INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2017-2020

SG16-TD359/PLEN

STUDY GROUP 16

Original: English

Question(s): 8/16 Geneva, 7-17 October 2019

TD

Source: Editor H.ILE-MMT a.i.

Title: Consent: H.430.4 (ex H.ILE-MMT) "Service configuration, media transport protocols and signalling information of MMT for Immersive Live Experience (ILE) systems" (New)

Purpose: Discussion

Contact: Jiro NAGAO
NTT
Japan
Tel: + 81 468-59-4582
Fax: +
E-mail: jiro.nagao.cd[at]hco.ntt.co.jp

Contact: Hideo IMANAKA
NTT
Japan
Tel: +81 422-36-7502
Fax:
Email: hideo.imanaka[at]ntt-at.co.jp

Keywords: ILE, MMT, media transport

Abstract: This TD contains the text of draft new Recommendation H.430.4 (ex H.ILE-MMT) "Service configuration, media transport protocols and signalling information of MMT for Immersive Live Experience (ILE) systems" that is proposed for Consent at this SG16 meeting.

This TD contains the text of draft new Recommendation H.430.4 (ex H.ILE-MMT) "Service configuration, media transport protocols and signalling information of MMT for Immersive Live Experience (ILE) systems" proposed for Consent at this meeting as an output of Q8/16 sessions held between 7-17 October 2019.

This output document has been developed based on the following documents, contribution and discussion results of the meeting:

– SG16-TD99/WP3: H.ILE-MMT “Service configuration, media transport protocols, signalling information of MMT for Immersive Live Experience systems”: Output draft from Q8/16 e-meeting (4 September 2019)

– SG16-C537-R1: H.ILE-MMT: Proposal of technical updates in H.ILE-MMT

– SG16-C538-R1: H.ILE-MMT: Proposal of editorial changes

CONTENTS

Page

List of Tables

Page

List of Figures

Page

Draft New ITU-T H.430.4 (ex H.ILE-MMT)

Service configuration, media transport protocol, signalling information of MMT for ILE systems

AAP Summary

ILE services are realized by several types of information such as video, audio, lighting and stage effects, and the information should be transferred synchronously from source site to viewing sites. This document identifies service configuration, system structure, media transport protocol and signalling information for Immersive Live Experience (ILE) systems using ISO/IEC 23008-1 (MPEG Media Transport). This specifies constrains to ISO/IEC 23008-1 for ILE systems.

Summary

This draft Recommendation identifies service configuration, media transport protocol, signalling information of MMT for Immersive Live Experience (ILE) systems, in order to provide ILE services.

Keywords

Immersive Live Experience, media transport, MPEG media transport, service configuration, descriptor, XML syntax

1 Scope

ILE systems consist of several devices such as cameras, displays and transmission networks from source site to viewing sites, as described in ITU-T Recommendation H.430.1, Requirements for immersive live experience (ILE) services. Synchronous signalling transmission including video and audio were studied in ISO/IEC JTC1/SC29/WG11 (MPEG), and ISO/IEC 23008-1 (MPEG Media Transport: MMT) is one of strong candidates for transporting synchronously several media from event site to remote sites on ILE systems. However, MMT might not consider transporting spatial information, such as X-Y-Z coordinate of objects, and stage effect information like lighting. In order to utilize MMT for ILE system, it needs to clarify the ILE profile of MMT, which includes some constraints of MMT specification, for example usage of optional attributes.

This draft ITU-T Recommendation identifies service configuration, system structure, media transport protocol and signalling information for Immersive Live Experience (ILE) systems using ISO/IEC 23008-1 (MPEG Media Transport). This specifies constrains to ISO/IEC 23008-1 for ILE systems.

The scope of this Recommendation includes:

– Service configuration and system structure for MMT-based ILE systems

– Media transport protocol for MMT-based ILE systems

– Signalling information for MMT-based ILE systems

2 References

The following ITU-T Recommendations and other references contain provisions, which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the

most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ISO/IEC 14496-3] ISO/IEC 14496-3 (2009), *Coding of audio-visual objects — Part 3: Audio*.

[ISO/IEC 14496-12] ISO/IEC 14496-12 (2015), *Coding of audio-visual objects — Part 12: ISO base media file format*.

[ISO/IEC 23008-1] ISO/IEC 23008-1 (2017), *Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 1: MPEG Media Transport (MMT)*.

[ITU-T H.265] Recommendation ITU-T H.265 | ISO/IEC 23008-2 (2019), *High efficiency video coding (HEVC).*

[H.430.1] Recommendation ITU-T H.430.1, *Requirements for immersive live experience (ILE) services*.

[BT.2074-1] Recommendation ITU-R BT.2074-1, *Service configuration, media transport protocol, and signalling information for MMT-based broadcasting systems*.

3 Terms and definitions

3.1 Terms defined elsewhere

This draft Recommendation uses the following terms defined elsewhere:

3.1.1 Immersive Live Experience (ILE) [ITU-T H.430.1]: A shared viewing experience which stimulates emotions within audiences at both the event site and remote sites, as if the ones at remote sites wandered into substantial event site and watched actual events in front of them, from high-realistic sensations brought by a combination of multimedia technologies such as sensorial information acquisition, media processing, media transport, media synchronization and media presentation.

3.2 Terms defined here

None.

4 Abbreviations

AAC	MPEG-4 advanced audio coding
ALS	MPEG-4 audio lossless coding
cc	closed caption
DMX	Digital Multiplex
HEVC	high efficiency video coding
ILE	Immersive Live Experience
MFU	Media Fragmentation Unit
MMT	MPEG Media Transport
MMTP	MMT protocol

MMT-SI	MMT signalling information
MPI	Media Presentation Information
MPT	MMT package table
MPU	media processing units
NAL	network abstraction layer
PA	package access
PI	Presentation Information
UTC	Coordinated Universal Time

5 Conventions

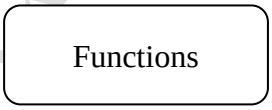
In this Recommendation:

– The keywords "is required to" indicate a requirement which must be strictly followed and from which no deviation is permitted, if conformance to this Recommendation is to be claimed.

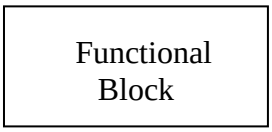
– The keywords "is recommended" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

– The keywords "can optionally" indicate an optional requirement that is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option, and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with this Recommendation.

– The keyword "functions" are defined as a collection of functionalities. It is represented by the following symbol in this Recommendation:



– The keyword "functional block" is defined as a group of functionalities that has not been further subdivided at the level of detail described in this Recommendation. It is represented by the following symbol in this Recommendation:



NOTE – In the future, other groups or other Recommendations may possibly further subdivide these functional blocks.

Frame borders of "functions" and "functional block", and relational lines among "functions" and "functional block" are drawn with solid lines or dashed lines. The solid lines mean required functionalities or relations. On the other hand, the dashed lines mean optional functionalities or relations.

6 Requirements to MMT for ILE services

Most of users want to watch sports games at real time and with high-realistic sensations. Real-time content delivery or live video streaming usually require time synchronization between transmitted video and audio. For displaying psudo-3D objects, it also requires time synchronization between the spatial information and the displayed objects. One of the key features on ILE services is live experience which can be realized by synchronization of multiple media, such as video, audio and spatial data stream of objects, video and audio streams of background, and other stage effects information.

In order to synchronize and transport video and audio combined with spatial information, MMT, an optimized protocol for media synchronization, can be utilized with the MMT Assets, describing the specific signalling information of the objects such as their three-dimensional size, position, and direction. This technology makes it possible to correlate the physical spatial parameters such as the size and position of the display device with Asset data (frame pixel data) so that the space can be reconstructed with high realism at the destination at the intended size. In addition, transmission of the Digital Multiplex (DMX) [b-DMX] signals, which are commonly used in production to control stage lighting and audio devices, as one of the MMT Assets enables realistic presentations that accurately synchronize remote stage equipment with the media.

MMT is defined in ISO/IEC 23008-1 (MPEG Media Transport) for streaming information, and it is one of the major technologies for synchronous media transport. In order to utilize MMT for ILE services, constrains to ISO/IEC 23008-1 for ILE systems need to be specified since MMT was designed for transporting multiple media synchronously and MMT does not focus on transmitting information required for ILE systems such as spatial information and stage effect information.

7 Service configuration and system structure of ILE

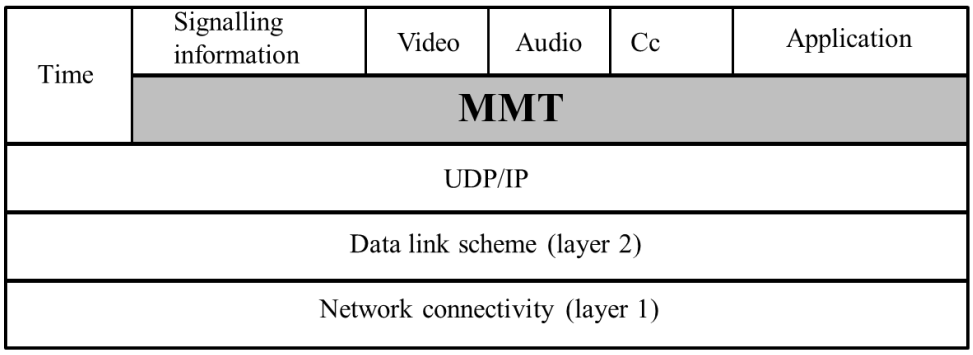
7.1 Service configuration of ILE

ISO/IEC 23008-1 specifies the MMT package as a logical structure of content. The MMT package includes presentation information (PI) and associated Assets that constitute content. For broadcasting services, ITU-R published Recommendation BT.2074-1, Service configuration, media transport protocol, and signalling information for MMT-based broadcasting systems. Because ILE services are not broadcasting services, [BT.2074-1] could not be utilized for ILE services as it is. For providing ILE services, various kinds of information collected from source sites, such as video stream data shot by multiple cameras, audio data collected by multiple microphones, the location information of objects and stage effect information including lighting control, need to be synchronously transmitted to one or more viewing sites, and reconstructed at the viewing sites.

In ISO/IEC 23008-1, an Asset is defined as a media component. An Asset is equivalent to a series of media processing units (MPUs). In ILE systems, one entertainment programme is an MMT package including one or more Assets and signalling information. A package access (PA) message is an MMT signalling information (MMT-SI), and the MMT package table (MPT) carried in the PA message identifies Assets constituting the ILE programme.

7.2 System structure

This section describes the general structure of MMT-based ILE systems. Figure 7-1 shows the protocol stack of MMT for ILE services based on the protocol stack for broadcasting services written in [BT.2074-1].



Based on BT.2074-0

Figure 7-1 – Protocol stack of MMT for ILE services

In ILE systems, most features of MMT-based broadcasting systems could be utilized. Media components, such as video, audio, closed caption (cc) and stage effects including lighting, are encapsulated into media fragment units (MFUs)/ MPUs. They are carried as MMT payloads in IP packets.

The systems also have MMT-SI. MMT-SI is signalling information on the structure of an event program. MMT-SI is carried in MMT payloads in MMTP packets over IP.

In order for the receiver terminals to synchronize with the event venue, time information in Coordinated Universal Time (UTC) is also delivered in IP packets.

7.3 Spatial information

One of the most important information which realizes immersiveness is spatial information, such as the spatial location of objects. Because the screen sizes of viewing sites differ, it is necessary to provide information which can be used for adjusting the projection parameters to the presentation environment at each viewing site.

In order to reconstruct atmosphere of source site at the viewing sites, the spatial information at the source site needs to be transferred. The information can be transferred either by sending the size and location of the screens, speakers, and other objects one by one, or simply by using reference space.

Reference space provides an easier way to describe the spatial information at the source site in a normalized format, which can be in turn accessed in order to adjust to the actual environment at the viewing sites. The reference space is shown in Fig. 7-2.

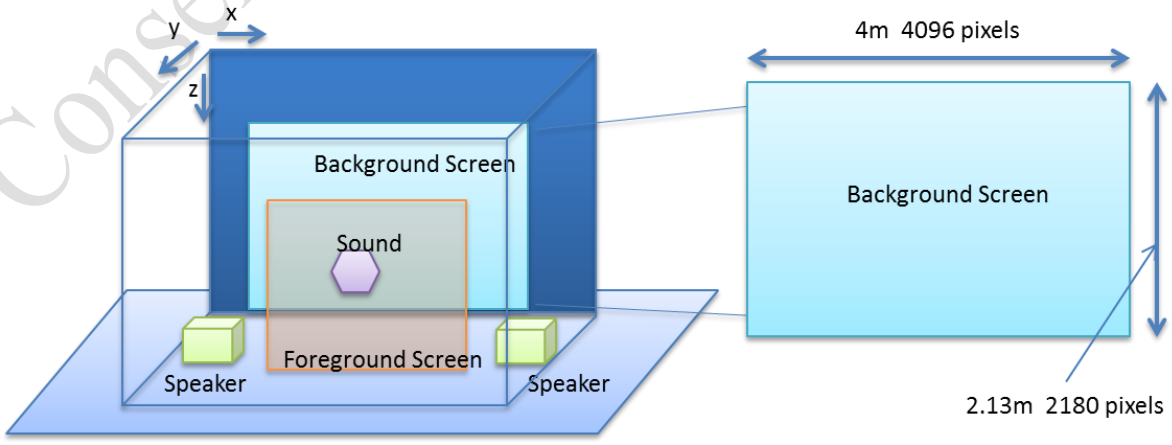


Figure 7-2 – Example of reference space describing the location and size of equipment

This kind of information could be transferred from the source site to the viewing sites by utilizing [ISO/IEC 23008-1] (MMT).

7.4 Lighting information

Most lighting devices can be controlled by DMX (Digital Multiplex: DMX512 referred to by ANSI standard USITT DMX512-A) [b-DMX], and there exist lighting devices which can be controlled by Art-Net (the specification of transporting DMX signals over UDP/IP) [b-Art-Net]. Art-Net may be highly compatible with MMT in terms of transmission over IP network. This sub-clause describes the way to encapsulate Art-Net by MMT.

When using DMX-capable devices, DMX signals are converted to Art-Net signals bilaterally, and the devices are connected to each other. The lighting information to be transferred to viewing sites is recommended to be treated as follows:

– At the event site
  - Converting DMX signals into Art-Net packets by using convertors
  - Receiving Art-Net packets by MMT servers
  - Creating MPU from Art-Net packets (or storing Art-Net packets into MPU), and creating MMT packets
  - Setting play time into MPU time stamp descriptors
  - Transferring MMT packets as MPU mode

– At the viewing sites
  - Reconstructing Art-Net packets from MMT packets
  - Playing Art-Net packet at designated time according to MPU time stamps
  - Controlling devices which are not capable Art-Net by DMX signals that are converted by convertors

8 Media transport protocols for ILE system

8.1 Encapsulation of multimedia data

In order to improve the interoperability of MMT-based ILE systems, the following constraint apply to carriage of multimedia data in MMT protocol (MMTP) packets.

8.1.1 MFU format for video and audio streams

[ISO/IEC 23008-1] specifies encapsulation format, but there are some variations of MFU format.

When a high efficiency video coding (HEVC) stream is transported in the MMTP, a network abstraction layer (NAL) unit is encapsulated into an MFU of the MMTP. If an HEVC encoder generates the byte stream format specified in Annex B of [ITU-T H.265], one start code prefix (0x000001) followed by one NAL unit is replaced with 4 bytes length information of the NAL unit (unsigned integer format). Each box defined in [ISO/IEC 14496-12] is selectively carried in MMT stream.

When an MPEG-4 advanced audio coding (AAC) stream or MPEG-4 audio lossless coding (ALS) stream is carried in the MMT protocol, one AudioMuxElement () specified in [ISO/IEC 14496-3] is encapsulated into an MFU of the MMTP for AAC stream and, a raw data stream is encapsulated into an MFU of the MMTP for ALS stream.

9 Signalling information for ILE system

[ISO/IEC 23008-1] specifies Signalling information to handle encapsulated Asset data in MMTP packets. However, these descriptors have a lot of flexibility. In order to handle Assets including spatial information for realizing ILE services and for ensuring interoperability of MMT-based ILE systems, the following constraints are applied to the signalling information for media descriptors.

9.1 Environment descriptor

For displaying objects at viewing sites of ILE services, a Presentation Information (PI) content in Media Presentation Information (MPI) tables is required to carry the spatial environment information. The syntax of the environment descriptor is provided in the Table 9-1, and the XML syntax sample of environment descriptor is shown in the following Table 9-2.

As one of media descriptor, the syntax of environment descriptor contains environmental information such as site and equipment information to reconstruct images at viewing sites. The information will be processed by media processing functions.

Table 9-1 – Syntax of environment descriptor

Name	Data	Form
environment		
@xmlns	Default	String
site	Site information	
@id	Site identification	String
@width	Width of site	decimal
@height	Height of site	decimal
@depth	Depth of site	decimal
@unit	Unit of size	string
equipments	Equipment info	
equipment	Equipment	
@id	Identification	string
@type	Type of equipment	string
position	Position	
location	XYZ location	
@x	x axis	decimal
@y	y axis	decimal
@z	z axis	decimal
@unit	Unit of location	String
rotation	Rotation information	
@x	Roll	Decimal
@y	Pitch	Decimal

	@z	Yaw	Decimal
	@unit	Unit of rotation	String
size		Size information	
	@width	Width	Decimal
	@height	Height	Decimal
	@depth	Depth	Decimal
	@unit	Unit of size	String
	offset	Offset	
	@left	Left offset	Decimal
	@right	Right offset	Decimal
	@top	Top offset	Decimal
	@bottom	Bottom offset	Decimal
	@unit	Unit of offset	String

Table 9-2 – XML Syntax sample of environment descriptor

```
<?xml version="1.0" encoding="UTF-8"?>
<environment xmlns="http://xxx.yyy.zz/mmt/artnet">
  <!—Environment information  -->
  <site id="main" width="1000" height="1500" depth="10000" unit="mm"/>

  <!—Equipments information -->
  <equipments>
    <equipment id="screen 1" type="screen">
      <position>
        <location x="10" y="20" z="130" unit="mm"/>
        <rotation x="10" y="20" z="30" unit="deg"/>
      </position>
      <size width="200" height="100" depth="1" unit="mm">
        <offset left="10" right="10" top="10" bottom="20" unit="mm"/>
      </size>
    </equipment>
    <equipment id="light" type="lighting">
      <position>
        <location x="110" y="120" z="1130" unit="mm"/>
        <rotation x="45" y="90" z="70" unit="deg"/>
      </position>
      <size width="1000" height="1000" depth="1000" unit="mm"/>
    </equipment>
  </equipments>
</environment>
```


9.2 Object recognition and location descriptor

In order to achieve auditory lateralization on large-sized screens at viewing sites of ILE services, it could use wave field synthesis. Object recognition information and location information should be transferred from source site to viewing sites, so that audio from the object could be homologized to

the displayed objects at the viewing sites. In addition, the objects displayed at viewing sites could be optimized to the different display facilities by using the object recognition and location information. An object recognition and location descriptor is transferred on MMT. In most cases, objects move around in the source site, so location information of the objects should be captured in each frame. In other words, object recognition and location information are necessary to reproduce the event at viewing sites.

As one of the media descriptors, the syntax of object recognition and location descriptor contains spatial information of objects to reconstruct audio direction aligned with displayed image of objects at the viewing sites. The information will be processed by media processing functions.

ILE has several viewing styles, and one of the typical ILE services is Amphitheatre (Arena style) viewing, which enables all audiences to share one stage from all directions. This kind of omni-directional viewing style needs special information such as X-Y-Z coordinate of moving objects in addition to background information for displaying objects.

Objects are moving around in the arena at event site, so it is required to reconstruct actual trajectory of moving objects. For this purpose, ILE system needs to transport tracking information through MMT. The tracking information is used to specify sound source and associate it with the object profiles. In order to reconstruct 3D images in omni-directional display, some other data are also required such as distance of objects sensed by laser.

The data schema of moving objects, which is transferred on MMT, is shown in Fig. 9-1.

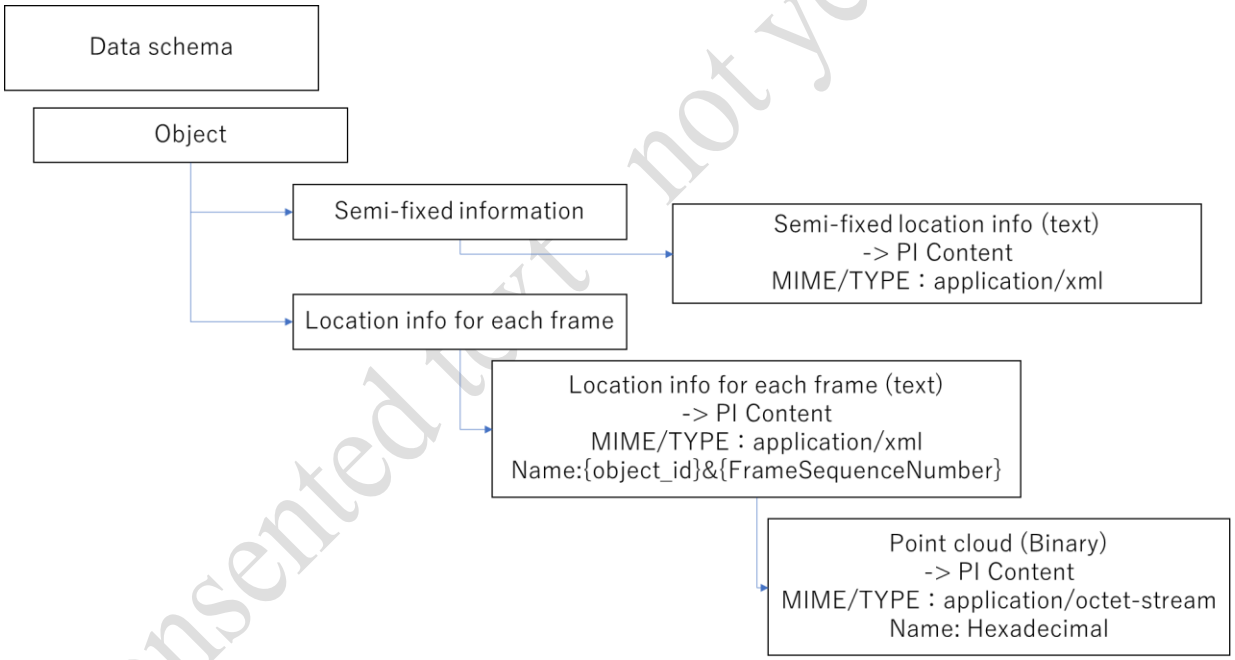


Figure 9-1 – Data schema of moving objects

Object information in each frame consists of object label, location information in the global coordinates, location information in the local coordinates, and location information in the image coordinates. Syntax of location information includes centre of mass, rectangle and point cloud. Camera ID is used to identify the camera which captured images, and the direction of shooting angle.

The syntax of location_data and XML syntax sample are shown in Table 9-3 and 9-4. The data schema in Fig. 9-1 has “Point cloud” for future development, whose syntax is not elaborated on this document.

Table 9-3 – Syntax of location_data

Name	Data	Form	Repetition
location_data	Three-dimensional location information of each object		1
timecode	Time at which location info is consumed, which is denoted by UTC for time synchronization between objects	String	1
object	Name of the target object	String	0..*
label	The label of target object, which can be used for object identification	String	0..*
global_locations	Location of the object in the world coordinates		0..1
center_point	Center of mass of the object		0..1
point	Position of the center of mass		1
x	X-coordinate	decimal	1
y	Y-coordinate	decimal	1
z	Z-coordinate	decimal	1
unit	Unit of the coordinates	String	1
rectangle	Rectangle covering the target object		0..1
point	Dimensions of the rectangle		2
x	X-coordinate (width)	decimal	1
y	Y-coordinate (height)	decimal	1
z	Z-coordinate (depth)	decimal	1
unit	Unit of the coordinate	String	1
local_locations	Location information in the local coordinate system		0..1
center_point	Center of mass		0..*
camera_id	ID of camera	decimal	1
point	3D coordinates		1
x	X-coordinate (width)	decimal	1
y	Y-coordinate (height)	decimal	1
z	Z-coordinate (depth)	decimal	1
unit	Unit of the coordinates	String	1
rectangle	Rectangle covering the target object		0..*
camera_id	ID of camera	decimal	1
point	3D coordinates		2
x	X-coordinate (width)	decimal	1
y	Y-coordinate (height)	decimal	1
z	Z-coordinate (depth)	decimal	1
unit	Unit of the coordinates	String	1
image_locations	Location information in the image coordinate system		0..1
rectangle	Rectangle covering the target object		0..*
camera_id	ID of camera	decimal	1
w	Width	decimal	1
h	Height	decimal	1
point	2D coordinates		2
x	X-coordinate (width)	decimal	1
y	Y-coordinate (height)	decimal	1
unit	Unit of the coordinates	String	1

Table 9-4 – XML syntax sample of location_data

```
<?xml version="1.0" encoding="UTF-8" ?>
     <location_data>
          <timecode>DF9F7CB944EF4217</timecode>
          <object>
               <label>Object0</label>
               <local_locations>
                    <center_point>
                         <camera_id>VC002</camera_id>
                         <point>
                          <x>392</x>
                          <y>129</y>
                          <z>6405</z>
                          <unit>mm</unit>
                          </point>
                    </center_point>
                    <rectangle>
                         <camera_id>VC002</camera_id>
                         <point>
                          <x>523</x>
                          <y>456</y>
                          <z>6331</z>
                          <unit>mm</unit>
                          </point>
                         <point>
                          <x>262</x>
                          <y>-197</y>
                          <z>6480</z>
                          <unit>mm</unit>
                          </point>
                    </rectangle>
               </local_locations>
               <image_locations>
                    <rectangle>
                         <camera_id>VC002</camera_id>
                         <w>1920</w>
                         <h>1080</h>
                         <point>
                          <x>2176</x>
                          <y>1349</y>
                          <unit>pixel</unit>
                          </point>
                         <point>
                          <x>2037</x>
                          <y>969</y>
                          <unit>pixel</unit>
                          </point>
                    </rectangle>
               </image_locations>
          </object>
     </location_data>
```

The semi-fixed information periodic_data contains size of event venue, object profile information, and camera information. The syntax of the periodic_data, which is stored separately in PI contents of MPI table, is shown in Table 9-5.

Table 9-5 – Syntax of periodic_data

Name	Data		Repetition
periodic_data	Semi-fixed location information		1
Site	Venue information		1

		Width	Width	decimal	1
		Height	Height	decimal	1
		Depth	Depth	decimal	1
		Unit	Unit	String	1
	Object		Object		1..*
		Label	Object label	decimal	1
		Profile	Profile	String	1
		Property	Property	String	1
	Camera		Camera		1..*
		Id	ID for camera	decimal	1

The timing and repetition of data sent from sensors might differ by sensor types, thus each application might require different conditions such as repetition and timing of receiving the sensed data. In addition, it might occur that some objects could not be detected in the cases where several objects exist. In these cases, simple processing, such as ignoring the missing information and waiting for all required data might cause lower accuracy of location information, and also longer processing time. In order to solve above problems, the timing and repetition of sending and receiving data needs to be aligned with the designated frame rate for synchronizing sensed data and image.

Bibliography

[b-DMX] ANSI E1.11 – 2008 (R2013), *Entertainment Technology USITT DMX512-A Asynchronous Serial Digital Data Transmission Standard for Controlling Lighting Equipment and Accessories*.

[b-Art-Net] Artistic License
http://www.artisticlicence.com