119th MPEG Turin, Italy, 17 - 21 July 2017, Meeting Report

Panos Kudumakis and Krishna Chandramouli
qMedia, Queen Mary University of London

Contents

1 Visual Identity Management Application Format

The basic objective of preserving privacy protection is to enable security and confidentiality in the multimedia content chain. Many usages of image/video communication services, social networking and video sharing platforms have led to an increasing works to protect user's privacy.

Traditionally, multimedia data security is achieved by methods of cryptography, which deals with encryption of data. This approach is called Naive Encryption Algorithm (NEA) and it treats the video bitstream as text data without paying attention on the compressed video structure. To this end, MPEG Common Encryption has been standardized in order to support encryption and key mapping methods in ISO Base Media File Format (ISO BMFF) files and in MPEG-2 Transport Streams. Consequently, bitstreams encrypted by those standards are decodable only after a correct decryption process even when only parts of the video are encrypted. Nevertheless, none of these formats allow signalling the encryption of a part of the picture (region), or indicating to the video decoder that the encrypted bitstream can be partially decoded.

Moreover, all the access control is provided and done globally without taking into account the image/video content and context. To restore citizens' confidence in online data collection practices, submitted media should be encrypted to protect privacy and can only be viewed with a limited access, that the user chooses: group of people, purpose of sharing, time, date, metadata, etc.

In order to provide privacy protection over processing and sharing of multimedia content, a flexible, effective and scalable mechanism is to provide users a way to express their control desires in a form that can be processed and monitored systematically, consistently and persistently throughout the lifecycle of the multimedia content. There is currently no standardized format to represent privacy description information (PDI), hindering the interoperability between secured systems.

MPEG-A Visual Identity Management Application Format specifies the standard representation of the set of signaling and data used in the process of preserving privacy for the storage or the sharing of image/video.

In particular, to protect privacy content, stored and/or shared media should be encrypted by the service's user and should only be viewed with a well-defined limited access (group of people, purpose of sharing, time, date, metadata, etc.). Consequently, some particular regions of the video (e.g., human faces, text data …) could only be seen by the authorized users, regardless of who captured and shared the video. Additionally, since multiple regions may need different protections (e.g. multiple faces shown in the video), there is also a need to manage different control access (i.e. different key identifiers) within the same video bitstream, potentially for each frame, as shown in Figure 1.

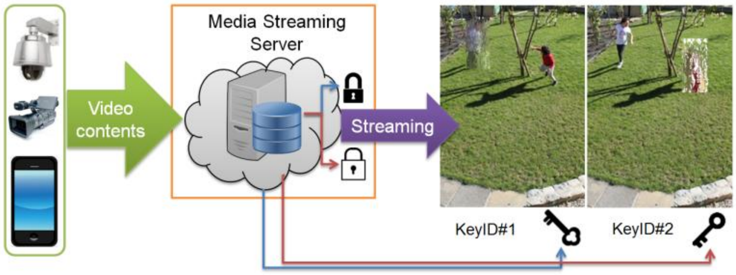


Figure 1 - Privacy Management in multimedia streaming applications

Figure 2 illustrates an example of framework for managing privacy of users when picture or video are taken and being shared among uses with different keys. The steps of Privacy protection mechanism can be expressed by the following walkthrough and graphically depicted in Figure 2:

1. User 1 :
    a. Capture a media
    b. Select part of the media considered as ‘private’ or/and let an application detect and recognize automatically faces
    c. Transmit information to the Privacy description Management
2. Privacy description Management:
    a. Get information and manage access control thanks to privacy politics defined by User 1 with the different relative User Descriptions, potentially taking into account Context descriptions and/or Service Descriptions through the Recommendation Engine.
    b. Send an Unique ID of the media, and a list of Encryption Keys associated to different part of the media
3. User 1:
    a. Generation (i.e. compression and encapsulation) of media file with an encryption scheme fed by the list of Encryption keys and their location.
    b. Storage or transmission to dedicated server for media sharing.
4. User 2:
    a. Get media file
    b. Send the ID of media and associated context and user description to Privacy description Management
5. Privacy description Management:
    a. Use the transmitted information through the Recommendation Engine to evaluate whether the user can be totally or partially authorized to render the associated media.
    b. Send Decryption keys adapted to the user who are allowed or not to see each part of the media.
6. User 2:
    a. Get Decryption keys
    b. Render (i.e. de-capsulate and decompressed) of media with an appropriated decryptions scheme depending on associated authorization.

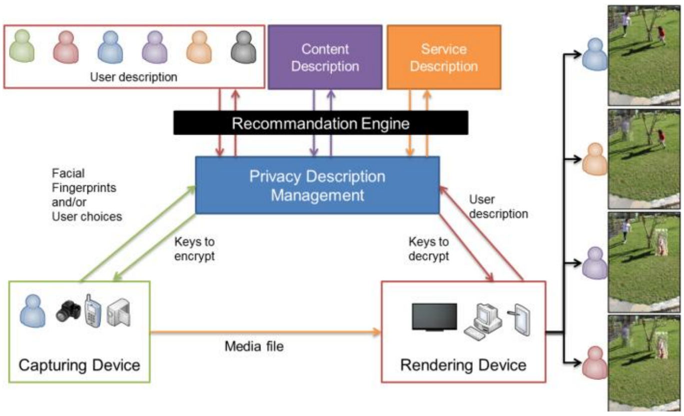


Figure 2 - Proposed Framework for privacy management of media

Note that ISO/IEC 15938-13:2015 Compact descriptors for visual search (CDVS) should be used to represent fingerprint of the face of a user in a picture. For video, CDVA might be used.

Output Documents
N16956 - WD of ISO/IEC 23000-21 Visual Identity Management AF

2 Sample Variants in ISO Base Media File Format

This part of ISO/IEC 23001 defines a framework for the carriage of Sample Variants in the ISO Base Media File Format (ISOBMFF), as specified by ISO/IEC 14496-12.

Sample Variants are typically used to provide forensic information in the rendered sample data that can, e.g. identify the Digital Rights Management (DRM) client. This variant framework is intended to be fully compatible with ISOBMFF and Common Encryption (CENC), as specified by ISO/IEC 23001-7, and agnostic to the particular forensic marking system used.

The Sample Variant framework uses three core constructs to define and carry Sample Variant data in ISOBMFF: Variant Constructors, Variant Byte Ranges and Variant Samples.

Figure 3 shows a scenario where a sample (Sample 2) has a number of Sample Variants. Figure 3 shows 3 samples which are encrypted in a series left to right, the middle of which has variants. The top row is a conceptual depiction of what is encoded using ISOBMFF and the bottom row shows what is output after Sample Variant processing. Access to samples is under the control of Key Identifiers (KIDs) as depicted in the top row of Figure 3. For Sample Variants, a hierarchy of KIDs is used to provide access to data, with the higher-level KIDs providing access to Sample Variant Metadata and the lower level KIDs providing access to media data.

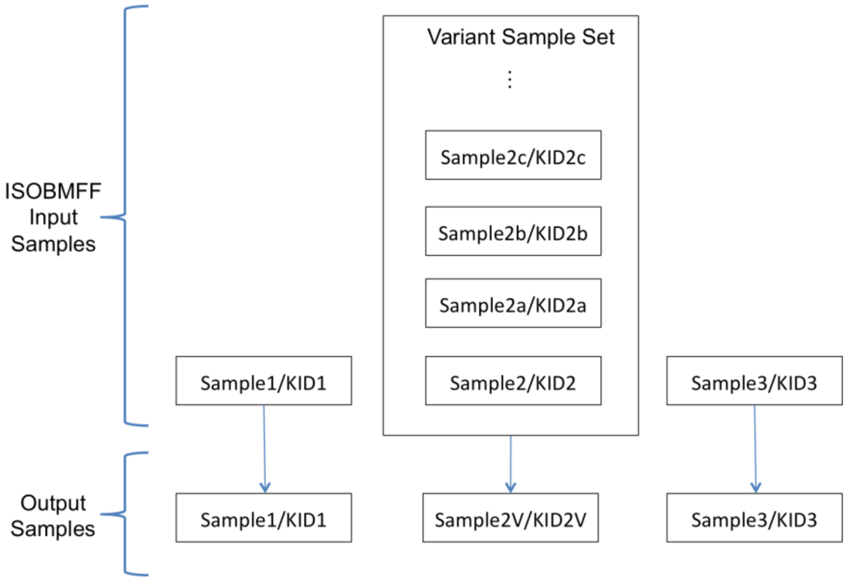


Figure 3 - Sample Variant structure

The control point for the use of the proposed framework is the content publisher:

- the content publisher will encode encrypted, compressed Sample Variant data into the ISOBMFF file and ensure that each set of Sample Variant data for a given sample time is encrypted with a key and signalled with a KID.

- the content publisher will work with the DRM to manage the release of KIDs/keys such that the playback path (the actual sample data used during playback) is controlled and the player can only decrypt and render the data that it has been authorized to render.

The decoder model for the processing of the file is shown in Figure 4. Control over if and how the Sample Variants are processed is critical to the Sample Variant decoding.

By operating in the encrypted/compressed domain, secure baseband link operation (e.g. dedicated, secure video pathways) is preserved and is intended to be fully compatible with CENC.

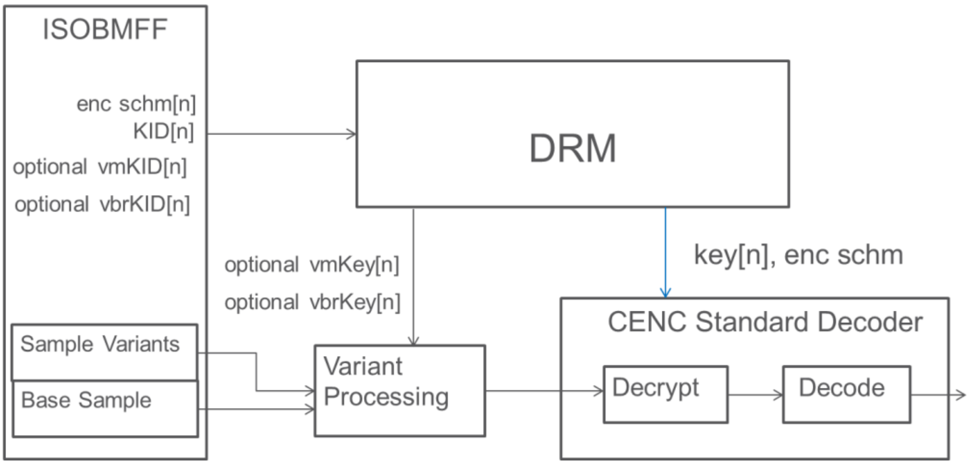


Figure 4 - Variant Decoder Model

The Sample Variant framework can be used in an application in which the content providers assigns different playback path of Sample Variants for each device as additional layer of content protection. By assigning the pre-defined playback path for each device, the content providers can detect the specific device which the content was played from.

Output Documents
N16961 - Text of ISO CD 23001-12 2nd edition Sample Variants
N16962 - Whitepaper for Sample Variants in the ISO base media format

3 MPEG-I Audio Architecture and Evaluation Procedures for 6DoF

It is understood that 23090-4, MPEG-I Audio will be based upon 23008-3:201x, MPEG-H 3D Audio Second Edition.

MPEG-H 3DA includes all coding technology that is necessary to carry 3D Audio content for 6DoF by supporting channels, objects and HOA. The core waveform carriage and essential metadata part of MPEG-H 3DA will be employed also by MPEG-I 6DoF. The “back-end” part of MPEG-H 3DA is dedicated to rendering these waveforms to any flexible output speaker configuration or binaurally to headphones. MPEG-H 3DA supports rotation of the presented sound scene in response to user interaction as indicated by (yaw, pitch, roll), or 3DoF. However, it is desired to support VR/AR applications that have full 6DoF of user interaction. To achieve this, additional user interaction as indicated by (x, y, z) translation has to be added to MPEG-H 3DA.

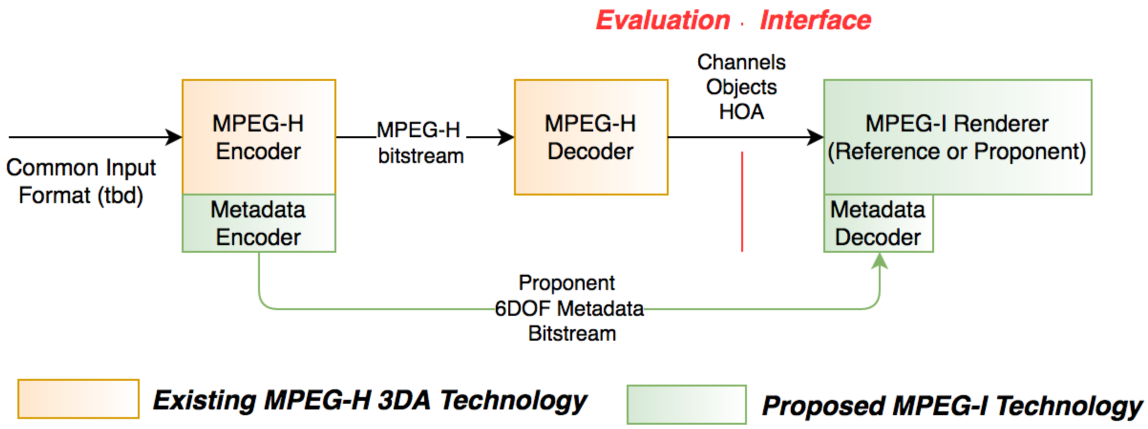
The VR and AR use-cases emphasize the importance of headphone rendering. In addition to headphone-based listening, a key component of the VR and AR 6DoF use-cases is the incorporation of *time-varying interactive rendering of linear translation with sufficiently low motion-to-rendered-output latency* compared to that currently supported by MPEG-H 3DA.

Thus, the new additional technology to be defined for MPEG-I 6DoF Audio consists of

- New interactive rendering technology, also supporting linear translation, for objects (including channel, i.e. objects with a fixed placement) and HOA which fully supports the VR/6DoF use cases.
- Additional metadata needed to support the VR/6DoF applications beyond what is already available from MPEG-H 3D Audio. This can be packaged in an extra bitstream container such that the content is still compatible to MPEG-H 3D Audio rendering.

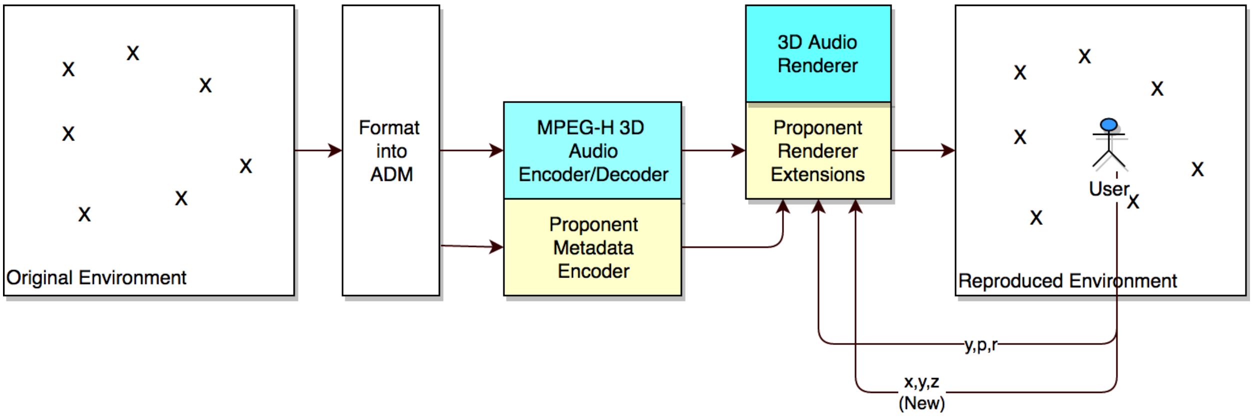
In view of this, evaluation of newly proposed technology consists of an evaluation of the decoding & interactive rendering process where each technology candidate would essentially perform

- decoding of a common reference MPEG-H bitstream representing the content elements as any combination of channels, objects or HOA
- decoding of newly defined individually generated VR/6DoF extension bitstreams containing the additional metadata that is needed for VR/6DoF application
- newly defined interactive rendering of the decoded item to the output devices (headphones and/or loudspeakers)


Evaluation · Interface
Common Input Format (tbd)
MPEG-H Encoder
Metadata Encoder
MPEG-H bitstream
MPEG-H Decoder
Channels Objects HOA
MPEG-I Renderer (Reference or Proponent)
Metadata Decoder
Proponent 6DOF Metadata Bitstream
Existing MPEG-H 3DA Technology
Proposed MPEG-I Technology

Note: The Reference renderer should not be fed by the decoded MPEG-H signals but directly from the common input format.

Alternative figure:



Note: The x,y,z position of the user may also need to be fed back to the encoder/front-end in order to ensure that an appropriate subset, taken from the global set, of audio sources is conveyed to the decoder/renderer to limit transmission bandwidth.

Testing Paradigm: Full Subject Motion and A/V Presentation

The overall goal of a VR/6DoF technology is to create a *convincing illusion* of "the user being there" with a feeling of total *immersion (presence)*. This feeling of immersion is generated by the user's brain if and only if all of the relevant cues reaching it are sufficiently consistent and consistent with the experience of daily reality. This includes the perception of body (and head) self-motion and the consistent low-latency reaction of the A/V presentation. In other words, it requires the action of body/head movement and reaction by the A/V presentation.

In addition to audio cues, the following cues contribute to the feeling of immersion:

- The influence of *head & body movement* are considerable (see the low quality of illusion with untracked / tracked binaural audio and / or video). Similarly, the physical use of hand-held controllers that may be used by the user to position virtual objects within the virtual space by dragging them around is very effective in creating a convincing illusion. In this case, it is the perception of physical position & motion of the arm/hand together with an appropriate rendered A/V response that are powerful supporting cues to the brain
- The influence of the *visual presentation* is considerable (and maybe even more powerful that the audio-only localization cues).
- Attempts to test with reduced body movement will significantly reduce the feeling of user presence or immersion (3D Audio cues presented alone are weak ones, movement cues are strong and visual cues are strong, too).
- Conversely, it is quite possible that visual cues may be so strong that they mask audio defects and artifacts that would be otherwise apparent in an audio only presentation or application (e.g. The McGurk Effect).

It is thus envisioned that ideally users should be permitted to exercise full and unhindered body motion (i.e. head movement, body movement and (x,y,z) translation) when systems under test are presented for subjective evaluation. Obviously, this requires real-time decoding and rendering systems.

Considering these facts, it is recognized that testing with visual presentation and full and unhindered user body movement is desirable in the interests of immersion and VR realism. It is, however, also recognized that certain 6DoF application scenarios may not include full body motion (e.g. sitting in front of a computer and navigating via joystick), and that rendering for a predefined user path may provide more reproducibility. It is also recognized that presentation of a matching visual component will require extra effort. More discussion on these issues are provided in the following output documents.

Output Documents
N17038 - MPEG-I Audio Architecture and Evaluation Procedures for 6DoF
N17060 - Overview of standards activities related to immersive media (v1)

4 Internet of Media Things and Wearables (IoMT)

IoMT use-cases are structured in four main categories, as follows:

- **Smart spaces: Monitoring and control with network of audio-video cameras**
  - Face recognition to evoke sensorial actuations
  - Human tracking with multiple network cameras
  - Networked digital signs for customized advertisement
  - Intelligent firefighting with IP surveillance cameras
  - Automatic video clip generation by detecting event information
  - Self-adaptive quality of experience for multimedia applications
  - Ultra wide viewing video composition
  - Digital signage and second screen use
  - Temporal synchronization of multiple videos for creating 360° or Multiple view video

- **Smart spaces: Navigation**
  - Blind person assistant system
  - Personalized navigation by visual communication
  - Personalized tourist navigation with natural language functionalities
  - Smart identifier: face Recognition on Smart Glasses
  - Smart advertisement: QR code recognition on smart glasses

- **Smart environments in smart cities**
  - Smart factory: Car maintenance assistance A/V system using smart glasses
  - Smart museum: Augmented visit museum using smart glasses
  - Smart house: Light control, vibrating subtitle, olfaction media content consumption
  - Smart car: Head-light adjustment and speed monitoring to provide automatic volume control

- **Smart collaborative health**
  - Increasing patient autonomy by remote control of left-ventricular assisted devices
  - Diabetic coma prevention by monitoring networks of in-body / near body sensors
  - Enhanced physical activity with smart fabrics networks
  - Medical assistance with smart glasses

The global IoMT interface is presented in the Figure 5, which identifies a set of interfaces, protocols and associated media-related information representations related to:

- User commands (setup info.) between a system manager and an MThing, cf. Interface 1.
- User commands (Setup info.) forwarded by an MThing to another MThing, possibly in a modified form (e.g., subset of 1), cf. Interface 1'.
- Sensed data (Raw or processed data) (compressed or semantic extraction) and actuation information, cf. Interface 2.
- Wrapped interface 2 (e.g. for transmission), cf. Interface 2'
- MThing characteristics, discovery, cf. Interface 3.

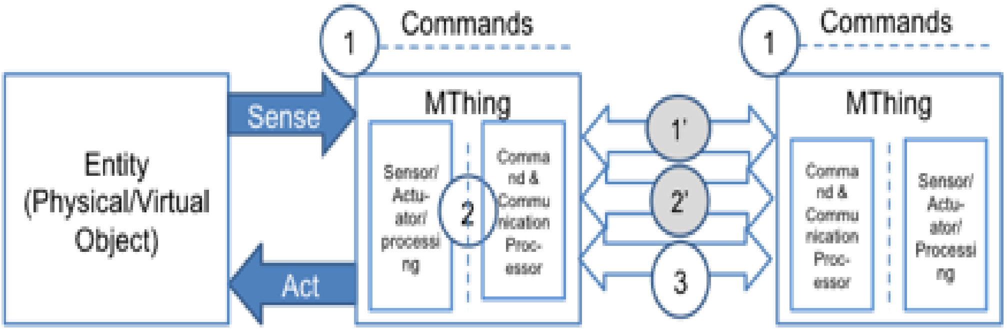

Commands
1
Entity (Physical/Virtual Object)
Sense
Act
MThing
Sensor/ Actu-ator/ processing
2
Command & Communication Processor
1'
2'
3
Commands
1
MThing
Command & Communication Processor
Sensor/ Actu-ator/ Processing

Figure 5 - IoMT Architecture

The following four IoMT core experiments are currently underway:

Hand Gesture Detection and Recognition
In m40389, a metadata schema for the description of general hand gesture has been proposed to support gesture-based wearable applications in the context of IoMT. It is assumed that gesture detection and gesture recognition more likely to be performed in separate PUs due to mainly the limitation of computational power of a PU. Under this assumption, the proposed metadata schema allows the description of detected hand contour and/or trajectory to be delivered from a detection PU to a recognition PU in an interoperability way.

Face Recognition
In m40451, a metadata schema for the description of face features to be used for face recognition in the wearable applications of IoMT. It is assumed that gesture detection and gesture recognition more likely to be performed in separate PUs due to mainly the limitation of computational power of a PU. Under this assumption, the proposed metadata schema allows the description of face features to be delivered from a detection PU to a recognition PU in an interoperability way.

QR Code Recognition
In m40453, a metadata schema for the description of ROI (region of interest) to be used for QR code recognition in the wearable applications of IoMT. It is assumed that gesture detection and gesture recognition more likely to be performed in separate PUs due to mainly the limitation of computational power of a PU. Under this assumption, the proposed metadata schema allows the description of ROI to be delivered from a detection PU to a recognition PU in an interoperability way.

Speech recognition, Speech synthesis, and Question Analysis
In m40473, a metadata schema for the description of speech type and question type to be used for speech recognition, speech synthesis and question analysis has been proposed to support speech/language related wearable applications in the context of IoMT. It is assumed that speech recognition, speech synthesis, question analysis and question answering more likely to be performed in separate PUs due to mainly the limitation of computational power of a PU. Under this assumption, the proposed metadata schema allows the description of speech and question to be delivered from a relevant PUs to the processing PUs in an interoperable way.

Output Documents
N17093 - Core experiment description of IoMT
N17094 - Working draft 2.0 of ISO/IEC 23093-1 IoMT Architecture
N17095 - Working draft 2.0 of ISO/IEC 23093-2 IoMT Discovery and Communication API
N17096 - Working draft 2.0 of ISO/IEC 23093-3 IoMT Media Data Formats and API

5 Media Orchestration

The media orchestration standard provides specification for the orchestration of media and metadata capture, processing and presentation across multiple devices. The functional components of the specification are (i) orchestration of media capture; (ii) orchestration of media presentation; and, (iii) orchestration of processing.

i. Orchestration of media capture is about metadata and control in terms of which device captures what, when and how. What to capture is about device location, orientation and capture capabilities, e.g. zoom capabilities. When to capture is about synchronization with other devices, as well as start and stop of capture. How to capture is about frame rate, resolution, microphone gain, white balance settings as well as codecs used, metadata delivered, and possible processing to be applied.

ii. Orchestration of media presentation is about metadata and control in terms of which device presents what, when and how. What to present is about what media to retrieve and which parts of that media should be presented. When to present is about presentation synchronization with other devices. How to present is about where exactly to play out something (e.g. positioning of a media part in a screen, positioning of an audio object in a room, and possible processing to be applied).

iii. Orchestration of processing is about metadata and control for applying processing to combinations of captured media and/or metadata. This includes single-media processing (e.g. media synchronization in case of transcoding), as well as processing of multiple media and/or metadata together (e.g. performing video stitching, changing arrangements of media in space and time, or automated editing and selection processes).

Furthermore, the specification supports temporal orchestration at both source and sink, extending the DVD CSS specification. The messaging and control is achieved through the DVD-CSS-WC specification. The timed metadata, which cannot be rendered independently and may affect rendering, processing or orchestration of the associated media data is extended from Part 2 and Part 5 of the MPEG-V specification.

Output Documents
N16963 - DoC on ISO/IEC CD 23001-13 Media Orchestration
N16964 - Text of ISO/IEC DIS 23001-13 Media Orchestration
N19953 - TuC on Media Orchestration
N19954 - Thoughts on Media Orchestration Reference Software and Conformance