|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligence  www.mpai.community |

**Public document**

|  |  |
| --- | --- |
| **N1311** | 2022/07/12 |
| **Source** | Requirements (ARA) and MMC-DC |
| **Title** | Technical Specification – Visual Object and Scene Description (MPAI-OSD) WD0.2 |
| **Target** | MPAI-34 |

|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding  by Artificial Intelligence www.mpai.community |

**MPAI Technical Specification**

**Visual Object and Scene Description**

**MPAI-OSD**

|  |
| --- |
| **WD0.2** |

**WARNING**

Use of the technologies described in this Technical Specification may infringe patents, copyrights or intellectual property rights of MPAI Members or non-members.

MPAI and its Members accept no responsibility whatsoever for damages or liability, direct or consequential, which may result from the use of this Technical Specification.

Readers are invited to review Annex 2 - Notices and Disclaimers.

© Copyright MPAI 2022-2023. All rights reserved.

**Visual Object and Scene Description**

**V1 (under development)**

[1 Introduction (Informative) 3](#_Toc139896287)

[2 Scope of Standard 5](#_Toc139896288)

[3 Terms and Definitions 6](#_Toc139896289)

[4 References 7](#_Toc139896290)

[4.1 Normative References 7](#_Toc139896291)

[4.2 Informative References 7](#_Toc139896292)

[5 Use Case Architectures 7](#_Toc139896293)

[5.1 Conversation About a Scene (CAS) 7](#_Toc139896294)

[5.1.1 Scope of Use Case 7](#_Toc139896295)

[5.1.2 Reference Architecture 7](#_Toc139896296)

[5.2 Human-Connected Autonomous Vehicle (CAV) Interaction (HCI) 8](#_Toc139896297)

[5.2.1 Scope of Use Case 8](#_Toc139896298)

[5.2.2 Reference architecture 9](#_Toc139896299)

[5.3 Avatar-Based Videoconference 10](#_Toc139896300)

[5.3.1 Transmission-side Client Architecture 10](#_Toc139896301)

[5.3.2 Server-side Architecture 11](#_Toc139896302)

[5.3.3 Virtual Secretary Architecture 13](#_Toc139896303)

[5.3.4 Receiving-side Client Architecture 14](#_Toc139896304)

[6 AI Modules 15](#_Toc139896305)

[6.1 Visual Scene Description 15](#_Toc139896306)

[6.1.1 Scope of the AIM 15](#_Toc139896307)

[6.1.2 Reference Architecture 16](#_Toc139896308)

[6.1.3 Input/output data 16](#_Toc139896309)

[6.1.4 AIM Metadata 16](#_Toc139896310)

[6.2 Spatial Object Identification (SOI) 16](#_Toc139896311)

[6.2.1 Scope of the AIM 16](#_Toc139896312)

[6.2.2 Reference Architecture 16](#_Toc139896313)

[6.2.3 Input/output data 17](#_Toc139896314)

[6.2.4 AIM Metadata 17](#_Toc139896315)

[6.3 Data Formats 17](#_Toc139896316)

[6.3.1 Physical Object Identifier 17](#_Toc139896317)

[6.3.2 Spatial Attitude 18](#_Toc139896318)

[6.3.3 Visual Scene Descriptors 18](#_Toc139896319)

[Annex 1 - MPAI-wide terms and definitions 19](#_Toc139896320)

[Annex 2 - Notices and Disclaimers Concerning MPAI Standards (Informative) 22](#_Toc139896321)

[Annex 3 - The Governance of the MPAI Ecosystem (Informative) 24](#_Toc139896322)

[Annex 4 - Patent declarations 26](#_Toc139896323)

# Introduction (Informative)

In recent years, Artificial Intelligence (AI) and related technologies have been introduced in a broad range of applications, have started affecting the life of millions of people and are expected to do so even more in the future. As digital media standards have positively influenced industry and billions of people, so AI-based data coding standards are expected to have a similar positive impact. Indeed, research has shown that data coding with AI-based technologies is generally *more efficient* than with existing technologies for, e.g., compression and feature-based description.

However, some AI technologies may carry inherent risks, e.g., in terms of bias toward some classes of users. Therefore, the need for standardisation is more important and urgent than ever.

The international, unaffiliated, not-for-profit MPAI – Moving Picture, Audio and Data Coding by Artificial Intelligence Standards Developing Organisation has the mission to develop *AI-enabled data coding standards*. MPAI Application Standards enable the development of AI-based products, applications, and services.

As a rule, MPAI standards include four documents: Technical Specification, Reference Software Specifications, Conformance Testing Specifications, and Performance Assessment Specifications.

The last type of Specification includes standard operating procedures to enable users of MPAI Implementations to make informed decision about their applicability based on the notion of Performance, defined as a set of attributes characterising a reliable and trustworthy implementation.

In the following, Terms beginning with a capital letter are defined in *Table 1* if they are specific to this Standard and in Table 13 if they are common to all MPAI Standards.

In general, MPAI Application Standards are defined as aggregations – called AI Workflows (AIW) – of processing elements – called AI Modules (AIM) – executed in an AI Framework (AIF). MPAI defines Interoperability as the ability to replace an AIW or an AIM Implementation with a functionally equivalent Implementation.

MPAI also defines 3 Interoperability Levels of an AIF that executes an AIW. The AIW and its AIMs may have 3 Levels:

*Level 1 –* Implementer-specific and satisfying the MPAI-AIF Standard.

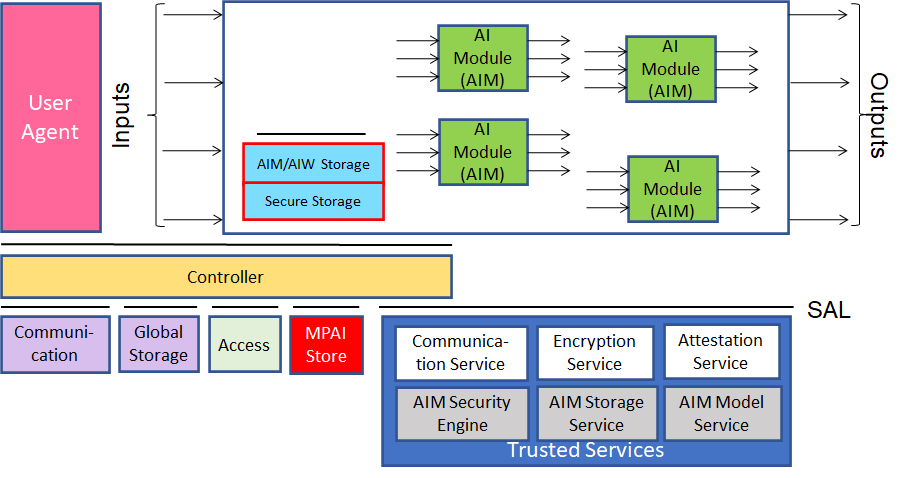
*Level 2 –* Specified by an MPAI Application Standard.

*Level 3 –* Specified by an MPAI Application Standard and certified by a Performance Assessor.

MPAI offers Users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability as the Interoperability Level of an Implementation moves from 1 to 3. Additional information on Interoperability Levels is provided in reference [1].

*Figure 1* depicts the MPAI-AIF Reference Model under which Implementations of MPAI Applic­ation Standards and user-defined MPAI-AIF Conforming applications operate.

MPAI Application Standards normatively specify the Syntax and Semantics of the input and output data and the Function of the AIW and the AIMs, and the Connections between and among the AIMs of an AIW.



*Figure 1 – The AI Framework (AIF) Reference Model and its Components*

In particular, an AIM is defined by its Function and data, but not by its internal architecture, which may be based on AI or data processing, and implemented in software, hardware or hybrid software and hardware technologies.

MPAI Standards are designed to enable a User to obtain, via standard protocols, an Implementation of an AIW and of the set of corresponding AIMs and execute it in an AIF Implementation. The MPAI Store in *Figure 1* is an entity from which Implementations are downloaded. MPAI Standards assume that the AIF, AIW, and AIM Implementations may have been developed by independent implementers. A necessary condition for this to be possible, is that any AIF, AIW, and AIM implementations be uniquely identified. MPAI has appointed an ImplementerID Registration Authority (IIDRA) to assign unique ImplementerIDs (IID) to Implementers.[[1]](#footnote-2)

A necessary condition to make possible the operations described in the paragraph above is the existence of an ecosystem composed of Conformance Testers, Performance Assessors, and an instance of the IIDRA and of the MPAI Store. Reference [1] provides an informative example of such ecosystem.

The chapters and the annexes of this Technical Specification are Normative, unless they are labelled as Informative.

# Scope of Standard

*Visual Object and Scene Description* (MPAI-OSD) is an MPAI Standard targeting the specification of visual object description and their localisation in a space. MPAI-OSD currently comprises:

1. Three Use Cases:
   1. “*Conversation About a Scene*” (CAS) where a human converses with a machine pointing at the objects scattered in a room and displaying Personal Status in their speech, face, and gestures while the machine responds displaying its Personal Status in speech, face, and gesture.
   2. *“Human-Connected Autonomous Vehicle Interaction”* (HCI) where humans stand outside of a Connected Autonomous Vehicle (CAV) and interact with its Human-CAV Interaction Subsystem (HCI) to be let in or sit in a CAV and converse with its HCI including about objects.
   3. “*Avatar-Based Videoconference*” (ABV) where avatars representing humans with their facial emotions and gestures hold a videoconference in a virtual space.
2. Two AIMs:
   1. “*Visual Scene Description*” provides the Body and Face Descriptors of the humans and the Physical Objects in a scene.
   2. “*Spatial Object Identification*” provides the Identifier of a Physical Object in a Physical Environment that a human indicates by pointing with a finger.

Each Use Case or Composite AIM normatively defines:

1. The Functions of the AIW (or Composite AIM) and of the AIMs.
2. The Connections between and among the AIMs
3. The Semantics and the Formats of the input and output data of the AIW (or Composite AIM) and the AIMs.

The word *normatively* implies that an Implementation claiming Conformance to:

1. An *AIW* (or *Composite AIM*), shall:
   1. Perform the AIW function specified in the appropriate Section of Chapter 0 (Chapter 5.3 for Composite AIM).
   2. All AIMs, their topology and connections should conform with the AIW Architecture specified in the appropriate Section of Chapter 0 (Chapter 5.3 for Composite AIM).
   3. The AIW and AIM input and output data should have the formats specified in the appropriate Subsection of Section 6.3.
2. An *AIM*, shall:
   1. Perform the AIM function specified by the appropriate section of Chapter 0 ((Chapter 5.3 for Composite AIM).
   2. Receive as input and produce as output data having the format specified in Section 6.3.

3. A data *Format*, the data shall have the format specified in Section 6.3.

Users of this Technical Specification should note that:

1. This Technical Specification does not mandate any Interoperability Level that is may define.
2. Implementers decide the Interoperability Level their Implementation satisfies.
3. Implementers can use the Reference Software of this Technical Specification to develop their Implementations.
4. The Conformance Testing specification can be used to test the conformity of an Implemen­tation to this Standard.
5. Performance Assessors can assess the level of Performance of an Implementation based on the Performance Assessment specification of this Standard.
6. Implementers and Users should consider the notices and disclaimers of Annex 2.

The current Version of MPAI-OSD has been developed by Requirements (XXX). MPAI expects to produce future MPAI-OSD Versions extending the scope of the Use Cases and/or add new Use Cases within the scope of Visual Object and Scene Description.

# Terms and Definitions

The terms used in this standard whose first letter is capital have the meaning defined in *Table 1*.

*Table 1* *– Table of terms and definitions*

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Descriptor | Coded representation of text, audio, speech, or visual feature. |
| Environment | A Physical or Virtual Space containing a Scene. |
| Identifier | The label uniquely associated with a human or an avatar or an object. |
| Object Descriptor | An individual attribute of the coded representation of an object in a Scene, including its Spatial Attitude. |
| Orientation | The 3 yaw, pitch, and roll (α,β,γ) angles of a representative point of an object in the Real and Virtual Space. |
| Point of View | The Spatial Attitude of a human or avatar looking at an Environment. |
| Position | The 3 coordinates (x,y,z) of a representative point of an object in the Real and Virtual Space. |
| Scene | An Environment populated by objects whether real or virtual. |
| Scene Presentation | The format used by an audio-visual renderer to render the Audio-Visual Scene internal to the machine from a selected Point of View. |
| Spatial Attitude | Position and Orientation and their velocities and accelerations of a Human and Physical Object in a Real or Virtual Environment. |
| Visual Object | Coded representation of Visual information with its metadata. A Video Object can be a combination of Video Objects. |

# References

## Normative References

This standard normatively references the following documents, both from MPAI and other stan­dards organisations. MPAI standards are publicly available at https://mpai.community/standards/resources/.

1. MPAI; Technical Specification; The governance of the MPAI ecosystem (MPAI-GME), V1.1; https://mpai.community/standards/mpai-gme/
2. MPAI; Technical Specification; MPAI Ecosystem Governance (MPAI-GME) V1.1; https://mpai.community/standards/mpai-gme/
3. MPAI; Technical Specification; AI Framework (MPAI-AIF) V1.1; https://mpai.community/standards/mpai-aif/
4. MPAI; Technical Specification; Multimodal Conversation (MPAI-MMC) V2; https://mpai.community/standards/mpai-aif/
5. MPAI; The MPAI Statutes; https://mpai.community/statutes/
6. MPAI; The MPAI Patent Policy; https://mpai.community/about/the-mpai-patent-policy/.

## Informative References

The references provided here are for information purpose.

1. Technical Specification: Connected Autonomous Vehicles (MPAI-CAV) – Architecture V1; https://mpai.community/standards/mpai-cav/

# Use Case Architectures

## Conversation About a Scene (CAS)

### Scope of Use Case

This Use Case addresses a human holding a conversation with a machine:

1. The machine perceives (sees and hears) an Environment containing a speaking human and some scattered objects.
2. The machine recognises the human’s Speech and obtains the human’s Personal Status by capturing Speech, Face, and Gesture.
3. The human converses with the machine indicating the object in the Environment s/he wishes to talk to or ask questions about using Speech, Face, and Gesture.
4. The machine understands which object the human is referring to and generates an avatar that:
   1. Utters Speech conveying a synthetic Personal Status that is relevant to the human’s Personal Status as shown by his/her Speech, Face, and Gesture, and
   2. Displays a face conveying a Personal Status that is relevant to the human’s Personal Status and to the response the machine intends to make.
5. The machine displays the Scene Presentation corresponding to how it perceives the Environment. The objects in the scene are labelled with the machine’s understanding of their semantics so that the human can understand how the machine sees the Environment.

### Reference Architecture

The Machine operates according to the following workflow:

1. Visual Scene Description produces Body Object and Physical Objects from Input Video.
2. Speech Recognition produces Recognised Text from Input Speech.
3. Spatial Object Identification produces Object ID from Physical Object and Body Object.
4. Personal Status Extraction produces Personal Status.
5. Language Understanding produces Meaning and Text (Language Understanding) from Personal Status, Recognised Text, and Physical Object ID.
6. Dialogue Processing produces Output Text and Output Personal Status from Personal Status, Meaning, and Text (Language Understanding).
7. Personal Status Display produces Machine Text, Machine Speech, Machine Avatar from Output Text, and Output Personal Status. Machine Speech and Machine Avatar are in formats that can be immediately rendered.
8. Scene Presentation uses the Visual Scene Descriptors to produce the Rendered Scene as seen from the user-selected Point of View. The rendering is constantly updated as the machine improves its understanding of the scene and its objects.

Note that the version of the Visual Scene Parsing AIM employed in this Use Case is designed to provide the Body Object, the Face Object and the Visual Scene Geometry.

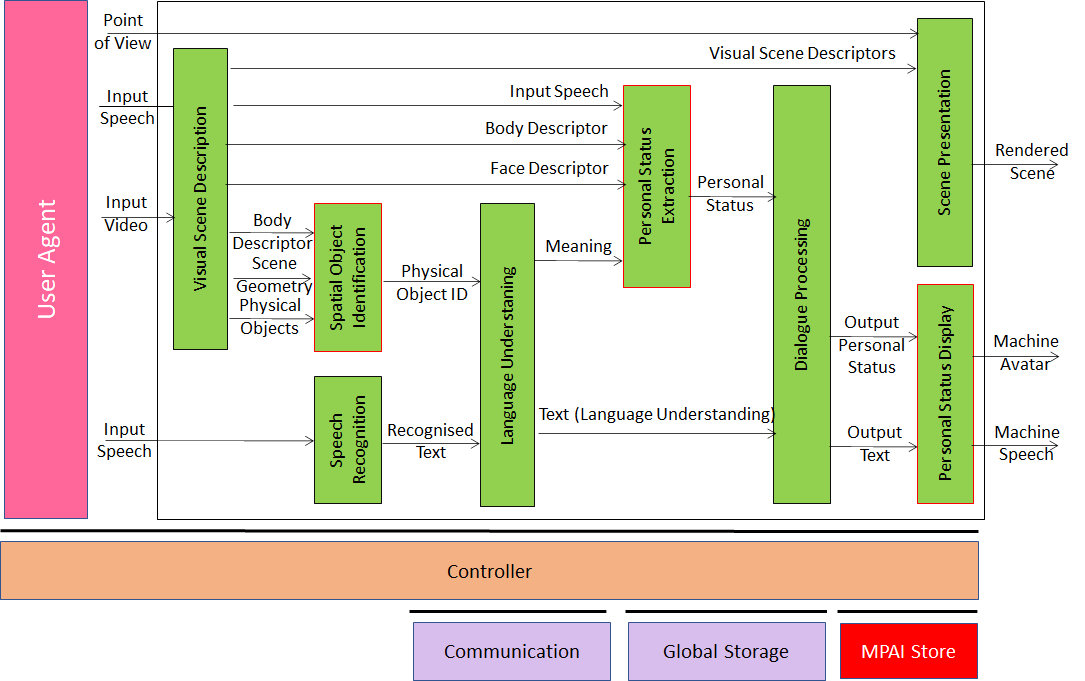


Figure 2 – Reference Model of Conversation About a Scene

## Human-Connected Autonomous Vehicle (CAV) Interaction (HCI)

### Scope of Use Case

This use case is part of the Connected Autonomous Vehicle (CAV) project [7]. A CAV is a system able to execute a command to move itself based on 1) analysis and interpretation of the data sensed by a range of onboard sensors exploring the environment and 2) information transmitted by other sources in range, such as other CAVs, traffic lights and roadside units.

Figure 3 depicts the four CAV subsystems.

Diagram

Description automatically generated

Figure 3 – The Connected Autonomous Vehicle Reference Model

**Human-CAV interaction (HCI)** recognises the human owner or renter, responds to humans’ commands and queries, converses with humans during the travel and may activate other Subsystems in response to humans’ requests. The data exchanged between the HCI, and the Autonomous Motion Subsystem (AMS) is depicted in Figure 4 but the requirements of the format of the data exchanged between HCI and AMS are not part of this document.

**Environment Sensing Subsystem (ESS)** acquires information from the Environment via a variety of sensors and produces a representation of the Environment (Basic Environment Representation), i.e., its best understanding of the Environment based on the sensed data.

**Autonomous Motion Subsystem (AMS)** computes the route to destination, uses different sources of information – CAV sensors, other CAVs and transmitting units – to produce a Full Environment Representation and issues commands to drive the CAV to the intended destination.

**Motion Actuation Subsystem (MAS)** provides non-electromagnetic and non-acoustical environment information¸ and receives and actuates motion commands in the physical world.

The CAV in Human-CAV Interaction is represented by an avatar with the following perceptible features:

1. Visual: head, face, and shoulders.
2. Audio: speech.

Both visual and audio features convey to the extent possible the Personal Status that would be displayed by a human driver in similar conditions.

### Reference architecture

Figure 4 represents the Human-CAV Interaction (HCI) Reference Model.

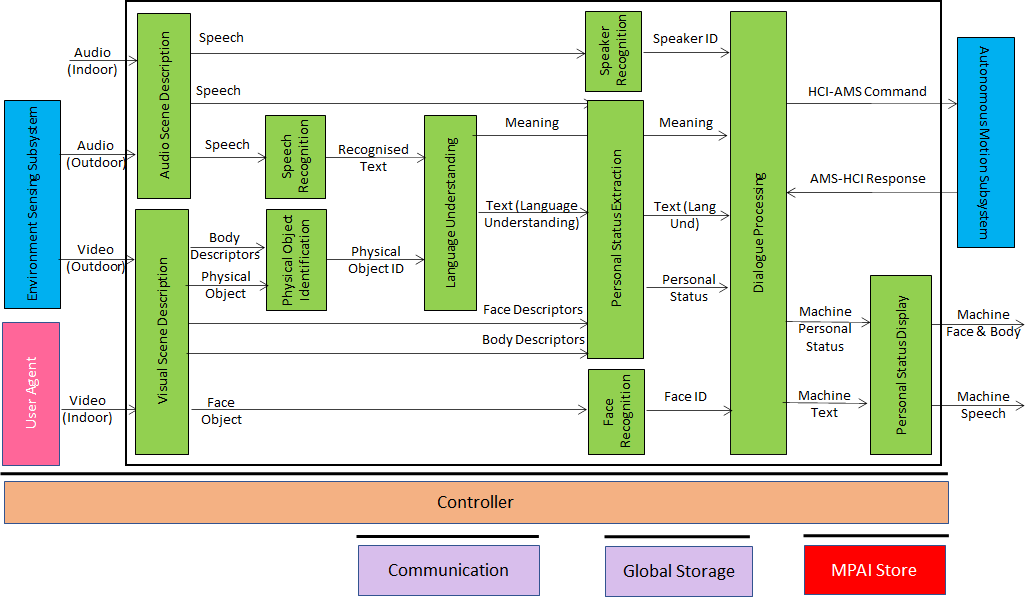


Figure 4 – Human-CAV Interaction Reference Model

## Avatar-Based Videoconference

### Transmission-side Client Architecture

*Figure 5* gives the normative architecture of the Client (Transmission side) AIW. Red text refers to data sent at meeting start.

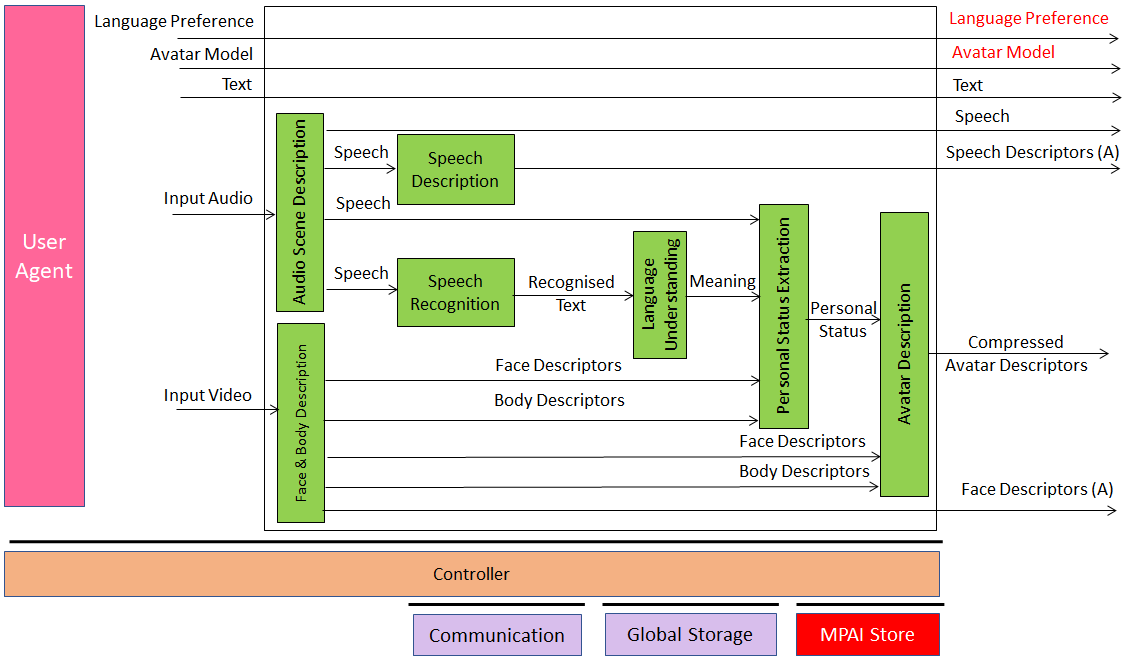
*Figure 5 – Reference Model of Avatar Videoconference Client (Transmission side)*

Table 2 gives the input and output data of the Client (Transmission side) AIW:

Table 2 - Input and output data of Client (Transmission side) AIW

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Text | Chat text used to communicate with Virtual Secretary or other participants |
| Language Preference | The language participant wishes to speak and hear at the videoconference. |
| Input Audio | Audio of participant’s Speech and Environment Audio. |
| Input Video | Video of participants’ upper part of the body. |
| Avatar Model | The avatar model selected by the participant. |
| **Output** | **Comments** |
| Language Preference | As in input. |
| Participant’s Speech | Speech as separated from Environment Audio. |
| Compressed Avatar Descriptors | Compressed Descriptors produced by Transmitting Client. |

Table 3gives the AI Modules of Client (Transmission side) AIW.

Table 3 - AI Modules of Client (Transmission side) AIW

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input** | **Output** |
| **Audio Scene Description** | Input Audio | Audio Scene Descriptors |
| **Visual Scene Description** | Input Video | Visual Scene Descriptors |
| **Audio Scene Parsing** | Audio Scene Descriptors | Audio Scene Geometry  Speech Objects |
| **Speech Recognition** | Speech Objects | Recognised Text |
| **Language Understanding** | Recognised Text | Language (Understanding Text)  Meaning |
| **Personal Status Extraction** | Recognised Text  Speech  Face Object  Human Object | Personal Status |
| **Avatar Description** | Meaning  Personal Status  Face Description  Gesture Description | Compressed Avatar Descriptors |

### Server-side Architecture

Figure 5 gives the architecture of Server AIW. Red text refers to data sent at meeting start.

A diagram of a model

Description automatically generated

Figure 6 – Reference Model of Avatar-Based Videoconference Server

Table 4 gives the input and output data of Server AIW.

Table 4 - Input and output data of Server AIW

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Participant Identities (xN) | Assigned by Conference Manager |
| Spatial Attitudes | (Initial) Assigned by Conference Manager |
| Speech Descriptors (A) (xN) | Participant’s Speech Descriptors for Authentication |
| Face Descriptors (A) (xN) | Participant’s Face Descriptors for Authentication |
| Selected Languages (xN) | From all participants |
| Speech (xN+1) | From all participants and Virtual Secretary |
| Text (xN+1) | From all participants and Virtual Secretary |
| Avatar Model (xN+1) | From all participants and Virtual Secretary |
| Compressed Avatar Descriptors (xN+1) | From all participants and Virtual Secretary |
| Summary | From Virtual Secretary |
| **Outputs** | **Comments** |
| Environment Model | From Server Manager |
| Avatar Model (xN+1) | From all participants and Virtual Secretary |
| Compressed Avatar Descriptors (xN+1) | Participants + Virtual Secretary Compressed Avatar D. |
| Participant ID (xN+1) | Participants + Virtual Secretary IDs |
| Speech (xN+1) | Participants + Virtual Secretary Speech |
| Text (xN+1) | Participants + Virtual Secretary Speech |

Table 5 gives the AI Modules of Server AIW.

Table 5 - AI Modules of Server AIW

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input** | **Output** |
| **Participant Authentication** | Speech Descriptors  Face Descriptors | Participant Authentication |
| **Text and Speech Translation** | Language Preferences  Text  Speech | Translates Text  Translated Speech |

### Virtual Secretary Architecture

Figure 7 depicts the architecture of the Virtual Secretary AIW. Red text refers to data sent only once at meeting start.

A screenshot of a computer

Description automatically generated

Figure 7 – Reference Model of Virtual Secretary

Table 6 gives the input and output data of Virtual Secretary Composite AIM.

Table 6 - I/O data of Virtual Secretary

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Text (xN) | Server | Remarks on the summary, etc. |
| Speech (xN) | Server | Utterances by avatars |
| Descriptor Decompression | Server | Separate for Face and Gesture |
| **Output data** | **To** | **Comments** |
| Summary | Avatars | Summary of avatars’ interventions |
| VS Speech | Avatars | Speech to avatars |
| VS Text | Avatars | Response to chat. |
| VS Compressed Avatar Descriptors | Avatars | Face to avatars |

Table 7 gives the AI Modules of the Virtual Secretary depicted in Figure 7.

Table 7 - AI Modules of Virtual Secretary

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input** | **Output** |
| **Speech Recognition** | Speech | Recognised Text |
| **Descriptor Decompression** | Compressed Avatar Descriptors | Face Descriptors  Gesture Descriptors |
| **Avatar Synthesis** | Compressed Avatar Descriptors | Body Descriptors  Face Descriptors |
| **Personal Status Extraction** | Speech  Body Descriptors  Face Descriptors. | Receives.  Personal Status |
| **Language understanding** | Recognised Text  Personal Status | Text (Language Understanding)  Meaning |
| **Summarisation** | Text (Language Understanding)  Meaning  Edited Summary | Summary |
| **Question & Dialogue Processing** | Text (Language Understanding)  Personal Status  Meaning | PS (Speech)  PS (Face)  PS (Gesture)  Output Text  Edited Summary. |
| **Personal Status Display** | PS (Speech)  PS (Face)  PS (Gesture)  Output Text | VS Text  VS Speech  VS Compressed Avatar Descriptors |

### Receiving-side Client Architecture

Figure 8 depicts the architecture of the Client (Receiving Side) AIW. Red text refers to data sent only once at meeting start.

A diagram of a video game

Description automatically generated

Figure 8 – Reference Model of Avatar-Based Videoconference Client (Receiving Side)

*Table 8* gives the input and output data of Client (Receiving Side) AIW.

*Table 8 – Input and output data of Client (Receiving Side) AIW*

|  |  |
| --- | --- |
| **Input** | **Comments** |
| Point of View | Participant-selected point to see visual objects and hear audio objects in the Virtual Environment. |
| Spatial Attitudes (xN+1) | Positions and Orientations of Avatars in the Environment |
| Participant IDs (xN) | Unique Participants’ IDs |
| Speech (xN+1) | Participant’s Speech (e.g., translated). |
| Environment Model | Environment Model. |
| Compressed Avatar Descriptors (xN+1) | Descriptors of Avatar’s body animation. |
| **Output** | **Comments** |
| Output Audio | Presented using loudspeaker (array)/earphones. |
| Output Visual | Presented using 2D or 3D display. |

Table 9 gives the AI Modules of Receiving Client AIW.

Table 9 - AI Modules of Client (Receiving Side)

|  |  |  |
| --- | --- | --- |
| **AIM** | **Input** | **Output** |
| **Audio Scene Creation** | Environment Model  Input Audio (xN)  VS Speech | Audio Scene |
| **Visual Scene Creation** | Environment Model  Compressed Avatar Descriptors (xN+1)  Spatial Attitudes (xN+1) | Visual Scene |
| **AV Scene Viewer** | Audio Scene  Visual Scene  Point of View | Output Audio  Output Video |

# AI Modules

It is useful to define combinations of AI Modules called Composite AI Modules for use in several MPAI-MMC Use Cases. This chapter specifies such Composite AI Modules using a format like the one adopted for Uses Cases. The specification of such Composite AI Modules is provided by [3].

## Visual Scene Description

### Scope of the AIM

The scope of the Visual Scene Description AIM is to

1. Acquire a visual scene.
2. Provide the following output:
   1. The Face Descriptors of a human in the scene.
   2. The Body Descriptors of a human in the scene.
   3. The Physical Objects in the scene.
   4. The Scene Geometry.

### Reference Architecture

Figure 9 depicts the AIM implementing the Visual Scene Description AIM.

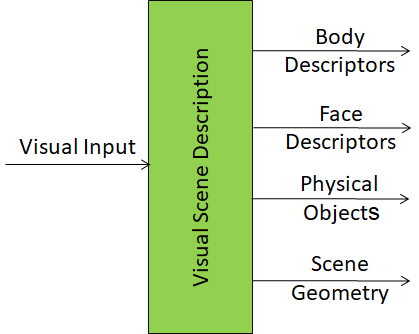


Figure 9 - Reference Model of the Visual Scene Description AIM

### Input/output data

*Table 10*gives the input/output data of Spatial Object Identification.

*Table 10 – I/O data of Spatial Object Identification*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Visual Scene | Another AIM or a Device |  |
| **Output data** | **To** | **Comments** |
| Body Descriptors | Downstream AIM | Interprets Body Descriptors |
| Face Descriptors | Downstream AIM | Interprets Face Descriptors |
| Physical Objects | Downstream AIM | Identifies Object |
| Scene Geometry | Downstream AIM | Used to localise human and objects |

### AIM Metadata

Specified in Annex Section 1.

## Spatial Object Identification (SOI)

### Scope of the AIM

The purpose of the Spatial Object Identification AIM is to provide the Identifier of a Physical Object in an Environment. The human indicates by pointing at the intended object with a finger.

### Reference Architecture

Figure 10 depicts the AIM implementing the Spatial Object Identification AIM.

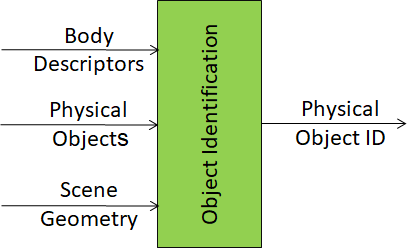


Figure 10 – Reference Model of the Spatial Object Identification AIM

### Input/output data

*Table 11* gives the input/output data of Spatial Object Identification.

*Table 11 – I/O data of Spatial Object Identification*

|  |  |  |
| --- | --- | --- |
| **Input data** | **From** | **Comment** |
| Body Descriptors | Visual Scene Description | There is a human pointing to an object |
| Physical Objects | Visual Scene Description | There are many scene objects |
| Scene Geometry | Visual Scene Description | Full description of the scene |
| **Output data** | **To** | **Comments** |
| Physical Object ID | Another AIM | Human points to one object only |

### AIM Metadata

Specified in Annex Section 1.

## Data Formats

The Data Formats specified in this Technical Specification are listed in *Table 12*.

The first column gives the name of the data Format, the second the subsection where the data Format is specified and the third the Use Case(s) making use of it.

*Table 12 – Data formats*

|  |  |  |
| --- | --- | --- |
| **Name of Data Format** | **Subsection** | **Use Case** |
| Physical Object Identifier | 6.3.1 | CAS |
| Spatial Attitude | 6.3.2 | CAS |
| HCI |
| Visual Scene Descriptors | 6.3.3 | ABV |
| CAS |
| HCI |

### Physical Object Identifier

A Physical Object is identified as follows.

#### Syntax

{

"$schema":"http://json-schema.org/draft-07/schema",

"definitions":{

"objectIdentifier":{

"type":"object",

"properties":{

"objectImageLabel":{

"type":"string"

},

"confidenceLevel":{

"type":"integer"

}

}

}

},

"type":"object",

"properties":{

"primary":{

"$ref":"#/definitions/ObjectIdentifierType"

},

"secondary":{

"$ref":"#/definitions/ObjectIdentifierType"

}

}

}

#### Semantics

| *Name* | *Definition* |
| --- | --- |
| objectIdentifier | Provides the description of the recognised object. |
| objectImageLabel | Indicates the recognised object’s label in the object image. |
| confidenceLevel | Indicates the confidence level of the object image label recognised by the “Video analysis”. |

### Spatial Attitude

Spatial Attitude is a real vector containing at least Position in a Cartesian Coordinate System. It may also contain Orientation and their Velocities and Accelerations, in the following order:

|  |  |  |
| --- | --- | --- |
| PosX | PosY | PosZ |
| OrientX | OrientY | OrientZ |
| PosXVel | PosYVel | PosZVel |
| OrientXVel | OrientYVel | OrientZVel |
| PosXAcc | PosYAcc | PosZAcc |
| OrientXAcc | OrientYAcc | OrientZAcc |

The vector

PosX, PosY, PosZ

OrientX, OrientY, OrientZ

Is also called Spatial Attributes.

### Visual Scene Descriptors

Visual Scene Descriptors are contained in a real vector with the following structure:

|  |
| --- |
| **Variable name** |
| Timestamp type |
| Timestamp value |
| Space type |
| Space value |
| Number of humans |
| Spatial Attitudes |
| Number of Objects |
| Spatial Attitudes |
| Human1 Body Descriptors |
| Human1 Head Descriptors |
| Human1 Face Descriptors |
| Human2 Body Descriptors |
| Human2 Head Descriptors |
| Human2 Face Descriptors |
| … |
| Physical Object1 |
| Physical Object2 |
| … |

The Timestamp type can have the following two values:

0 – Absolute Time

1 – Relative Time

The Space type can have the following two values:

0 – Absolute space (0,0,0) point is point with latitude = 0, longitude = 0, and altitude = 0)

1 – Relative Space (0,0,0) point is point with assigned (x.y.z) values.

1. MPAI-wide terms and definitions

The Terms used in this standard whose first letter is capital and are not already included in *Table 1* are defined in **Error! Reference source not found.***.*

Table 13 - MPAI-wide Terms

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Access | Static or slowly changing data that are required by an application such as domain knowledge data, data models, etc. |
| AI Framework (AIF) | The environment where AIWs are executed. |
| AI AIMName (AIM) | A data processing element receiving AIM-specific Inputs and producing AIM-specific Outputs according to according to its Function. An AIM may be an aggregation of AIMs. |
| AI Workflow (AIW) | A structured aggregation of AIMs implementing a Use Case receiving AIW-specific inputs and producing AIW-specific outputs according to the AIW Function. |
| Application Standard | An MPAI Standard designed to enable a particular application domain. |
| Channel | A connection between an output port of an AIM and an input port of an AIM. The term “connection” is also used as synonymous. |
| Communication | The infrastructure that implements message passing between AIMs |
| Composite AIM | An AIM aggregating more than one AIM. |
| Component | One of the 7 AIF elements: Access, Communication, Controller, Internal Storage, Global Storage, Store, and User Agent |
| Conformance | The attribute of an Implementation of being a correct technical Implem­entation of a Technical Specification. |
| Conformance Tester | An entity Testing the Conformance of an Implem­entation. |
| Conformance Testing | The normative document specifying the Means to Test the Conformance of an Implem­entation. |
| Conformance Testing Means | Procedures, tools, data sets and/or data set characteristics to Test the Conformance of an Implem­en­tation. |
| Connection | A channel connecting an output port of an AIM and an input port of an AIM. |
| Controller | A Component that manages and controls the AIMs in the AIF, so that they execute in the correct order and at the time when they are needed |
| Data Format | The standard digital representation of data. |
| Data Semantics | The meaning of data. |
| Ecosystem | The ensemble of actors making it possible for a User to execute an application composed of an AIF, one or more AIWs, each with one or more AIMs potentially sourced from independent implementers. |
| Explainability | The ability to trace the output of an Implementation back to the inputs that have produced it. |
| Fairness | The attribute of an Implementation whose extent of applicability can be assessed by making the training set and/or network open to testing for bias and unanticipated results. |
| Function | The operations effected by an AIW or an AIM on input data. |
| Global Storage | A Component to store data shared by AIMs. |
| Internal Storage | A Component to store data of the individual AIMs. |
| Identifier | A name that uniquely identifies an Implementation. |
| Implementation | 1. An embodiment of the MPAI-AIF Technical Specification, or 2. An AIW or AIM of a particular Level (1-2-3) conforming with a Use Case of an MPAI Applic­ation Standard. |
| Implementer | A legal entity implementing MPAI Technical Specifications. |
| ImplementerID (IID) | A unique name assigned by the ImplementerID Registration Authority to an Implementer. |
| ImplementerID Registration Authority (IIDRA) | The entity appointed by MPAI to assign ImplementerID’s to Implementers. |
| Interoperability | The ability to functionally replace an AIM with another AIW having the same Interoperability Level |
| Interoperability Level | The attribute of an AIW and its AIMs to be executable in an AIF Implem­en­tati­on and to:   1. Be proprietary (Level 1) 2. Pass the Conformance Tes­ting (Level 2) of an Applic­ation Standard 3. Pass the Perform­ance Testing (Level 3) of an Applic­ation Standard. |
| Knowledge Base | Structured and/or unstructured information made accessible to AIMs via MPAI-specified interfaces |
| Message | A sequence of Records transported by Communication through Channels. |
| Normativity | The set of attributes of a technology or a set of technologies specified by the applicable parts of an MPAI standard. |
| Performance | The attribute of an Implementation of being Reliable, Robust, Fair and Replicable. |
| Performance Assessment | The normative document specifying the Means to Assess the Grade of Performance of an Implementation. |
| Performance Assessment Means | Procedures, tools, data sets and/or data set characteristics to Assess the Performance of an Implem­en­tation. |
| Performance Assessor | An entity Assessing the Performance of an Implementation. |
| Profile | A particular subset of the technologies used in MPAI-AIF or an AIW of an Application Standard and, where applicable, the classes, other subsets, options and parameters relevant to that subset. |
| Record | A data structure with a specified structure |
| Reference Model | The AIMs and theirs Connections in an AIW. |
| Reference Software | A technically correct software implementation of a Technical Specific­ation containing source code, or source and compiled code. |
| Reliability | The attribute of an Implementation that performs as specified by the Application Standard, profile and version the Implementation refers to, e.g., within the application scope, stated limitations, and for the period of time specified by the Implementer. |
| Replicability | The attribute of an Implementation whose Performance, as Assessed by a Performance Assessor, can be replicated, within an agreed level, by another Performance Assessor. |
| Robustness | The attribute of an Implementation that copes with data outside of the stated application scope with an estimated degree of confidence. |
| Scope | The domain of applicability of an MPAI Application Standard |
| Service Provider | An entrepreneur who offers an Implementation as a service (e.g., a recommendation service) to Users. |
| Standard | The ensemble of Technical Specification, Reference Software, Confor­man­ce Testing and Performance Assessment of an MPAI application Standard. |
| Technical Specification | (Framework) the normative specification of the AIF.  (Application) the normative specification of the set of AIWs belon­ging to an application domain along with the AIMs required to Im­plem­ent the AIWs that includes:   1. The formats of the Input/Output data of the AIWs implementing the AIWs. 2. The Connections of the AIMs of the AIW. 3. The formats of the Input/Output data of the AIMs belonging to the AIW. |
| Testing Laboratory | A laboratory accredited to Assess the Grade of Performance of Implementations. |
| Time Base | The protocol specifying how Components can access timing information |
| Topology | The set of AIM Connections of an AIW. |
| Use Case | A particular instance of the Application domain target of an Application Standard. |
| User | A user of an Implementation. |
| User Agent | The Component interfacing the user with an AIF through the Controller |
| Version | A revision or extension of a Standard or of one of its elements. |

1. Notices and Disclaimers Concerning MPAI Standards (Informative)

The notices and legal disclaimers given below shall be borne in mind when [downloading](https://www.mpai.community/resources/) and using approved MPAI Standards.

In the following, “Standard” means the collection of four MPAI-approved and [published](https://www.mpai.community/resources/) docum­ents: “Technical Specification”, “Reference Software” and “Conformance Testing” and, where applicable, “Performance Testing”.

Life cycle of MPAI Standards

MPAI Standards are developed in accordance with the [MPAI Statutes](https://mpai.community/statutes/). An MPAI Standard may only be developed when a Framework Licence has been adopted. MPAI Standards are developed by especially established MPAI Development Committees who operate on the basis of consensus, as specified in Annex 1 of the [MPAI Statutes](https://mpai.community/statutes/). While the MPAI General Assembly and the Board of Directors administer the process of the said Annex 1, MPAI does not independently evaluate, test, or verify the accuracy of any of the information or the suitability of any of the technology choices made in its Standards.

MPAI Standards may be modified at any time by corrigenda or new editions. A new edition, however, may not necessarily replace an existing MPAI standard. Visit the [web page](https://mpai.community/resources/) to determine the status of any given published MPAI Standard.

Comments on MPAI Standards are welcome from any interested parties, whether MPAI members or not. Comments shall mandatorily include the name and the version of the MPAI Standard and, if applicable, the specific page or line the comment applies to. Comments should be sent to the [MPAI Secretariat](mailto:secretariat@mpai.community). Comments will be reviewed by the appropriate committee for their technical relevance. However, MPAI does not provide interpretation, consulting information, or advice on MPAI Standards. Interested parties are invited to join MPAI so that they can attend the relevant Development Committees.

Coverage and Applicability of MPAI Standards

MPAI makes no warranties or representations of any kind concerning its Standards, and expressly disclaims all warranties, expressed or implied, concerning any of its Standards, including but not limited to the warranties of merchantability, fitness for a particular purpose, non-infringement etc. MPAI Standards are supplied “AS IS”.

The existence of an MPAI Standard does not imply that there are no other ways to produce and distribute products and services in the scope of the Standard. Technical progress may render the technologies included in the MPAI Standard obsolete by the time the Standard is used, especially in a field as dynamic as AI. Therefore, those looking for standards in the Data Compression by Artificial Intelligence area should carefully assess the suitability of MPAI Standards for their needs.

IN NO EVENT SHALL MPAI BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

MPAI alerts users that practicing its Standards may infringe patents and other rights of third parties. Submitters of technologies to this standard have agreed to licence their Intellectual Property according to their respective Framework Licences.

Users of MPAI Standards should consider all applicable laws and regulations when using an MPAI Standard. The validity of Conformance Testing is strictly technical and refers to the correct implementation of the MPAI Standard. Moreover, positive Performance Assessment of an implementation applies exclusively in the context of the [MPAI Governance](https://mpai.community/standards/mpai-gme/) and does not imply compliance with any regulatory requirements in the context of any jurisdiction. Therefore, it is the responsibility of the MPAI Standard implementer to observe or refer to the applicable regulatory requirements. By publishing an MPAI Standard, MPAI does not intend to promote actions that are not in compliance with applicable laws, and the Standard shall not be construed as doing so. In particular, users should evaluate MPAI Standards from the viewpoint of data privacy and data ownership in the context of their jurisdictions.

Implementers and users of MPAI Standards documents are responsible for determining and complying with all appropriate safety, security, environmental and health and all applicable laws and regulations.

Copyright

MPAI draft and approved standards, whether they are in the form of documents or as web pages or otherwise, are copyrighted by MPAI under Swiss and international copyright laws. MPAI Standards are made available and may be used for a wide variety of public and private uses, e.g., implementation, use and reference, in laws and regulations and standardisation. By making these documents available for these and other uses, however, MPAI does not waive any rights in copyright to its Standards. For inquiries regarding the copyright of MPAI standards, please contact the [MPAI Secretariat](mailto:secretariat@mpai.community).

The Reference Software of an MPAI Standard is released with the [MPAI Modified Berkeley Software Distribution licence](https://mpai.community/about/licence/). However, implementers should be aware that the Reference Software of an MPAI Standard may reference some third party software that may have a different licence.

1. The Governance of the MPAI Ecosystem (Informative)

**Level 1 Interoperability**

With reference to *Figure 1* MPAI issues and maintains a standard – called MPAI-AIF – whose components are:

1. An environment called AI Framework (AIF) running AI Workflows (AIW) composed of inter­connected AI Modules (AIM) exposing standard interfaces.
2. A distribution system of AIW and AIM Implementation called MPAI Store from which an AIF Implementation can download AIWs and AIMs.

A Level 1 Implementation shall be an Implementation of the MPAI-AIF Technical Specification executing AIWs composed of AIMs able to call the MPAI-AIF APIs.

|  |  |
| --- | --- |
| Implementers’ benefits | Upload to the MPAI Store and have globally distributed Implementations of   * AIFs conforming to MPAI-AIF. * AIWs and AIMs performing prop­rietary functions executable in AIF. |
| Users’ benefits | Rely on Implementations that have been tested for security. |
| MPAI Store’s role | * Tests the Conformance of Implementations to MPAI-AIF. * Verifies Implementations’ security, e.g., absence of malware. * Indicates unambiguously that Implementations are Level 1. |

**Level 2 Interoperability**

In a Level 2 Implem­entation, the AIW shall be an Implementation of an MPAI Use Case, and the AIMs shall con­form with an MPAI Applicati­on Standard.

|  |  |
| --- | --- |
| Implementers’ benefits | Upload to the MPAI Store and have globally distributed Implementations of   * AIFs conforming to MPAI-AIF. * AIWs and AIMs conforming to MPAI Application Standards. |
| Users’ benefits | * Rely on Implementations of AIWs and AIMs whose Functions have been reviewed during standardisation. * Have a degree of Explainability of the AIW operation because the AIM Func­tions and the data Formats are known. |
| Market’s benefits | * Open AIW and AIM markets foster competition leading to better products. * Competition of AIW and AIM Implementations fosters AI innovation. |
| MPAI Store’s role | * Tests Conformance of Implementations with the relevant MPAI Standard. * Verifies Implementations’ security. * Indicates unambiguously that Implementations are Level 2. |

**Level 3 Interoperability**

MPAI does not generally set standards on how and with what data an AIM should be trained. This is an important differentiator that promotes competition leading to better solutions. However, the performance of an AIM is typically higher if the data used for training are in greater quantity and more in tune with the scope. Training data that have large variety and cover the spec­trum of all cases of interest in breadth and depth typically lead to Implementations of higher “quality”.

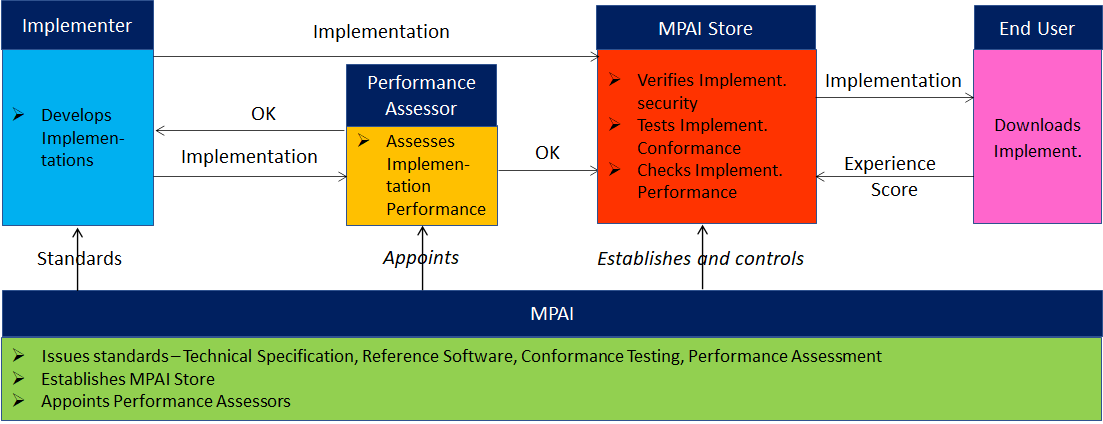
For Level 3, MPAI normatively specifies the process, the tools and the data or the characteristics of the data to be used to Assess the Grade of Performance of an AIM or an AIW.

|  |  |
| --- | --- |
| Implementers’ benefits | May claim their Implementations have passed Performance Assessment. |
| Users’ benefits | Get assurance that the Implementation being used performs correctly, e.g., it has been properly trained. |
| Market’s benefits | Implementations’ Performance Grades stimulate the development of more Performing AIM and AIW Implementations. |
| MPAI Store’s role | * Verifies the Implementations’ security * Indicates unambiguously that Implementations are Level 3. |

**The MPAI ecosystem**

The following *Figure 11* is a high-level description of the MPAI ecosystem operation applicable to fully conforming MPAI implementations:

1. MPAI establishes and controls the not-for-profit MPAI Store (step 1).
2. MPAI appoints Performance Assessors (step 2).
3. MPAI publishes Standards (step 3).
4. Implementers submit Implementations to Performance Assessors (step 4).
5. If the Implementation Performance is acceptable, Performance Assessors inform Implementers (step 5a) and MPAI Store (step 5b).
6. Implementers submit Implementations to the MPAI Store (step 6); The Store Tests Confor­mance and security of the Implementation.
7. Users download Implementations (step 7).



*Figure 11 – The MPAI ecosystem operation*

1. Patent declarations

The MPAI Multimodal Conversation (MPAI-MMC) Technical Specification has been developed according to the process outlined in the MPAI Statutes [4] and the MPAI Patent Policy [6].

The following entities have agreed to licence their standard essential patents reading on the MPAI Multimodal Conversation (MPAI-MMC) Technical Specification according to the MPAI-AIF Framework Licence [6]:

|  |  |  |
| --- | --- | --- |
| **Entity** | **Name** | **Email address** |
|  |  |  |
|  |  |  |
|  |  |  |

1. AIW and AIM Metadata of MMC-CAS

# AIW metadata for CAS

# AIW metadata for CAS

1. AIW and AIM Metadata of CAV-HCI

# AIW metadata for HCI

# AIW metadata for HCI

1. AIW and AIM Metadata of MMC-ABV

# AIW metadata for Transmitting Client

# AIM metadata for Transmitting Client

# AIW metadata for Transmitting Client

# AIM metadata for Transmitting Client

# AIW metadata for Server

# AIM metadata for Server

# AIW metadata for Virtual Assistant

# AIM metadata for Virtual Assistant

# AIW metadata for Receiving Client

# AIM metadata for Receiving Client

1. AIM Metadata of Visual Scene Description

# AIW metadata for VSD

1. AIM Metadata of Physical Object Identification

# AIW metadata

1. At the time of publication of this standard, the MPAI Store was assigned as the IIDRA. [↑](#footnote-ref-2)