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

|  |  |
| --- | --- |
| N73 | 2020/11/18 |
| Source | Miran Choi |
| Title | MPAI-MMC Functional Requirements work programme |
| Target | MPAI Members |

# Introduction

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an [international association](http://mpai.community/) with the mission to develop *AI-enabled data coding standards*. Artificial Intelligence (AI) technologies have shown they can offer *more efficient data coding* than existing technol­ogies.

MPAI has analysed six use cases covering applic­ation areas benefiting from AI technol­ogies. Even though use cases are disparate, each of them can be implemented with a combination of processing modules performing functions that concur to achieving the inten­ded result.

MPAI has assessed that, leaving it to the market to develop individual implementations, would multiply costs and delay adoption of AI technologies, while moduleswith standard interfaces, combined and executed within the MPAI-specified AI Framework, will favour the emergence of horizontal markets where *proprietary and competing* module implemen­tations exposing *standard interfaces* will reduce cost, promote adoption and incite progress of AI technologies. MPAI calls these modules *AI Modules (AIM)*.

This paper describes the current plans to develop the MPAI “MultiModal Conversation” standard (MPAI-MMC) to enable human-machine conversation that emulates human-human conversation in completeness and intensity using AI.

Chapter 2 introduces the MPAI-MMC features. Chapter 3 provides summary information on the advanced IT environment that will execute MPAI-MMC applications. Chapter 4 identifies the items that will likely be the object of the MPAI-MMC standard.

# MPAI-MMC features

Owing to the recent advancement of AI technologies, natural language processing started to be widely used in various applications. One useful application is the conversational partner which provides the user with information, entertains, chats and answers questions through the speech interface. For the application to provide a better service to the user, more than just a speech inter­face should be included. For example, emotion recognizer and gesture interpreter are needed for improved multi-modal interfaces.

MPAI Multi-modal conversation (MPAI-MMC) aims to enable human-machine conversation that emulates human-human conversation in completeness and intensity by using AI.

The following list gives MMC examples of a conversation between a human user and a computer /robot. The user input can be voice, text or image or combination of different inputs. Considering emotion of the human user, MMC will output responses in a text, speech, music depending on the user’s needs.

* Chats: “I am bored. What should I do now?” - “You look tired. Why don’t you take a walk?”
* Question Answering: “Who is the famous artist in Barcelona?” – “Do you mean Gaudi?”
* Information Request: “What’s the weather today?” – “It is a little cloudy and cold.”
* Action Request: “Play some classical music, please” - “OK. Do you like Brahms?”

So far, the AIMs required by the following application areas have been considered for possible standardisation by MPAI-MMC:

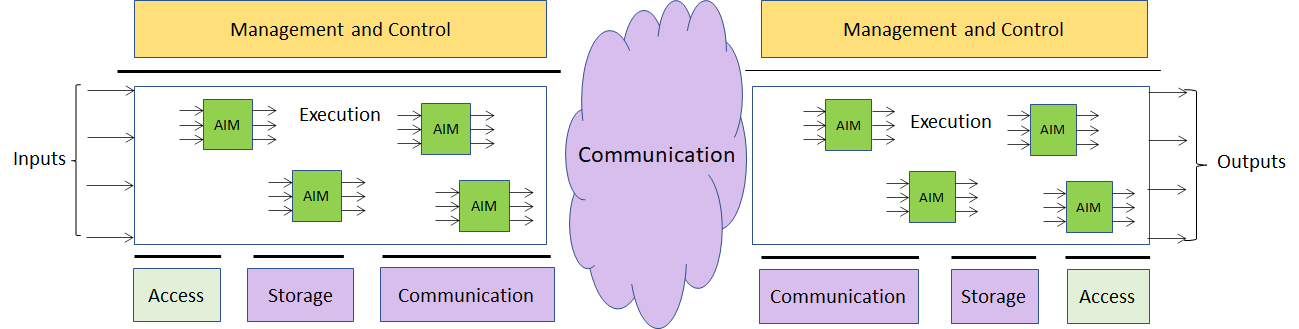
1. Conversation with emotion: a human-machine conversation system where the computer can recognize emotion in the user’s speech to produce a reply
2. Multimodal Question Answering: a human-machine Question Answering system where the human asks questions to the computer presenting an image
3. Personalized Automatic Speech Translation: a system that recognizes a voice uttered in a language by a speaker, converts the recognized voice into another language through automatic translation, and outputs a converted voice as text-type subtitles or as a synthesized voice

# AI Framework

Most MPAI applications considered so far can be implemented as a set of AIMs – AI/ML and even traditional data processing based with standard interfaces assembled in suitable topologies to achieve the specific goal of an application and executed in an MPAI-defined AI Framework. MPAI is making all efforts to iden­tify processing modules that are re-usable and upgradable without necessarily changing the logic of the application.

MPAI plans on completing the development of a 1st generation AI Framework called MPAI-AIF in July 2021.

The MPAI-AIF Architecture is given by *Figure 1*



*Figure 1 –The MPAI-AIF Architecture*

Where

1. *Management and Control* manages and controls the AIMs, so that they execute in the correct order and at the time when they are needed.
2. *Execution* is the environment in which combinations of AIMs operate. It receives external inputs and produces the requested outputs both of which are application specific interfacing with Management and Control and with Communication, Storage and Access.
3. *AI Modules* (AIM) are the basic processing elements receiving processing specific inputs and producing processing specific
4. *Communication* is required in several cases and can be implemented, e.g. by means of a service bus and may be used to connect with remote parts of the framework
5. *Storage* encompasses traditional storage and is used to e.g. store the inputs and outputs of the individual AIMs, data from the AIM’s state and intermediary results, shared data among AIMs.
6. *Access* represents the access to static or slowly changing data that are required by the application such as domain knowledge data, data models, etc.

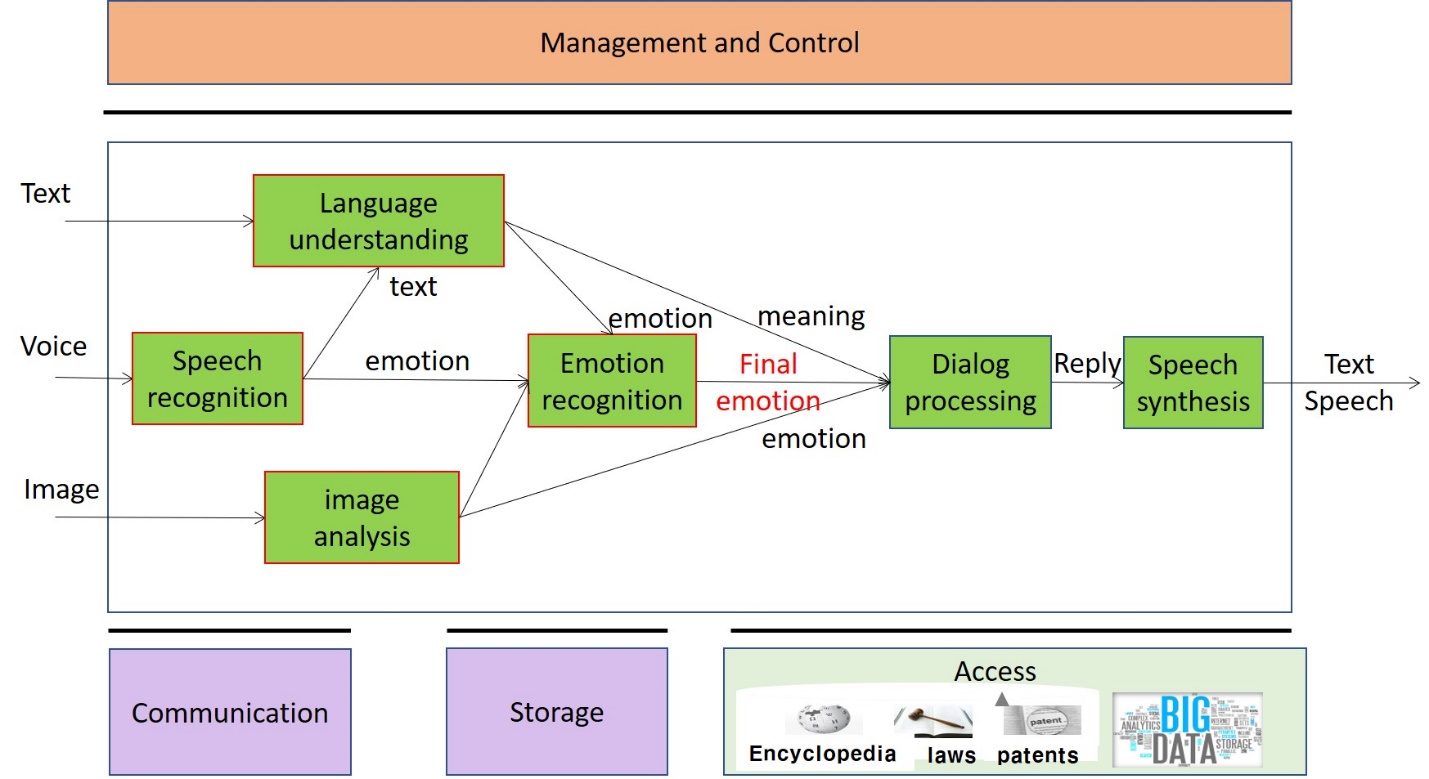
# MPAI-MMC work plan

In this chapter there is currently one application area with its relevant AI Modules (AIM) identified and described, and their inputs/outputs summarily specified.

## Conversation with emotion

One example of MPAI-MMC is the case of conversation with emotion instance. When people talk, they use multiple modalities: speech, facial expression, text, sign languages and gesture. Emotion is one of the key features to understand the meaning of the utterances that are made by the speaker. Therefore, a conversation system should have the capability to recognize emotion to understand the user’s speech and produce the reply as the output.

The AIMs implied by a multi-modal conversation system would look approximately as presented in *Figure 2*. The interaction between different AIMs are described including a language understanding module, a speech recognition module, an image analysis module, an emotion recognition module, a dialog processing module, and a speech synthesis module.



*Figure 2 – Conversation with emotion*

The following *Table 1* lists the AIMs and their inputs and outputs.

*Table 1 – AI Modules interactions*

|  |  |  |  |
| --- | --- | --- | --- |
| **AI Module** | **Input** | **Output** | **External data** |
| Language understanding (LU) | Text | Emotion | Emotion ontology |
|  | Text from ER | Meaning |  |
| Speech recognition (SR) | Voice | Text | Emotion ontology |
|  |  | Emotion |  |
| Speech synthesis (SS) | Reply from DP |  |  |
| Emotion recognition (ER) | Emotion from LU | Final emotion | Emotion ontology, Emotion model |
|  | Emotion from SR |  |  |
|  | Emotion from IA |  |  |
| Image analysis (IA) | Image | Emotion | Emotion ontology |
| Dialog processing (DP) | Meaning from LU | Reply | Dialog model, Dialog Knowledge Base |
|  | Final emotion from ER |  |  |
|  | Emotion from IA |  |  |

In the following subsections each AIM is analysed in detail.

### Language understanding

|  |  |
| --- | --- |
| **Function** | To analyse natural language in a text format to produce its meaning and emotion included in the text |
| **Inputs** | Text  Text from Emotion Recognition |
| **Outputs** | Emotion  Meaning |
| **External data** | Emotion ontology |

### Speech recognition

|  |  |
| --- | --- |
| **Function** | To analyse the voice input and generate text output and emotion it takes |
| **Inputs** | Voice |
| **Outputs** | Text  Emotion |
| **External data** | Emotion ontology |

### Speech synthesis

|  |  |
| --- | --- |
| **Function** | To produce speech from the input text |
| **Inputs** | Reply from Dialogue Processing in the text form |
| **Outputs** | Speech |
| **External data** |  |

### Emotion recognition

|  |  |
| --- | --- |
| **Function** | To determine the final emotion from multi source emotions |
| **Inputs** | 1. Emotion from Language Understanding 2. Emotion from Speech Recognition 3. Emotion from Image Analysis |
| **Outputs** | Final Emotions with proportions (ex. 80% happy with 20% surprise) |
| **External data** | Emotion ontology, Emotion model |

### Image analysis

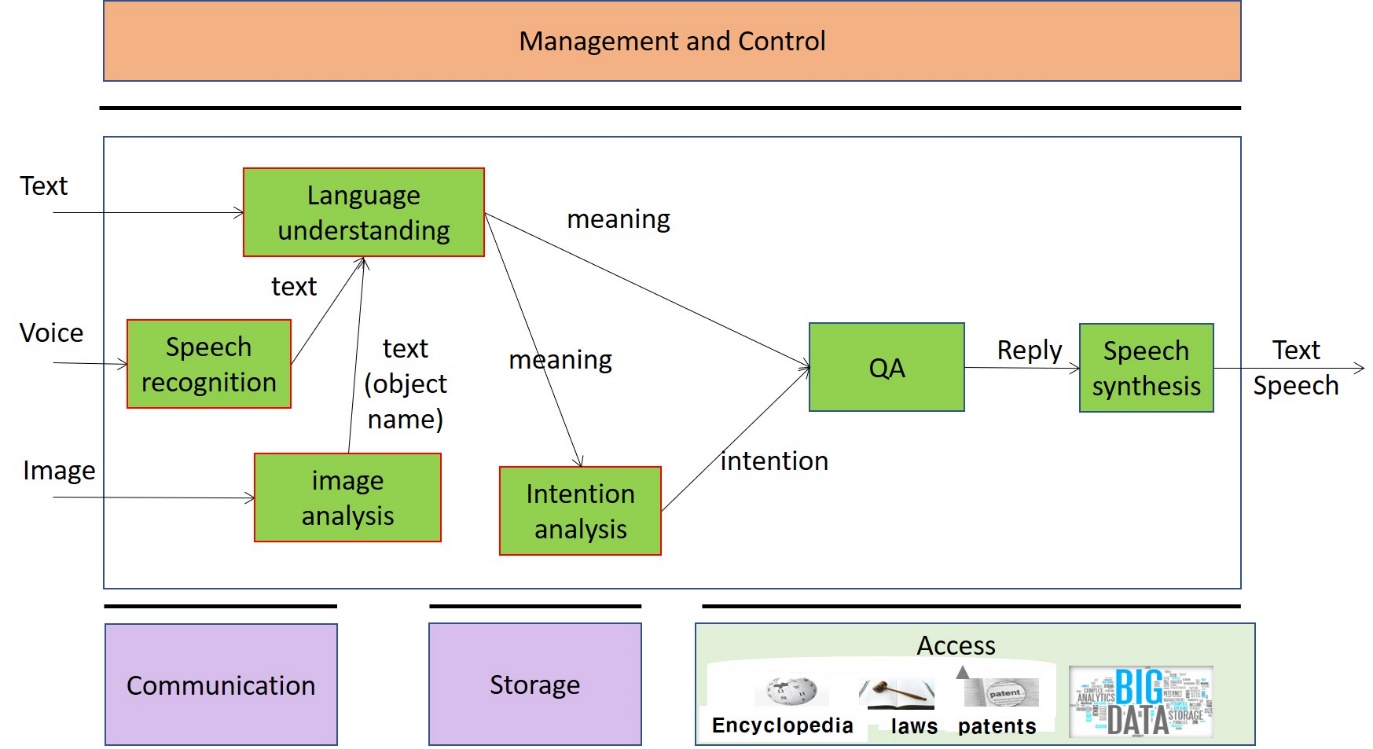
|  |  |
| --- | --- |
| **Function** | To analyse image and produce the emotion it takes |
| **Inputs** | Image |
| **Outputs** | Emotion |
| **External data** | Emotion ontology |

### Dialog processing

|  |  |
| --- | --- |
| **Function** | To analyse user’s utterance and produce a reply based on the user’s intention and emotion |
| **Inputs** | 1. Meaning from Language Understanding 2. Final emotion from Emotion Recognition 3. Emotion from Image Analysis |
| **Outputs** | Reply in natural language in the text form |
| **External data** | Dialog model, Dialog Knowledge Base |

## Multimodal Question Answering

Question Answering System (QA) is a technology that answers to a user’s question presented in natural language. Current QA system only deals with the case where input is in “text” form or “speech” form. However, more attention is paid these days to the case where mixed inputs such as speech with a image are presented to the system. For example, a user can ask a question about a picture which contains some specific tool as in “Where can I buy this tool?” showing the picture of the tool. In that case, the QA system should process the question in a text along with the image and should find out the answer to the question. *Figure 3* illustrates the multimodal question an­swering system with several AIMs to deal with the example question.



*Figure 3* – *Multimodal Question Answering*

The following *Table 2* lists the AI Modules and their inputs and outputs.

*Table 2 – AI Module interactions*

|  |  |  |  |
| --- | --- | --- | --- |
| **AI Module** | **Input** | **Output** | **External data** |
| Language understanding (LU) | Text, text from SR | Meaning | Dictionaries, Language model |
|  | Text from IA |  |  |
| Speech recognition (SR) | Voice | Text | Acoustic model, Language model |
| Speech synthesis (SS) | Answer from QA | Speech |  |
| Intention analysis (IA) | Meaning from LU | Intention | Intention ontology, Intention model |
| Question Answering (QA) | Meaning from LU | Answer | Wikipedia, question ontology |
|  | Intention from IA |  |  |
| Image analysis (IA) | Image | Object name | Image DB |

In the following subsections each AIM is analysed in detail.

### Language understanding

|  |  |
| --- | --- |
| **Function** | To analyse natural language in a text format to produce its meaning. |
| **Inputs** | Text from input, speech recognition and image analysis |
| **Outputs** | Meaning |
| **External data** | Dictionaries, Language model |

### Speech recognition

|  |  |
| --- | --- |
| **Function** | To analyse the voice input and generate text output |
| **Inputs** | Voice |
| **Outputs** | Text |
| **External data** | Acoustic model, Language model |

### Speech synthesis

|  |  |
| --- | --- |
| **Function** | To produce speech from the input text |
| **Inputs** | Answers from Question Answering in the text form |
| **Outputs** | Speech |
| **External data** |  |

### Intention Analysis

|  |  |
| --- | --- |
| **Function** | To determine the intention from the sentence meaning |
| **Inputs** | Meaning from Language Understanding |
| **Outputs** | Intention |
| **External data** | Intention ontology, Intention model |

### Image analysis

|  |  |
| --- | --- |
| **Function** | To analyse image and produce the object name in focus |
| **Inputs** | Image |
| **Outputs** | text |
| **External data** | Image DB |

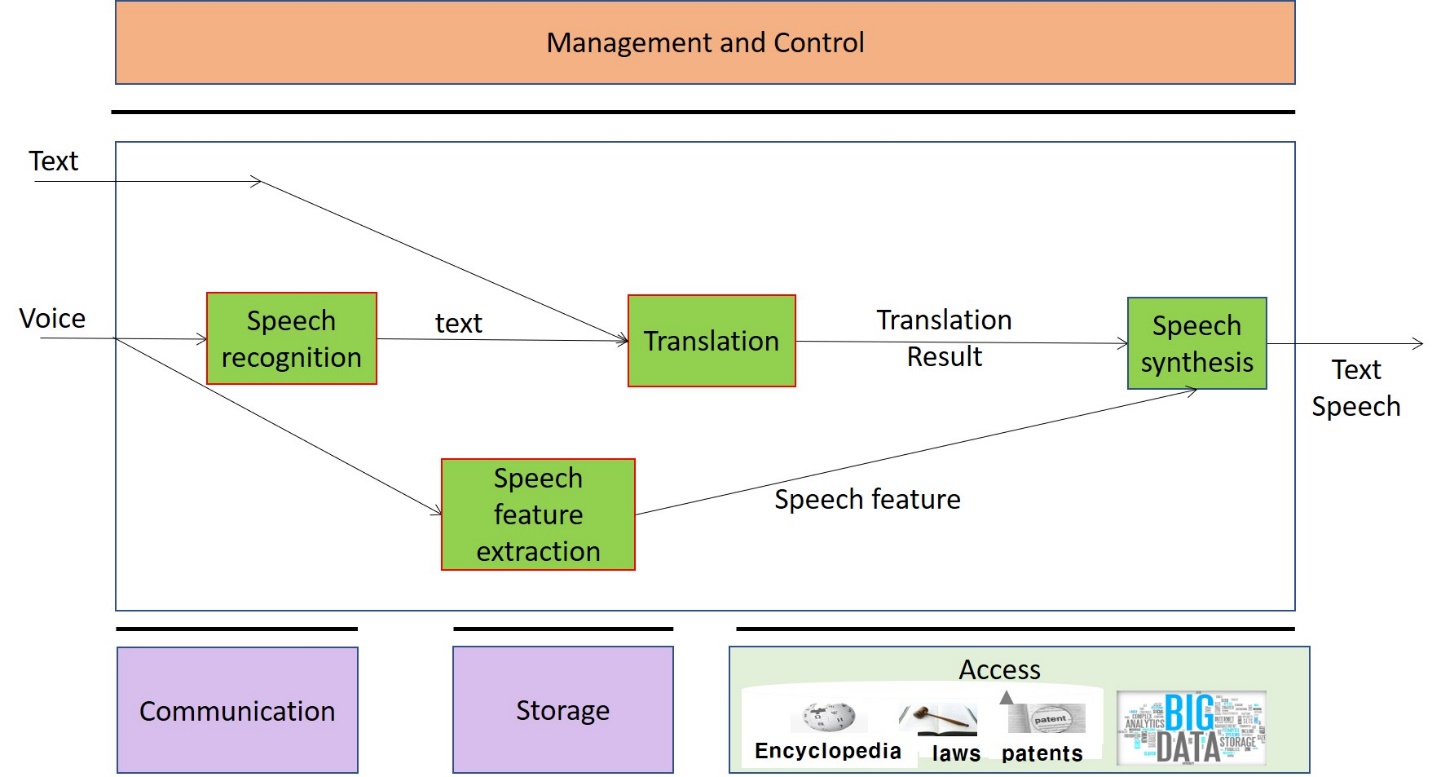
### Question Answering

|  |  |
| --- | --- |
| **Function** | To analyse user’s question and produce the reply based on the user’s intention |
| **Inputs** | Meaning from Language understanding  Intention from Intention analysis |
| **Outputs** | Answer in natural language in the text form |
| **External data** | Wikipedia, question ontology |

## Personalized Automatic Speech Translation

Automatic speech translation technology denotes technology that recognizes a voice uttered in a language by a speaker, converts the recognized voice into another language through automatic translation, and outputs a converted voice as text-type subtitles or as a synthesized voice. Recently, as interest in voice synthesis among main technologies for automatic interpretation increases, personalized voice synthesis instead of simple communication is being researched. Personalized voice synthesis denotes technology that outputs a target language through voice recognition and automatic translation, as a synthesis voice similar to a tone (or an utterance style) of a speaker.

The AI Modules implied by a personalized automatic speech translation system would look approximately as presented in Figure 4. The interaction between different AIMs are described including a speech recognition module, a speech feature extraction module, a translation module and a speech synthesis module.



*Figure 4* – *Personalized Automatic Speech Translation*

The following *Table 3* lists the AI Modules and their inputs and outputs.

*Table 3 – AI Module interactions*

|  |  |  |  |
| --- | --- | --- | --- |
| **AI Module** | **Input** | **Output** | **External data** |
| Speech recognition (SR) | Voice | Text | Acoustic model, Language model |
| Speech feature extraction (SF) | Voice | Speech features | Speech feature DB |
| Translation (TR) | Text input, text from SR | Text (translation result) |  |
| Speech synthesis (SS) | Text from TR | Text or personalized speech |  |
|  | Speech features from SF |  |  |

In the following subsections each AIM is analysed in detail.

### Speech recognition

|  |  |
| --- | --- |
| **Function** | To analyse the voice input and generate text output |
| **Inputs** | Voice |
| **Outputs** | Text |
| **External data** | Acoustic model, Language model |

### Speech feature extraction

|  |  |
| --- | --- |
| **Function** | To extract speech features such as tones, intonation, intensity, pitch, emotion, intensity or speed from the input voice  To encode personal voice features |
| **Inputs** | voice |
| **Outputs** | Speech features, hidden variable from the personal voice features |
| **External data** | Speech feature DB |

### Translation

|  |  |
| --- | --- |
| **Function** | To convert from the source language to the target language automatically |
| **Inputs** | Text in a source language which is the output of the speech recognition |
| **Outputs** | Text of translation results in target language |
| **External data** |  |

### Speech synthesis

|  |  |
| --- | --- |
| **Function** | To produce speech from the input text |
| **Inputs** | Translation result in the text form, speech features, hidden variable from the personal voice features |
| **Outputs** | Personalized Speech in target language |
| **External data** |  |

# Conclusions

The document in its current form is work in progress. MPAI intends to add more details to the existing document to enable MPAI to issue a Call for Technologies. MPAI may also add more usage exam­ples.

When the document will be considered sufficiently mature, MPAI will issue a Call for Technol­ogies requesting MPAI members and industry members to submit proposals for:

1. *Data formats* suitable as inputs and outputs of the identified AIMs
2. Possible *alternative partitioning* of the AIMs implementing the example cases providing
   1. Arguments in support of the proposed partitioning
   2. Detailed specifications of the inputs and outputs of the proposed AIMs
3. New *usage* *examples* fully described as in the final version of this document.

Respondents will be asked to state in their submissions their intention to adhere to the Framework Licence developed for MPAI-MMC when licencing their technologies if included in the MPAI-MMC standard. Please note that “a Framework Licence is the set of conditions of use of a license without the values, e.g. currency, percent, dates etc.”. The *Framework Licence* willgive the MPAI-MMC standard a *clear IPR licensing* framework.

The MPAI-MMC Framework Licence will be developed, as for all other MPAI Framework Licences, in compliance with the gener­ally accepted principles of competition law.