3D format group had continuously been involved in the development of these extensions and many verbal and written exchanges, including a successful joint workshop<sup>2</sup>, and comments improved the extensions to fit into the Khronos glTF2.0 specification. At the time of publication of this paper, Khronos is in the final review stages to merge MPEG extensions into the main glTF branch <https://github.com/KhronosGroup/glTF>.

## Storage and Transport Formats

In addition to the glTF extensions, ISO/IEC 23090-14 also defines carriage formats related to delivery of scene description data as well as to delivery of data related to glTF2.0 extensions. The carriage format is based on MPEG File Format, i.e. the ISO/IEC 14496-12 ISO-BMFF standard. To facilitate delivery of the scene description to a client, ISO/IEC 23090-14 defines how glTF files and related data can be provided as non-timed and timed (i.e. as samples of track) data encapsulated in an ISO-BMFF file.

A number of extensions, including `MPEG_scene_dynamic`, `MPEG_mesh_linking`, and `MPEG_animation_timing`, indicate that a particular form of timed data is provided to a Presentation Engine during the consumption of the scene, and the Presentation Engine is

<sup>2</sup> Please check <http://mpeg-sd.org/workshop.html>

expected to act based on continuously updated information. ISO/IEC 23090-14 defines the format of the timed data for each of the extensions, the way it can be encapsulated in an ISOBMFF file and how the timing of scene updates relate to media timing.

The `MPEG_media` extension enables referencing of external media streams that are delivered over protocols, such as RTP/SRTP or MPEG-DASH. In order to support referencing media streams without actually knowing the values for the protocol scheme, hostname, or port, ISO/IEC 23090-14 defines a new URL scheme. The scheme requires the presence of a stream-identifier in the query part. However, it does not dictate a specific type of identifier. It allows for the usage of the Media Stream Identification scheme (RFC5888), a labeling scheme (RFC4575), or a 0-based indexing scheme.

## Future Work

With the key requirements for an immersive scene description solution addressed in the first edition of ISO/IEC 23090-14 MPEG-I Scene Description, the MPEG-I Scene Description breakout group focus has now shifted to address more advanced use cases and requirements. Notable new functionalities on the roadmap for the second edition include interactivity, AR anchoring, user, lighting and avatar representation, haptics support, and providing support of immersive visual and audio codecs. The extensions are expected to be technically completed by the end of 2023 and will be published shortly after.

*Interactivity* complements the immersive user experience by allowing users to interact with the virtual environment and the objects therein. Each node of the scene graph may be made interactive by associating some interactivity triggers and associated actions. Different types of triggers and actions will be defined to support different forms of interactivity.

*AR anchoring* aims at enabling authors to create and distribute AR experiences, where virtual scene content is composited with the user's real world. The virtual content is anchored to the real world through anchor points that can be associated with XR spaces and tracked throughout the whole experience.

The *user* needs to be *represented* in a scene, to move and interact with objects and other participants. The representation of the user can be achieved at different levels, ranging from a simple hand to a full body representation (aka *avatar*). Any 3D model can be used to represent users and virtual humans. To ease interoperability and transferability from one environment to another, a generic humanoid model used for the representation of a virtual (or real) user is specified. The user avatar representation considered in the Scene Description will define such a humanoid model, with its associated semantic and topology. The model activity corresponds to a gender-neutral face with both a male and female body model, which can be fully customized.

Regarding MPEG-I *Haptics* format representation and coding, ISO/IEC 23090-31 provides a compression scheme for haptics. It supports various haptic perceptions (vibrations, pressure, temperature, velocity, kinesthetic) and rendering devices. Haptics experience can be

associated with objects and characters to feel the physical properties of an object or some feedback from the interaction with objects and characters. The physical properties can be associated with a 3D object in the scene graph as well as the haptic feedback with interactions defined.

When it comes to inserting visual information in a captured real-world environment such as in AR applications, *lighting* is a fundamental cue to provide a realistic experience to the user. In a VR context, accurate lighting models allow to achieve a high-level of realism which is also key for many VR applications that offer the experience of “being there” to the user. To this end, the integration of lighting information for both real and virtual scenes into scene description documents is currently under study for inclusion in the next edition of the MPEG-I Scene Description standard.

Finally, *MPEG-I immersive codecs*, namely ISO/IEC 23090-5 - Visual Volumetric Video-based coding (V3C) and Video-based compression of Point-Cloud (V-PCC) and ISO/IEC 23090-12 - MPEG Immersive Video (MIV), provide a compression scheme for volumetric content. Such codecs are used to compress 3D objects, which can be positioned in the scene graph. Support for V3C compressed objects will be part of Amendment 1 to ISO/IEC 23090-14. Support for other immersive media codecs, including MPEG-I audio, is expected to follow.

## Conclusions

In this paper, we described the main principles behind the work of the MPEG Systems working group on scene description, including the architecture, key extensions, and formats. MPEG-I scene description builds on Khronos’ glTF2.0 to offer a scene description solution that addresses the needs of a wide variety of dynamic immersive applications. MPEG-I scene description is based on an architectural design that separates media handling from rendering, thus allowing for flexible media access and media representation to be coupled with simple and reliable rendering. With emerging use cases and requirements related to Metaverse experiences and other immersive use cases, the work on MPEG-I scene description provides core technologies in its first edition and relevant extensions in the second edition including AR anchoring, interactivity, lighting, and haptics.

## Resources

### Standards

ISO/IEC 23090-14, Information technology – Coded representation of immersive media – Part 14: Scene description (at stage of publishing available as FDIS text)

### Reference Software

ISO/IEC 23090-24, Information technology – Coded representation of immersive media – Part 24: Conformance and reference software for scene description (at stage of publishing available as WD text)

**Public Information and Repositories**

Asset	Location name
Web Page	<a href="https://mpeg-sd.org">https://mpeg-sd.org</a>
Public comments	<a href="https://github.com/MPEGGroup/Scene-Description">https://github.com/MPEGGroup/Scene-Description</a>
Repository	<a href="https://gitlab.com/mpeg-i/scene-description">https://gitlab.com/mpeg-i/scene-description</a>
Reference SW	<a href="https://gitlab.com/mpeg-i/scene-description/mpegtrimesh">https://gitlab.com/mpeg-i/scene-description/mpegtrimesh</a>
Conformance SW	<a href="https://gitlab.com/mpeg-i/scene-description/conformance">https://gitlab.com/mpeg-i/scene-description/conformance</a>
Scenarios	<a href="https://gitlab.com/mpeg-i/scene-description/scenarios">https://gitlab.com/mpeg-i/scene-description/scenarios</a>
Test vectors	<a href="https://gitlab.com/mpeg-i/scene-description/test-vectors">https://gitlab.com/mpeg-i/scene-description/test-vectors</a>
Test assets	<a href="http://mpegfs.int-evry.fr/mpegcontent/ws-mpegcontent/MPEG-I/Part14-SceneDescriptions">http://mpegfs.int-evry.fr/mpegcontent/ws-mpegcontent/MPEG-I/Part14-SceneDescriptions</a>