Text

Description automatically generatedISO/IEC JTC 1/SC 29/WG 03 N01506

**ISO/IEC JTC 1/SC 29/WG 03  
MPEG Systems   
Convenorship: KATS (Korea, Republic of)**

**Document type:** Output Document

**Title:** Technologies under Consideration for ISO/IEC 14496-12

**Status:** Approved

**Date of document:** 2025-05-06

**Source:** ISO/IEC JTC 1/SC 29/WG 03

**No. of pages:** 94 (with cover page)

**Email of Convenor:** young.L @ samsung . com

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg3>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 03 MPEG SYSTEMS**

**ISO/IEC JTC 1/SC 29/WG 03 N01506**

**April 2025, Teleconference**

|  |  |
| --- | --- |
| **Title** | **Technologies under Consideration for ISO/IEC 14496-12** |
| **Source** | **WG 03, MPEG Systems** |
| **Status** | **Approved** |
| **Serial Number** | **25042** |

**Abstract**

The document contains following technologies under consideration for the ISO base media file format (ISO/IEC 14496-12):

[1 Multiplexed timed metadata tracks 3](#_Toc197460402)

[2 On MIME type parameters 9](#_Toc197460403)

[3 Stronger defaulting in Track Runs 10](#_Toc197460404)

[4 Segment Index and Level Assignment 19](#_Toc197460405)

[5 Generic sub-picture track grouping extensions 34](#_Toc197460406)

[6 Integrating new codecs 37](#_Toc197460407)

[7 MovieFragmentHeaderBox update 39](#_Toc197460408)

[8 Sample Run Sample Group 40](#_Toc197460409)

[9 Improvements to movie fragments and dynamic tracks 43](#_Toc197460410)

[10 Undetermined mdat size 55](#_Toc197460411)

[11 MetaBox compatibility with QT 56](#_Toc197460412)

[12 Update to the Segment Index Box 57](#_Toc197460413)

[13 Signaling poisoned data for AI training 57](#_Toc197460414)

[14 Empty Track References 58](#_Toc197460415)

[15 Sample auxiliary information 59](#_Toc197460416)

[16 Allow for a track reference ID to be 0 66](#_Toc197460417)

[17 Generic sample extensions 67](#_Toc197460418)

[18 Compressed MetaBox (MPEG #146, issue #151)) 71](#_Toc197460419)

[19 On handling invalid depth bands (MPEG #150, issue#350)) 72](#_Toc197460420)

[20 On supporting high bit-depth depth maps (MPEG #150, issue#354)) 73](#_Toc197460421)

[21 On Alpha, Depth and Gain map metadata (MPEG #150, issue #355)) 74](#_Toc197460422)

[22 Inventory of technologies to store a sequence of image-encoded frames (MPEG #150, m72634) 75](#_Toc197460423)

[References 78](#_Toc197460424)

[23 On multilayer mapping with ISOBMFF structures (MPEG #150, Issue #357) 78](#_Toc197460425)

[24 On Generic Sample Extensions (MPEG #150, Issue #358 #360) 80](#_Toc197460426)

[25 On cross-track synchronisation (MPEG #150, Issue #359) 92](#_Toc197460427)

# Multiplexed timed metadata tracks

## Basic Design

The basic design is in the 7th edition Amendment 1. However, we have possible extensions on the table; inline associations, sample groups, and structurally-dependent metadata.

## Extensions

## Carrying inline associations

### General

In some circumstances, it can be useful to carry inline definitions. This allows for more flexibility, at the expense of risking that a reader will encounter metadata items it does not recognize and did not expect.

The support is fairly simple; a box in the sample entry to warm that inline carriage may occur, and then a value item box in the sample data that carries both a MetadataKeyBox (the declaration) and a value box.

Tracks without inline keys offer a few advantages:

* A client can determine the entire set of keys that may be present in the track allowing the client to ignore the entire track if no keys of interest to the client are present.
* Space is optimized as keys are carried once in the sample entry and values have only a box header to frame their data and associate them with their key. Inline key/value boxes carry a MetadataKeyBox so if multiple inline keys are present in the same access units, they do not share the key with sibling boxes or with other access units.
* If the sample group optimization technique described earlier is used, the client can determine which samples contain values for particular keys.

Tracks signaling the presence of inline keys offer a few advantages:

* If a new key is required (e.g., for new key, data type, locale, etc. combination), an inline key/value box can be written at any time. This contrasts to the non-inline key case where the set of keys must be known a priori.
* A client does not need to be able to enumerate all combinations of key properties (key, datatype, locale, etc.) to write the track.

With that said, it is recommended that tracks be created without using inline keys. Inline keys can be seen as an optional fallback facility to be used when needed.

A movie may contain a combination of metadata tracks, some using inline keys and some not using inline keys. Where the set of keys can be known a priori, non-inline key- bearing tracks can be used. Where that’s not possible, inline keys can be used.

### Sample entry

The optional MetadataInlineKeysPresentBox indicates if inline key/value boxes might occur in corresponding access units. If MetadataInlineKeysPresentBox is absent or present with a presence indicator of 0, no inline key/value boxes (value boxes with a local\_id of 0xFFFFFFFF) should occur in the access units.

The box is defined as:

aligned(8) class MetadataInlineKeysPresentBox extends Box('keyi') {   
 unsigned int(8) inlineKeyValueBoxesPresent;  
}

inlineKeyValueBoxesPresent is a Boolean integer that should be set to a non-zero value (the value 1 is preferred) if inline key/value boxes are known to be present or might be present in the access units associated with this sample entry.

If MetadataInlineKeysPresentBox is present but inlineKeyValueBoxesPresent is set to 0, access units should be treated as though no MetadataInlineKeysPresentBox is attached to the sample entry. Whether MetadataInlineKeysPresentBox is absent or inlineKeyValueBoxesPresent is set to 0, access unit metadata values having a local id of 0xFFFFFFFF should be ignored.

This approach allows a sample entry to reserve space for and include a MetadataInlineKeysPresentBox but to rewrite just the inlineKeyValueBoxesPresent field to 0 to signal there are no inline key/value boxes present.

If all sample values include inline keys, a MetadataKeyTableBox shall still be present although it may be empty (i.e., it contains no MetadataKeyBoxes). It is also possible to have a combination of some known keys signaled in the MetadataKeyTableBox and some inline key/values signaled with a MetadataInlineKeysPresentBox.

### Sample data item

If the access units associated with the BoxedMetadataSampleEntry contain inline key/value metadata, each inline item is carried in a box with a local\_key\_id of 0xFFFFFFFF and conforming to the type MetadataInlineKeyValueAUBox defined as:

aligned(8) class MetadataInlineKeyValueAUBox extends MetadataAUBox(0xFFFFFFFF) {  
 MetadataKeyBox inline\_key; // local\_key\_id of '1key'   
 MetadataAUBox inline\_value; // local\_key\_id of '1val'  
}

inline\_key is a MetadataKeyBox where local\_key\_id is set to ‘1key’ (for “one key”).

inline\_value is a MetadataAUBox where local\_key\_id is set to ‘1val’ (for “one value”).

The MetadataInlineKeyValueAUBox can be viewed as a MetadataAUBox with two differences:

* It is a container box carrying two boxes, one a MetadataKeyBox holding the key and the other a MetadataAUBox holding the value for the metadata item.
* It has a local\_key\_id (or box type) with the special value 0xFFFFFFFF. All inline key/value boxes share this special local\_key\_id of 0xFFFFFFFF regardless of the contained value’s key.

Because a MetadataInlineKeyValueAUBox carries both the key and the value using that key, this box alone is sufficient to carry what would otherwise require a MetadataAUBox and an associated BoxedMetadataSampleEntry with a MetadataKeyTableBox having the same local\_key\_id as the MetadataAUBox. This allows any non-inline key and associated value to be converted to an inline key/value box. The reverse transform (inline key/value box to non-inline value and sample entry) is possible, too.

While possible, writing a MetadataInlineKeyValueAUBox declaring a key that’s also declared within the MetadataKeyTableBox (i.e., it carries a duplicate MetadataKeyBox) is strongly discouraged. The presence of a MetadataInlineKeysPresentBox signaling the presence of inline keys defeats optimizations that are possible when all available keys are declared within the MetadataKeyTableBox and no inline keys are used. Using inline keys should be reserved for cases when the keys to be written cannot be known at the time the sample entry is constructed. Section xx also discusses the use of inline keys.

## Using sample groups to optimize key searches

### General

This section describes an optional mechanism to optimize searches for metadata track access units containing particular key/value pairs. While this mechanism’s support is not required, it allows a reader to locate access units with values for a key without having to read each access unit in the track and scan for the key’s value. This can be useful for some kinds of metadata (i.e., values that don’t occur in every metadata access unit) but for others (e.g., GPS metadata) it may not add much value. Different tracks in the same movie may choose to use or not to use this optimization.

If inline key/value boxes are used, the mechanism described here can be used to locate those access units with inline key/value boxes. This can be useful in limiting the scan for metadata items with keys only found in inline key/value boxes.

The following subsections describe the details.

### Sample group overview

A metadata track conforming to this specification may optionally make use of the SampleGroupDescriptionBox and SampleToGroupBox constructs to optimize searching for access units containing particular keys or inline keys. This can be characterized as having a ‘key search sample group.’

### Optimizing search with a sample group

In a metadata track containing one or more sample entries, the MetadataKeyTableBox() in the BoxedMetadataSampleEntry can be used to determine possible keys present in the track’s AUs. If a key is not present in the MetadataKeyTableBox(), it is known that the key doesn’t exist in any AUs. It doesn’t however indicate which samples have particular keys (and associated values). Therefore, to determine which metadata keys are present in the track requires an exhaustive search of AUs (associated with that sample entry) in the metadata track.

While it would be possible to create a track with sample entries for each combination of keys present in the track and only associate the samples with that combination with the particular sample entry, having many sample entries may not be ideal or easily done. An alternative (described here) is to define a new kind of sample group that indicates the keys present in each AU.

The new sample group consists of a SampleGroupDescriptionBox holding a new group description for each new combination of keys present in AUs. If all AUs consist of the same four keys, for example, there would be one group description with these four keys. If the set of keys varied, there need only be as many descriptions as there are different sets of keys present in AUs.

A client looking for AUs with a particular key (or keys) would first consult the sample entry (or sample entries if there are more than one) and determine if the key is present in the set of possible keys (via MetadataKeyTableBox()). If this succeeds, the client would check if the optional sample group exists, and finding this to be the case, the client would walk through the SampleToGroupBox checking if the corresponding sample group description contains the key. As these operations require only information present in the MovieBox(), direct reading and processing of AUs is unnecessary.

NOTE: While “key” is used here as being present in the sample group description, an equivalent, more compact identifier is used.

### The keysearch sample group

For this specification, an optional sample group known as a “key search sample group” is defined. It consists of SampleGroupDescriptionBox and SampleToGroupBox having the grouping type ‘keyp’.

The SampleGroupDescriptionBox will contain variable-sized SampleGroupDescriptionEntries, each of type MetadataKeySearchGroupEntry. MetadataKeySearchGroupEntry is defined as:

class MetadataKeySearchGroupEntry() extends SampleGroupDescriptionEntry ('keyp') {  
 unsigned int(32) entry\_count;  
 unsigned int(32) local\_key\_ids\_array[entry\_count];   
}

entry\_count is a 32-bit unsigned integer holding the number local key ids that follow in local\_key\_ids\_array[].

local\_key\_ids\_array is an array of 32-bit integers corresponding to the local\_key\_id field used in the associated MetadataKeyTableBox() and the local key ids used in associated metadata track access units. A value of 0 is reserved and can be used to mark an array entry as absent. A value of 0xFFFFFFFF indicates the associated access units all contain one or more inline key/value boxes.

Each sample group description entry signals the presence of one or more keys from the key table found in the sample entry associated with the sample(s). Access units associated with this sample group description shall have corresponding metadata values with these same keys.

Each key in use is signaled by using the 32-bit integer value of the local\_key\_id field associated with the MetadataKeyTableBox entry. This local key id is also used in access units as the type of Box holding the corresponding value.

If two samples differ in the keys present, they cannot share the same sample group description. A sample group description for each combination should be created.

NOTE 1 While not strictly required, it is recommended that the order of local\_key\_ids be the same as the order of local key ids in the MetadataKeyTableBox of the sample entry. This can be followed by 0xFFFFFFFF if there is an inline key present. This prevents group descriptions with the same set of keys but differing only in key order from creating multiple, trivially different sample group descriptions.

NOTE 2 There is no relationship between the order of keys in the MetadataKeySearchGroupEntry and the order of values for those keys in the associated access unit(s).

A version 0 SampleGroupDescriptionBox should not be used.

Finally, if a sample group spans multiple sample entries with different sets of keys, the local key ids present in the sample entries spanned must be compatible (i.e., the local\_key\_id must be present in each MetadataKeyTableBox and the corresponding key table entry must be the same). An easy way to accomplish this is not to have samples from different sample entries share the same MetadataKeySearchGroupEntry.

## Structurally dependent metadata

### General

If the metadata values in a metadata track are dependent upon another track in a way that a change in the other track may invalidate some or all metadata item values, the dependent metadata items and metadata track itself are termed structurally dependent upon the other track. Structurally dependent metadata tracks are linked to the track upon which they are structurally dependent using a track-reference of type ‘cdep’. The ‘cdep’ track-reference should be used in addition to the ‘cdsc’ track reference because the ‘cdep’ track-reference’s purpose is to indicate tracks needing attention if another track is transformed (e.g., a video track being scaled or cropped during a reencode to a new file where the metadata will also be copied). Tracks having a ‘cdsc’ track-reference without a ‘cdep’ track-reference can be passed through directly so long as there are no other conditions restricting pass-through. Metadata tracks having a ‘cdep’ track-reference may also need to have metadata items transformed or deleted due to the change in the other track (e.g., the video track).

### MetadataStructuralDependencyBox

The interpretation of a metadata value may depend upon a detail of another track. For example, a geometric value such as a region of interest may be interpreted in the coordinate system of a video track. If the other track is transformed (e.g., scaled or cropped in the case of video), the metadata value may become invalid. To signal which metadata items are structurally dependent, a MetadataStructuralDependencyBox() should be present in MetadataKeyBox() for such metadata items.

The metadata track should also have a ‘cdep’ track reference to the other track upon which values are structurally dependent. This reference is used to determine which metadata tracks might need attention if the target track is manipulated.

If the other track is transformed, currently one of three things can occur to the dependent metadata items:

* + If the values can be transformed in response to the change in the other track, metadata values can be read, transformed, and written in their corrected form. This will typically be limited to being performed in a process that reads the source movie and writes a new one. This requires that the code performing this transformation be able to understand the change in the other track and how to transform the dependent metadata values.
  + If the values cannot be transformed, the metadata item can be deleted by removing the MetadataKeyBox() from the MetadataKeyTableBox() (i.e., setting the local id to 0 and optionally removing the MetadataKeyBox()) and optionally removing metadata values from associated access units. Again, this will typically be limited to being performed in a process that reads the source movie and writes a new one. This removal should be avoided if possible but is available for cases where the transform is not understood, the metadata values are not understood, or the change cannot be applied to understood values.
  + Another option is to mark the structurally dependent value as invalid without rewriting access units or deleting the metdata item from the key table. This allows the presence of metadata items known to have become invalid. Clients reading structurally dependent but invalid items may choose to ignore these or do whatever they see fit. They should not treat them as valid.

The MetadataStructuralDependencyBox() is a Box with this definition:

aligned(8) class MetadataStructualDependencyBox extends Box('sdpd') {   
 MetadataStructuralDependencyInfoBox info;  
}

MetadataStructuralDependencyInfoBox is a FullBox having one currently defined flag value. This flag can be used to mark a key table entry as invalid under the structural dependency.

Other children boxes may be introduced in the future.

The MetadataStructuralDependencyInfoBox is a FullBox with this definition:

aligned(8) class MetadataStructualDependencyInfoBox   
 extends FullBox('sdpi', 0, flags) {  
};

flags can have the lowest order bit (i.e., 0x000001) set to indicate the structural dependency is invalid. If this bit is clear, the metadata item and associated values should be considered valid.

Newly written MetadataStructuralDependencyBox() should have this flag in their contained MetadataStructuralDependencyInfoBox() be clear. New flags may be introduced in the future.

NOTE: In the future, other children boxes of MetadataStructuralDependencyBox() may be introduced. The current thinking is that these will help clients understand under which kinds of transforms the values might remain valid. For example, a video scaling where metadata values use a normalized range of 0...1 might not need to be transformed. If the metadata values used pixels, they might however need to be transformed.

# On MIME type parameters

## Introduction

This section discusses issues and solutions for signalling of important video information, as part of the MIME type parameters, for HDR/WCG video, and video with display orientation changes. Hereafter, important video information refers to video information that may be used for content selection, e.g., selection of a video track or a part thereof for consumption.

### Video with display orientation changes

AVC and HEVC both support video content for which the decoder side should apply a transformation of rotation and/or flipping to the cropped decoded picture prior to display, indicated by the display orientation SEI message. Such video is also referred to as video with display orientation changes. Video with display orientation changes need special post-decoding rendering processing to generate desirable viewing experience.

### Signalling of HDR/WCG information in ISOBMFF

The HDR/WCG information can be signalled using the ColourInformationBox defined in clause 12.1.5 of the ISOBMFF specification, for example with the colour\_type equal to 'nclx', in which case the most important HDR/WCG information would be carried in the fields colour\_primaries, transfer\_characteristics, matrix\_coefficients, and full\_range\_flag.

In addition, certain HDR/WCG solutions make use of dynamic metadata conveyed in SEI messages.

### The restricted scheme design in ISOBMFF

The restricted scheme design in ISOBMFF is for handling of situations where the file author requires certain actions on the player or renderer, to enable players to simply inspect a file to find out such requirements for rendering a bitstream and stops legacy players from decoding and rendering files that require further processing. The mechanism applies to any type of video codec.

The mechanism is similar to the content protection transformation where sample entries are hidden behind generic sample entries, 'encv', 'enca', etc., indicating encrypted or encapsulated media. The analogous mechanism for restricted video uses a transformation with the generic sample entry 'resv'. The method may be applied when the content should only be decoded by players that present it correctly.

The restricted scheme is specified in clauses 8.15.1 to 8.15.3 of the ISOBMFF specification.

## Problems and discussions

[Ed. (FM): Issues and solutions in this section were initially introduced as Items 6 (first part) and 7 from m40373 (MPEG #118 meeting)]

The following problems related to the MIME type parameters and signalling of HDR/WCG video, and video with display orientation changes were observed:

1. For video with display orientation changes, a special restricted scheme is missing.
2. There lacks a mechanism to include important video information for HDR/WCG video as part of the MIME type parameters.

## Proposal

The following methods are proposed to solve problems:

1. To solve the first problem, a new restricted scheme type, 'vdoc', is defined, which, when used, indicates that the track carries video with display orientation changes. No further information is provided, and the SchemeInformationBox may be absent in the RestrictedSchemeInfoBox.
2. To solve the second problem, a new optional MIME type parameter 'hdrinfo' is defined to contain the important information of HDR/WCG video. The format of this optional MIME type parameters is a single value or a comma-separated list of values, where each value consists of one or more dot-separated elements. A value of the 'hdrinfo' parameter contains four fields, in the form of "elment1.elment2.elment3.elment4", where the four elements 1 to 4 are the hexadecimal representations of the fields colour\_primaries, transfer\_characteristics, matrix\_coeffs, and full\_range\_flag, respectively, as defined in clause 12.1.5 of the ISOBMFF specification.
3. We note also the approach chosen by the EVC standard as discussed in m70275

The 'codecs' parameter string for the EVC codec is as follows:

<sample entry 4CC>.<key1><value1>.<key2><value2>.….<keyN><valueN>

# Stronger defaulting in Track Runs

## Discussion

Many file writers operate by parsing the high-level syntax of a given input video bitstream and generating the file format metadata from the information of the bitstream. A backward-compatible approach could be achieved by letting advanced clients

* omit the downloading of MovieFragmentBoxes, and
* create the MovieFragmentBoxes in the client side by parsing the high-level syntax of the received MediaDataBoxes.

At the same time, the transmitted track run data is reduced to 0 bytes, i.e. ultimate compression of MovieFragmentBox metadata is achieved.

## Overview

A set of indications is proposed based on which a player can conclude that it is able to process the media data without the MovieFragmentBox. Thus, even though the MovieFragmentBox is available, the player does not need to fetch and process it. Hence, this is a backward compatible approach for avoiding the delivery of MovieFragmentBox.

The following indications are proposed:

* Either of the following indications is included in the ISOBMFF:
  + Indicate with a box flag in a data reference box that all MediaDataBoxes that are referenced through the data reference entry are "tigthly packed", i.e. contain samples for a single track only in decoding order without unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format. See Section 3.4.
  + Indicate with a box flag of the SegmentIndexBox that the media data box(es) carrying the data for the described segment are "tightly packed". See Section 3.6.
* The following indication is included in the NAL unit file format (ISO/IEC 14496-15):
  + Indicate with a box in the sample entry how picture timing can be derived. The box includes a multiplication factor that applies to convert picture order count differences to composition times. For a picture that starts a new coded video sequence, the box indicates the composition time difference from the last picture of the previous coded video sequence. See Section 3.5.
* For usage with DASH, either of the following approaches can be used:
  + Extensions of the SegmentIndexBox indicate the referenced metadata size (mainly MovieFragmentBoxes) or offset(s) to the media data and can be used to conclude the byte ranges of MediaDataBoxes. See alternative approaches in Sections 3.6 and 3.7.
  + The following indications are included in the DASH MPD: Indicate the byte ranges or URLs for requesting the MediaDataBoxes only with a new MPD attribute. See Section 3.8.

A player can operate as follows:

* Conclude from the flag indicating "tightly packed" media data boxes that reception of MovieFragmentBoxes is not necessary.
* Use the indicated byte ranges or URLs for requesting MediaDataBoxes only.
* Generate the MovieFragmentBoxes based on parsing the high-level syntax of the bitstream in the received MediaDataBoxes or directly decodes and plays the bitstream without deriving the file format metadata. In this operation, the information of the TrackRunBox is concluded based on the received MediaDataBox for a movie fragment as follows:
  + By carrying out the access unit boundary determination as specified in AVC or HEVC, the sample sizes and the sample count can be derived.
  + Picture composition timing may be provided in the bitstream (picture timing SEI message) or it is concluded that composition times are proportional to picture order count. Values of sample\_duration are derived accordingly.
  + VCL NAL unit type can be used to determine sample flags or the sample flags can be set to indicate an unknown status.

## Analysis

### Responses to comments at MPEG#127

In the following, we copy the comments from the File Format minutes of MPEG#127 (labeled Cx) and provide our answers to them (Ax).

C1. It seems that in the case that the URLs (e.g. in an MPD) resolve to "pure media data" one would need a new MIME type (not the one for an ISO segment). In that case, is it really in scope for the file format? (See bullet below).

A1. The scheme is primarily intended for on-demand streaming (e.g. ISO base media file format on-demand profile of DASH) where all SegmentIndexBox(es) are placed before any MovieFragmentBox. No changes in the segment formats are proposed and hence no new MIME types are needed either. The media data is selectively fetched using HTTP GET requests with byte ranges that are concluded from the SegmentIndexBox(es) with extensions specified in the proposed scheme.

C2. The 'pain' is not the file format overhead, but the implementation complexity (and edge cases). We need to evaluate this.

A2. We implemented options for providing media data offsets in MP4Box and tested the reader compatibility of the options with several readers. See Section 3.3.2 for details.

C3. This relies on getting somehow getting a segment index (either in-band, e.g. after the moov box, or out of band) at the client:

* because the representation is ftyp-moov-mdat-mdat-mdat…
* because the representation is ftyp-moov-moof+mdat-moof+mdat-moof+mdat… and the sidx tells you the byte-range requests in each segment to omit the moof box

A3. Right, we assume that sidx(es) are placed before any moof, which is required e.g. in ISO base media file format on-demand profile of DASH. Thus, the file structure would be ftyp-moov-sidx(es)-moof+mdat-moof+dat-moof+dat…

### Tests on reader compatibility on extended SegmentIndexBox and SegmentMediaOffsetBox

This section provides results of the reader compatibility tests of the options in the TuC.

Tests were carried out by segmenting a video clip with one AVC media track by differently modified MP4Box programs. Files were then served to players from an HTTP server.

Three different MP4Box versions were used:

* Unmodified MP4Box
* Modified MP4Box which adds SegmentMediaOffsetBox ('smof') after SegmentIndexBox
* Modified MP4Box which sets flags to 1 in SegmentIndexBox and adds media\_data\_offset fields to the end of the SegmentIndexBox.

As seen in Table 1, extended SegmentIndexBox did not introduce any degraded functionality in any of tested players, compared to the unmodified input files.

In browsers, Dash.js reference client relies on browser-side Media Source Extensions (MSE) to handle segment data parsing. Both Firefox and Chromium browsers use MSE implementations which stop segment processing with an error, if they encounter an unknown root-level box. This prevents playback of files which contain SegmentMediaOffsetBox.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Media player | Unmodified | Added 'smof' box | Extended 'sidx' box | Notes |
| Vlc (3.0.8) | OK | OK | OK |  |
| MP4Client (GPAC 0.8.0) | OK | OK | OK | Log messages about extra bytes in 'sidx' box. |
| ffplay (ffmpeg 4.1.3) | OK | OK | OK |  |
| Android Exoplayer (2.10.4) | OK | OK | OK |  |
| Dash.js reference client (3.0.0) running in Firefox browser (69.0.1) | OK | Not playing. Error because of unknown root-level box. | OK |  |
| Dash.js reference client (3.0.0) running in Chromium browser (76.0.3809.100) | OK | Not playing. Error because of unknown root-level box. | OK |  |
| Dash.js reference client (3.0.0) running in Safari browser (13.0.1) | OK | OK | OK |  |

Table : Playback test results

## Flag in data reference box

The following is proposed to be added into clause 8.7.2.1 of ISOBMFF:

When (flags & 0x000002) is greater than 0 in a data reference box, all MediaDataBoxes that are referenced through the data reference box are "tightly packed", i.e.:

* + MediaDataBoxes contain samples for a single track only.
  + The samples are in decoding order within a MediaDataBox.
  + MediaDataBoxes contain no unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format.

## Box in 14496-15

*Add the following clause 4.13:*

**4.13 Indicating composition times directly proportional to picture order counts**

**4.13.1 Definition**

Box Type: 'reti'  
Container: Sample Entry   
Mandatory: No  
Quantity: Zero or one

The presence of this box specifies that the composition time difference of any two consecutive pictures in output order in the same coded video sequence is directionally proportional their picture order count difference.

**4.13.2 Syntax**

class RelativeTimingBox extends FullBox ('reti', version=0, flags=0) {  
 unsigned int(32) poc\_unit\_duration;  
 unsigned int(32) cvs\_start\_interval;  
}

**4.13.3 Semantics**

poc\_unit\_duration specifies the composition time difference that corresponds to a picture order count difference equal to 1.

cvs\_start\_interval specifies the composition time difference of the first picture of each coded video sequence, in output order, relative to the last picture of the previous coded video sequence, in output order.

NOTE: In practice, cvs\_start\_interval is the sample duration of the last picture of each coded video sequence.

## Extension of the segment index box

### Overview

Figure 1 below illustrates new versions of the extended segment index box ‘sidx’. In these new versions of the segment index box, when indexing fragments (i.e. reference\_type=0), two indexes can be stored per fragment (instead of a single one currently: referenced\_size).



Figure : New version of ‘sidx’

As illustrated in the new ‘sidx’ syntax below, the first index is associated with the actual data of the considered fragment while the second index is associated with the metadata of this fragment.

*Add the following at the end of clause 8.16.3.1 (definition of SegmentIndexBox):*

The flags field has the following semantics:

(flags & 1) equal to 1 specifies that the referenced segments are constrained as follows:

* + There is a single referenced MediaDataBox or IdentifiedMediaDataBox for references with reference\_type equal to 0.
  + The referenced MediaDataBox or IdentifiedMediaDataBox contains samples for a single track only.
  + The samples are in decoding order within the referenced MediaDataBox or IdentifiedMediaDataBox.
  + The referenced MediaDataBox or IdentifiedMediaDataBox contains no unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format.

NOTE 1: Since encryption requires extra data to be stored with samples, it is not possible to use (flags & 1) equal to 1 with encrypted media data.

NOTE 2: When (flags & 1) is equal to 1, and media samples are either self-framing or of constant size (indicated by flags & 4), and the sample times are either predictable (indicated by flags & 2) or calculable, it is possible to process the referenced MediaDataBox or IdentifiedMediaDataBox without the MovieFragmentBox. Clause J.2.5 provides background and rationale for using flags and contains an example of a file structure.

(flags & 2) equal to 2 specifies that the referenced segments are constrained so that default\_sample\_duration of TrackExtendsBox applies to each sample and that sample\_composition\_time\_offset is equal to 0 for each sample.

(flags & 4) equal to 4 specifies that the referenced segments are constrained so that default\_sample\_size of TrackExtendsBox applies to each sample.

*Change the syntax of the* SegmentIndexBox *in clause 8.16.3.2 to the following (i.e., replacing* 0 *with* flags *in the box header, and adding the parts conditioned by the value of* flags*):*

aligned(8) class SegmentIndexBox extends FullBox('sidx', version, flags) {  
 unsigned int(32) reference\_ID;  
 unsigned int(32) timescale;  
 if (version==0) {  
 unsigned int(32) earliest\_presentation\_time;  
 unsigned int(32) first\_offset;  
 }  
 else {  
 unsigned int(64) earliest\_presentation\_time;  
 unsigned int(64) first\_offset;  
 }  
 unsigned int(16) reserved = 0;  
 unsigned int(16) reference\_count;  
 for(i=1; i <= reference\_count; i++)  
 {  
 bit (1) reference\_type;  
 unsigned int(31) referenced\_size;  
 unsigned int(32) subsegment\_duration;  
 bit(1) starts\_with\_SAP;  
 unsigned int(3) SAP\_type;  
 unsigned int(28) SAP\_delta\_time;  
 }  
 if (flags & 1)  
 for(i=1; i <= reference\_count; i++)  
 if (reference\_type == 0) // reference\_type of the same i value  
 unsigned int(32) media\_data\_offset;  
}

*Add the following to the end of clause 8.16.3.3:*

media\_data\_offset specifies the offset to the start of the referenced MediaDataBox or the IdentifiedMediaDataBox of a subsegment from the start of the subsegment.

## Alternative improvements to the SegmentIndexBox

### Option 1

To support the different indexing modes, the semantics of reference\_type is extended as follows (highlighted in yellow):

* when set to 1 indicates that the reference is to a SegmentIndexBox; otherwise the reference is to media content as follows:
* when set to 0 indicates content including both metadata and media data (e.g., in the case of files based on this document, to a MovieFragmentBox);
* when set to 2 indicates content including metadata only (e.g., in the case of files based on this document, one or more MovieFragmentBox);
* when set to 3 indicates content including media data only (e.g., in the case of files based on this document, one or more MediaDataBox or IdentifiedMediaDataBox);

if a separate index segment is used, then entries with reference type 1 or 2 are in the index segment, and entries with reference type 0 or 3 are in the media file;

In this option a new version of the segment index box requires two bits for the representation of the reference\_type as illustrated below. The referenced\_size field in the new version is interpreted according to the following values of the reference\_type:

* When set to 0, the referenced\_size is the distance in bytes from the first byte of the referenced index to the first byte of the next referenced index (moof) item.
* When set to 1, the referenced\_size is the distance in bytes from the first byte of the referenced index to the first byte of the next referenced index (sidx) item.
* When set to 2, referenced\_size is the distance in bytes from the first byte of the referenced metadata item to the first byte of the next referenced index metadata item, or in the case of the last entry, the end of the referenced index metadata item.
* When set to 3, referenced\_size is the distance in bytes from the first byte of the referenced data item to the first byte of the next referenced index data item, or in the case of the last entry, the end of the referenced index data item .

The value of subsegment\_duration of each entry with reference\_type equal to 2 or 3 corresponds to the duration of the indexed sub-segment. When the reference\_type is set to 1, the semantics of the subsegment\_duration is the same as in ISOBMFF Table J.3.

### Option 2

#### Syntax

aligned(8) class SegmentIndexBox extends FullBox('sidx', version, flags) {  
 unsigned int(32) reference\_ID;  
 unsigned int(32) timescale;  
 if (version==0 || new\_version) {  
 unsigned int(32) earliest\_presentation\_time;  
 unsigned int(32) first\_offset;  
 } else { //version =1 || new\_version  
 unsigned int(64) earliest\_presentation\_time;  
 unsigned int(64) first\_offset;  
 }  
 unsigned int(16) reserved = 0;  
 unsigned int(16) reference\_count;  
 if (new version)  
 unsigned int(16) subpart\_count;  
 for(i=1; i <= reference\_count; i++) {  
 bit (1) reference\_type;   
 unsigned int(31) referenced\_size;  
 if (new\_version) {  
 for (j=1; j<=subpart\_count;j++) {  
 unsigned int(32) data\_reference\_offset;   
 // may be controlled by a flags value   
 unsigned int(32) referenced\_data\_size;   
 }  
 }  
 unsigned int(32) subsegment\_duration;   
 bit(1) starts\_with\_SAP;  
 unsigned int(3) SAP\_type;  
 unsigned int(28) SAP\_delta\_time;  
 }  
}

#### Semantics (for the new fields, the other remaining unchanged)

Data\_reference\_offset indicates in bytes from where, in a file or in a segment file, the indexed data start. The offset corresponds to the first byte of the file or to the first byte of the considered segment file.

referenced\_data\_size indicates a size in bytes for a contiguous byte range in the data part of the referenced fragment

subpart\_count indicates the number of data blocks (for example tiles) for the current subsegment.

This extended '*sidx*' box can also be combined with '*sidx*' boxes of the current version, for example as in the hierarchical or daisy-chain schemes defined in ISO/IEC 14496-12

## Attributes in DASH MPD

The following is proposed to be added into the SegmentBase element (clause 5.3.9.2.2 of DASH).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | @mediaOnlyRange | O | specifies the byte range that consists only of the media data, such as MediaDataBox, applicable to all Media Segments of the Representation. When used with ISOBMFF Media Segments, the indicated byte range shall start with a box.  The byte range shall be expressed and formatted as a byte-range-spec as defined in RFC 7233, Clause 2.1.. It is restricted to a single expression identifying a contiguous range of bytes. |

The following is proposed to be added into the SegmentList.SegmentURL element (clause 5.3.9.3.2 of DASH).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | @mediaOnlyRange | O | specifies the byte range within the resource identified by the @media that consists only of the media data, such as MediaDataBox. When used with ISOBMFF Media Segments, the indicated byte range shall start with a box.  The byte range shall be expressed and formatted as a byte-range-spec as defined in RFC 7233, Clause 2.1.. It is restricted to a single expression identifying a contiguous range of bytes. |

The following is proposed to be added into the SegmentTemplate element (clause 5.4.9.4.2 of DASH).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | @mediaOnly | O | specifies the template to create the Media Segment List where Media Segments only consist of the media data, such as MediaDataBoxes. |

## Example usage of the segment index box

*Add the following clause J.2.5:*

**J.2.5 Simple one-level indexing of "tightly packed" media**

When the flags field of the SegmentIndexBox is set so that (flags & 1) is equal to 1, the media data is "tigthly packed", i.e. a single MediaDataBox or IdentifiedMediaDataBox contain samples for a single track only in decoding order without unused bytes, sample auxiliary information, metadata, or any other information that does not belong to the sample format. A legacy client simply omits the flags field and the SegmentIndexBox syntax conditional on the values of the flags field. A client taking advantage of the "tightly packed" media could operate as follows:

* Conclude that since the SegmentIndexBox has (flags & 1) equal to 1, the reception of MovieFragmentBoxes is not necessary.
* Omit the downloading of MovieFragmentBoxes and only download the media data by deriving a byte range from the media\_data\_offset given in the SegmentIndexBox. Thus, a bitrate saving in the transmitted data is achieved.
* Create the MovieFragmentBoxes in the client side either by parsing the high-level syntax of the received media data or, when (flags & 2) and (flags & 4) are set, use the default values given in of TrackExtendsBox. For example, in case of AVC or HEVC, the information in the TrackRunBox could be concluded as follows:
  + Deriving the sample sizes and the sample size by carrying out the access unit boundary determination as specified in AVC or HEVC.
  + Deriving sample composition times from picture timing SEI messages present in the bitstream or concluding that composition times are proportional to picture order count.
  + Deriving sample flags from the VCL NAL unit types or setting sample flags to indicate an unknown status.

Figure J.2 shows an example that is aligned with the structure presented in Figure J.1. All entries of the top level SegmentIndexBox point to segments comprising one or more movie fragments, i.e. reference\_type is equal to 0. The values of referenced\_size and subsegment\_duration of each entry are calculated as in Table J.1 above. The dashed double-ended arrows in Figure J.2 indicate the values of media\_data\_offset.



Figure J.. Simple segment index including media\_data\_offset values (dashed arrows).

# Segment Index and Level Assignment

Issues: [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/123*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/123) *and* [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/42*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/42)

See also [*https://github.com/MPEGGroup/FileFormat/issues/12*](https://github.com/MPEGGroup/FileFormat/issues/12)

## Discussion

*https://github.com/MPEGGroup/FileFormat/issues/12*

The proposal is based on the following observations:

1. signaling of IDR byte-range is very similar to the range concepts of ‘ssix’
2. avoid modifying ‘sidx’ box, both for backward compatibility issues (comment from TuC, section 15) and because we index subsegments
3. signaling multiple byte-ranges for the same level in ‘ssix’ seems reasonable (for example, two IDRs in a subsegment)
4. usage of ‘ssix’ with ‘leva’ is not always desirable, especially since:

* level assignment may be dependent on sample group description and sample to group mapping, which is not always available (‘moof’ not yet fetched).
* ‘leva’ requires level to be present in increasing order in the ‘mdat’, which does not allow multiple byte-ranges for a given level.
* ‘leva’ cannot be updated on the fly (present in 'mvex'), all levels to describe have to be known at the start of the session

We therefore would like to introduce a new design of ‘ssix’, fixing the above shortcomings. Moreover, while redesigning ‘ssix’, we also considered the use case of spatial indexing for tile tracks in a file.

## Proposal

The proposal defines a way to use multiple byte ranges per level in ‘ssix’/’leva’, and multiple byte ranges with ‘ssix’ without ‘leva’ through predefined level assignments.

Yellow-highlighted corresponds to text (Part-12) move

Green comes from above proposal

Blue are changes as proposed in TuC.

*In 8.8.13.1 replace*

"Within a fraction, data for each level shall appear contiguously. Data for levels within a fraction shall appear in increasing order of level value. All data in a fraction shall be assigned to levels. "

*with*

"When version 0 of the LevelAssignmentBox is used, within a fraction, data for each level shall appear contiguously, and data for levels shall appear in increasing order of level value. All data in a fraction shall be assigned to levels.

When version 1 or more of the LevelAssignmentBox is used, data for each level need not be stored contiguously and data for levels may be stored in random order of level value. Some data in a fraction may have no level assigned, in which case the level is unknow but is not a level from the levels defined by the LevelAssignmentBox."

*In 8.8.13.1 remove*

“

When padding\_flag is equal to 1 this indicates that a conforming fraction can be formed by concatenating any positive integer number of levels within a fraction and padding the last MediaDataBox by zero bytes up to the full size that is indicated in the header of the last MediaDataBox. The use of padding\_flag is deprecated.

”

*In 8.8.13.2 replace*

aligned(8) class LevelAssignmentBox extends FullBox('leva', 0, 0)

*with*

aligned(8) class LevelAssignmentBox extends FullBox('leva', version, 0)

*In 8.8.13.3 replace*

“padding\_flag equal to 1 indicates that a conforming fraction can be formed by concatenating any positive integer number of levels within a fraction and padding the last MediaDataBox by zero bytes up to the full size that is indicated in the header of the last MediaDataBox. When

padding\_flag is equal to 0 this is not assured.

”

with

“

padding\_flag deprecated, should be set to 0.

”

*Replace 8.16.4.1 with*

“

The SubsegmentIndexBox provides a mapping from levels (as specified by the LevelAssignmentBox or as indicated in the box itself) to byte ranges of the indexed subsegment. In other words, this box provides a compact index for how the data in a subsegment is ordered according to levels into partial subsegments. It enables a client to easily access data for partial subsegments by downloading ranges of data in the subsegment.

~~Each byte in the subsegment shall be explicitly assigned to a level, and hence the range count shall be 2 or greater.~~ ~~If the range is not associated with any information in the level assignment, then any level that is not included in the level assignment may be used.~~

There shall be 0 or 1 SubsegmentIndexBoxes per each SegmentIndexBox that indexes only leaf subsegments, i.e. that only indexes subsegments but no segment indexes. A SubsegmentIndexBox, if any, shall be the next box after the associated SegmentIndexBox. A SubsegmentIndexBox documents the subsegments that are indicated in the immediately preceding SegmentIndexBox.

In general, the media data constructed from the byte ranges is incomplete, i.e. it does not conform to the media format of the entire subsegment.

For leaf subsegments based on this document (i.e. based on movie sample tables and movie fragments):

* ~~Each level shall be assigned to exactly one partial subsegment, i.e. byte ranges for one level shall be contiguous.~~
* ~~Levels of partial subsegments shall be assigned by increasing numbers within a subsegment, i.e., samples of a partial subsegment may depend on any samples of preceding partial subsegments in the same subsegment, but not the other way around. For example, each partial subsegment contains samples having an identical temporal level and partial subsegments appear in increasing temporal level order within the subsegment.~~
* For version 0 of the box, each level shall be assigned to exactly one partial subsegment and in increasing order of level value, i.e. byte ranges for one level shall be contiguous and samples of a partial subsegment may depend on any samples of preceding partial subsegments in the same subsegment, but not the other way around. This implies that all data for a given level require a single range to be retrieved.
* For version 1 or more of the box, multiple byte ranges, possibly discontinuous, with the same level may be described. This implies that all data for a given level may require multiple byte ranges to be retrieved.

//editor's note: the next notes correspond to the previously existing last 2 bullets of the spec but they are informative or repetitions from leva.

Note 1: When a partial subsegment is accessed in this way, for any assignment\_type other than 3, the final MediaDataBox may be incomplete, that is, less data is accessed than the length indication of the MediaDataBox indicates is present. The length of the MediaDataBox may need adjusting, or padding used. ~~The padding\_flag in the LevelAssignmentBox indicates whether this missing data can be replaced by zeros. If not, the sample data for samples assigned to levels that are not accessed is not present, and care should be taken not to attempt to process such samples.~~

Note 2: The data ranges corresponding to partial subsegments include both MovieFragmentBoxes and MediaDataBoxes. The first partial subsegment, i.e. the lowest level, will correspond to a MovieFragmentBox as well as (parts of) MediaDataBox(es), whereas subsequent partial subsegments (higher levels) may correspond to (parts of) MediaDataBox(es) only.

For version 0 of this box, the presence of the LevelAssignmentBox in the movie is mandatory, and the LevelAssignmentBox shall have a version equal to 0.

*Editor's note:* the current v0 spec is unclear, it does not explicitly mandate leva with ssix, maybe we should keep this.

Note 3: assignment\_type equal to 0 (specified in the LevelAssignmentBox) can be used, for example, together with the temporal level sample grouping ('tele') when frames of a video bitstream are temporally ordered within subsegments; assignment\_type equal to 2 can be used, for example, when each view of a multiview video bitstream is contained in a separate track and the track fragments for all the views are contained in a single movie fragment. assignment\_type equal to 3 can be used, for example, when audio and video movie fragments (including the respective MediaDataBoxes) are interleaved. The first level can be specified to contain the audio movie fragments (including the respective MediaDataBoxes), whereas the second level can be specified to contain both audio and video movie fragments (including all MediaDataBoxes).

For version 1 of this box, the presence of the LevelAssignmentBox is only mandatory for level\_assignment\_type 0, in which case the LevelAssignmentBox shall have a version of 1.

“

*Replace 8.16.4.2 with*

1. aligned(8) class SubsegmentIndexBox extends FullBox('ssix', version, flags) {  
   if (version==0) {  
    unsigned int(32)subsegment\_count;  
    for( i=1; i <= subsegment\_count; i++) {  
    unsigned int(32)range\_count;  
    for ( j=1; j <= range\_count; j++) {  
    unsigned int(8) level;   
    unsigned int(24) range\_size;   
    }  
    }  
   } else {  
    unsigned int(16)subsegment\_count;  
    unsigned int(1)lsc;   
    unsigned int(1)incomplete;   
    unsigned int(2)lbs;   
    unsigned int(2)rbs;   
    unsigned int(2)reserved;   
    unsigned int(8)level\_assignment\_type;  
    for( i=1; i <= subsegment\_count; i++) {  
    unsigned int(lsc ? 32 : 16)range\_count;  
    for ( j=1; j <= range\_count; j++) {  
    unsigned int((lbs+1)\*8) level;   
    unsigned int((rbs+1)\*8) range\_size;   
    }  
    }  
   }  
   }

*Replace 8.16.4.3 with*

subsegment\_count is a positive integer specifying the number of subsegments for which partial subsegment information is specified in this box. subsegment\_count shall be equal to reference\_count (i.e., the number of movie fragment references) in the immediately preceding SegmentIndexBox.

lsc if set, indicates that the number of indexed ranges within a partial subsegment is coded on 32 bits, otherwise the number of indexed ranges within a partial subsegment is coded on 16 bits.

incomplete if set, indicates that the last range of a given subsegment may ~~not cover the entire~~ end before the last byte of the subsegment, in which case assignment of remaining bytes to level is unknown but the remaining bytes should not correspond to any level listed in the box.

lbs gives the number of bytes, minus 1, used in coding the level field.

rbs gives the number of bytes, minus 1, used in coding the range field.

level\_assignment\_type gives the associated semantics of the indicated level.

* 0: the level value corresponds to the level indicated in the leva box. If the range is not associated with any information in the level assignment, then any level that is not included in the level assignment may be used. This value shall only be used when the leva box version is 1 or more.
* 1: the level value corresponds to a dependency level.
  + Level 0 indicates that the byte range contains:
    - exactly one or more file-level boxes (e.g. MovieFragmentBox) other than a media data container box (e.g. MediaDataBox or IdentifiedMediaDataBox), and/or
    - zero or at most one box header (8 or 16 bytes) of a media data container box which shall correspond to the last 8 or 16 bytes of the byte range
  + Level 1 indicates same type of data as level 0 but having a dependency on the previous preceding byte range with level L0 (e.g. one single box header (8 or 16 bytes) of a media data container box, the media data container box containing data described by the preceding MovieFragmentBox)
  + Level 2 means the data is independently decodable (SAP 1, 2 or 3). Byte range with level 1 immediately preceding if specified and the first preceding byte range with level 0 are required to process the data.
  + Level N, with N>2, requires data from the preceding byte ranges with lower levels (level N-1 and below) to be processed, stopping at the previous preceding byte range with level 2 if specified, otherwise at the first byte range in the box. Each first byte range with level 0 or 1 preceding any required byte range (level 2 to N) is required to process the data.
* 2: the level value corresponds to a multitrack dependency level. In this mode, lbs shall be 1 or more (i.e., at least 16 bits to code the level). The first 8 bits of the level field give the dependency level value, with the same values and semantics as level\_assignment\_type 1. The remaining less significant bits of the level field give a trackID, which shall identify a track present in the indexed subsegment for level values other than 0 and 1, and shall be 0 if level value is 0 or 1. In this mode, each range with level N>1 consists only of data from the identified track~~, possibly with some meta-data boxes (movie fragments, etc...). The~~ and the level value only gives dependency information within the track. This allows cross-track indexation within a same level.
* 3->0xFF: ISO reserved

range\_count specifies the number of partial subsegment levels into which the media data is grouped. For version 0 of the box, this value shall be greater than or equal to 2 and each byte in the subsegment shall be explicitly assigned to a level. For version 1 or more of this box, this value may be 0 or more, and the described ranges may lead to a size smaller than the subsegment if and only if incomplete is set to 1.

range\_size indicates the size of the partial subsegment. This value shall not be 0, except for the last entry for which the value 0 may be used ~~in the last entry~~ to indicate the remaining bytes of the segment, to the end of the segment.

level specifies the level to which this partial subsegment is assigned.

For level\_assignment\_type 1 or above, byte ranges assigned to levels other than 0 or 1 shall not contain file-level box headers. Typically, the header of a media data container box (e.g. MediaDataBox or IdentifiedMediaDataBox) is in level 0 or 1 while data may be in level 2 or more.

NOTE For level\_assignment\_type 1, since level N depends only from level N-1 and below, a direct mapping of temporal sublayers to levels will not always be possible in case frames from one temporal sublayer depend on preceding frames from the same temporal sublayer in another byte range.

## Examples of use

### Single track indexing

A picture containing shape

Description automatically generated

Figure - level\_assignment\_type=1 with 2 IDRs and no explicit range for moof

A picture containing shape

Description automatically generated

Figure - level\_assignment\_type=1 with 2 IDRs and explicit range for moof

A picture containing text

Description automatically generated

Figure - level\_assignment\_type=1 with low latency DASH segment, 2 chunks and no explicit range for 'moof

### Multi-track indexing

A picture containing text

Description automatically generated

Figure - level\_assignment\_type=2, 2 moof+traf with 1 IDR each and no explicit range for moof

A picture containing shape, rectangle

Description automatically generated

Figure - level\_assignment\_type=2, 1 moof with 2 trafs, 1 IDR/traf and no explicit range for moof

A picture containing rectangle

Description automatically generated

Figure - level\_assignment\_type=2, 1 moof with 2 trafs, 1 IDR/traf and explicit explicit range for moof

## Alternative approach

Discussion <http://mpeg.expert/software/MPEG/Systems/FileFormat/isobmff/-/issues/218>

### Alternative approach

The SubsegmentIndexBox('ssix') provides a mapping from levels (as specified by the LevelAssignmentBox('leva')) to byte ranges of the indexed subsegment. In the current ISOBMFF specification **Error! Reference source not found.**, assignment\_type from the LevelAssignmentBox box indicates the mechanism used to specify the assignment to a level:

* 0 or 1: sample groups are used to specify levels (for value 1: a parameterized sample group (i.e. with grouping type parameter))
* 2 or 3: level assignment is by track
* 4: the respective level contains the samples for a sub-track

In order to interpret the signification of levels when sample groups are used to specify levels, a reader will have to process the MovieFragmentBox('moof') to know whether sample group descriptions are added/modified in the fragment, and to know which samples are associated to which sample group description entry. If sample group description entries are inserted in the fragment, ssix cannot map to such group description entry. Then, Levels for inserted sample group description entries cannot be described with ssix/leva..

The suggested approach in section 9.2 of the ISOBMFF TuC is the possibility to define level assignments within SubsegmentIndexBox('ssix') (with version > 0), so that no dependencies on a static LevelAssignmentBox('leva') are required. The proposed text is quite complete and keeps the same approach of level assignment type. However, if sample B depends on sample A and a file reader needs to know exact level dependencies (i.e. can B be decoded if A is not?), the samples will have to be in separate levels. This gives a potentially very complex hierarchy of samples (almost as many levels as there are samples in one GOP), and may not be sufficient for some use cases (c.f. below).

Additionally, when doing low-latency streaming, we want to push data as fast as possible without having to wait for the complete segment. However, using SubsegmentIndexBox('ssix') mandates SegmentIndexBox('sidx') (to get the number of entries), and the SegmentIndexBox('sidx') mandates a size and a duration per entry, which are not known until the end of the segment. In other words, SegmentIndexBox('sidx')/SubsegmentIndexBox('ssix') is useless in low-latency DASH/HLS.

Finally, when doing broadcast/multicast ABR, the player must decide whether a file requires repair or not. Having levels (using SubsegmentIndexBox('ssix')/LevelAssignmentBox('leva') supposing non low-latency) does help but is insufficient. Indeed, if one level is incomplete, the player still needs to understand the samples impacted by the losses to take a decision (repair loss or drop sample(s)). Some bytes missing in one level can invalidate:

* the entire byte range(s) for that level
* a subset of the samples in that level
* only the last (in decoding order) sample in that level

The repair decision will not be the same in these cases, but cannot be inferred from the SubsegmentIndexBox('ssix') description. For example, samples belonging to the same sample group description entry (hence level) may still be dependent on each other.

### proposal

Rather than establishing a (complex) hierarchy of byte ranges in SubsegmentIndexBox('ssix') which still requires the MovieFragmentBox('moof') to be downloaded to understand the meaning of the desired byte ranges (described by the mapping to sample group), we suggest properly defining sample dependencies within a track or track fragment.

The existing SampleDependencyTypeBox('sdtp') box (and equivalent fragment flags) gives a very high level on the structure of a stream, i.e. is the sample independent (IDR)?, is it depended on (IDR, P, lower Bs)? or are no other samples depending on this one (leaf B)?.

This allows for leaf samples to be simply discarded if needed. However, dependencies of intermediate samples once leaf samples are discarded is unknown:

S1d=0 S2d=1 S3d=1,2 S4d=2,3 S5d=3,4 S6d=1 S7d=1,6 S8d=6,7 S9d=7,8

*Note: SXd=y,z* reads as Sample X with dependency to samples Y and/or Z.

In this example, S5 and S9 are “leaf” samples (not depended on) and can be discarded.

If the player needs to drop other frames in the pattern, it cannot guess from the SampleDependencyTypeBox('sdtp') that S4 and S8 can be discarded as well, once S5 and S9 have been dropped.

Thus, having an explicit list of dependencies allows the file reader to know exactly which samples are impacted when not decoding one sample (by choice or due to losses in the transmission).

This has the benefit of being independent from the coding type and gives a complete description of sample dependencies.

It is important to note that dependency patterns depend on the encoder decisions (e.g. GOP structure), and can vary from GOP to GOP or can be a fixed subset. Fixed subset advocates for usage of sample groups, when dynamic variations would rather go for simple box.

Another benefit is that this proposal may be combined with ssix/leva for on-demand or non low-latency live cases, level assigned to the ‘'sdep' sample group.

### Possible syntax (updated from 65337)

Add a new section 10.13 Sample Dependency Sample Group

**Definition**

Group Types: 'sdep'  
Container: Sample Group Description Box ('sgpd')  
Mandatory: No  
Quantity: Zero or more

This sample group provides explicit coding dependencies of samples towards other samples in the same track or in referenced tracks.

Dependencies are either described by a relative distance from the mapped sample (offset\_from\_reference=0) or from the last previous sample not mapped to the SampleDependencyGroupEntry(offset\_from\_reference=1).

The listed dependencies shall only contain the direct dependencies, i.e. if sample A depends on sample B which in turn depends on sample C, only sample B shall be listed as a dependency to sample A.

The version of the SampleGroupDescriptionBox for the 'sdep' sample group shall be greater than or equal to 1.

**Syntax**

class SampleDependencyGroupEntry extends SampleGroupDescriptionEntry ('sdep')  
{  
 unsigned int(1) offset\_from\_reference;   
 unsigned int(1) has\_inter\_deps;  
 unsigned int(14) num\_dependencies; // intra-track

if (has\_inter\_deps)  
 unsigned int(16) num\_inter\_dependencies; // inter-track  
 else

num\_inter\_dependencies = 0;

if (num\_dependencies < 0x3FFF)

signed int(16) depended\_sample\_num\_diff[num\_dependencies];

for (i=0; i<num\_inter\_dependencies; i++){  
 unsigned int(16)track\_ref\_index;  
 signed int(16)depended\_inter\_sample\_num\_diff;  
 }  
}

**Semantics**

In this subclause, the following terms are defined:

* A previous reference sample of track *A* refers to the previous sample in track *A* with no mapping to the SampleDependencyGroupEntry, meaning for example the previous IDR, BLA or CRA.
* A translated sample number is the sample number of the sample with the same decoding time (as the sample being described) in the referenced track if present, or one plus the sample number of the sample immediately preceding the decoding time.

offset\_from\_reference indicates the base sample (i.e. the sample used as reference for the sample offsets) from which is computed the depended\_sample\_num\_diff and depended\_inter\_sample\_num\_diff values. offset\_from\_reference=0 means values (for depended\_sample\_num\_diff and depended\_inter\_sample\_num\_diff) are relative to the sample number of the sample being described for dependencies in the track (also denoted mapped sample) or to the translated sample number in the referenced track for inter-track dependencies. offset\_from\_reference=1 means values (for depended\_sample\_num\_diff and depended\_inter\_sample\_num\_diff) are relative to the sample number of the previous reference sample in the track or relatively to the translated sample number in the referenced track for inter-tracks dependencies.

has\_inter\_deps indicates if the samples mapped to this entry depend from sample(s) of another track (or other tracks). Value 0 means dependency only from sample(s) of the same track. When has\_inter\_deps=0, num\_inter\_dependencies is inferred to be equal to 0. Value 1 means that samples mapped to this entry depends of sample(s) from another track (or other tracks).

num\_dependencies indicates the number of samples that the described sample depends on. Value 0 means no dependency to any other sample in the track. Value 0x3FFF (i.e. all bits set to 1) means dependencies are unknown, in which case has\_inter\_deps shall be 0. Samples that are not mapped to any sample group entry indicating sample dependencies are considered with no dependencies.

num\_inter\_dependencies indicates the number of samples that the described sample depends on in other tracks. Value 0 means no dependency to any other sample in other tracks.

depended\_sample\_num\_diff indicates the value used to locate a sample’s reference in the same track. If offset\_from\_reference=0, the value indicates the difference between the sample number of the sample being described and the sample depended on, and the value shall be strictly positive (i.e. a value of a value of 2 indicates that sample with number N depends on sample with number N-2). If offset\_from\_reference=1, the value is the difference between the previous reference sample *Pref* in the track of the sample being described and the sample number of the sample depended on, a negative value indicating a sample before *Pref*, a positive value indicating a sample after *Pref* and a value of 0 meaning *Pref*.

track\_ref\_index is the index in the track reference of type 'tdep' providing the track\_ID of the referenced track. Value 1 indicates the first entry. Value 0 is reserved.

depended\_inter\_sample\_num\_diff indicates value used to locate a sample’s reference in the referenced track. If offset\_from\_reference=0, the value indicates the difference between the translated sample number of the sample being mapped to this entry and the sample number of the sample depended on, and the value shall be strictly positive (i.e. a value of 2 indicates that sample with number N depends on sample with number N-2). If offset\_from\_reference=1, the value is the difference between the sample number of the previous reference samplet *PITref* in the referenced track and the sample number of the sample depended on, a negative value indicating a sample before *PITref*, a positive value indicating a sample after *PITref* and a value of 0 meaning *PITref*.

### Impact in TuC

#### Section 9 of TuC:

We suggest to remove all text related to version 1 of SubsegmentIndexBox('ssix'), including figures.

However, we note that current specification does not mandate the presence of the LevelAssignmentBox('leva') when SubsegmentIndexBox('ssix') is used, and files already exist that don’t use the LevelAssignmentBox('leva') with SubsegmentIndexBox('ssix'). But in such case, the meaning of level values is unspecified.

We therefore suggest to make clear in the specification that the LevelAssignmentBox('leva') is optional, and when LevelAssignmentBox('leva') is absent, the following default level assignment is recommended:

* Level 0 indicates that the byte range contains exactly one or more file-level boxes (e.g. MovieFragmentBox) other than a media data container box (e.g. MediaDataBox or IdentifiedMediaDataBox),
* Level 1 indicates that the data is independently decodable (SAP 1, 2 or 3) and may start with a MovieFragmentBox, and only the first preceding byte range with level 0, if present, is required to process the data,
* Level N, with N > 1, indicates other data and requires data from the preceding byte ranges with lower levels (level N-1 and below) to be processed. The last occuring preceding byte range with level 0, if present, and the last occuring preceding byte range with level 1 are required to process a byte range with level N>1.

We also suggest to remove the padding\_flag that is useless in LevelAssignmentBox('leva') box.

We also suggest to specify that non-contiguous byte ranges for a same level may exist when LevelAssignmentBox('leva') is absent.

#### Section 7 of TuC

We also note that section 7 also deals with indexing for stronger defaulted fragments and we believe this has been addressed in a slightly different way in OMAF, maybe this should be removed as well.

## Questions/remarks

It is noted that the proposed 'sdep' sample group is very similar to 'refs' sample group in HEIF. It would be good to investigate the relationship between the two.

It would be interesting to check the behaviors of existing common players when files with ssix and no leva, or levels not matching leva.

## Updated alternative syntax (from m70233)

### Context

Direct sample dependencies in most if not all codecs are usually defined using differential sample count. This idea was exploited in *sdep* approach (m65337/TuC), by coding a number of samples to rewind from (offset\_from\_reference = 0) or by coding the number of samples since the latest reference (offset\_from\_reference = 1), e.g. since the previous IDR, BLA or CRA. However, modern codecs construct the dependency list based on differential sample IDs (where sample IDs are the picture order counts (POCs)) applied to a POC itself computed from a difference with previous frame. The use of relative references (*SampleID\_diff*) makes it possible to:

* reuse the description of the dependency lists (in refs\_diff and sdep) AND reuse the description of the sample identifiers, which could be used to define an entry describing sample dependencies, either an entry in a box or a sample group description entry in a sample group.
* reuse patterns, which could be used to repeat or refer to a previous pattern

The coding of sample references as difference from sample count, as used in *sdep,* prevents from using the first frames as a pattern, while with sample references coded as differences from the current sample ID (corresponding to its POC), an earlier pattern (green) is possible.

Additionally, in many video sequences, it is quite frequent to have frames with varying references in-between reused patterns (e.g. predictive (P) or clean random access (CRA) pictures), and for which no repeating pattern could be computed, e.g. :

*A - PATTERN1(N) - B - C - PATTERN1(N) - D - E.*

This is a typical use case that the compact sample to group 'csgp' cannot properly handle, since each unique entry in the middle of a repeated pattern *P* requires the coding of the pattern *P* each time it is used again, and replicating the sample description indexes.

Tests in m70233 show that a combined approach of differential sample ID coding and optimized pattern representations reduce dependency signaling down to 2.4 bytes per sample on average for fragmented use cases.

### Compact Direct Sample References

#### Definition

Box Type: 'cdrf'  
Container: SampleTableBox or TrackFragmentBox  
Mandatory: No  
Quantity: Zero or one (per container)

CompactDirectSampleReferencesBox provides explicit coding dependencies of samples towards other samples. It associates with each sample of a track or track fragment:

* a sample identifier (sample ID) coded as an absolute sample ID or as a difference with the sample ID of a previous sample;
* a dependency list of sample references coded as a difference with the sample ID.

The listed dependencies should only contain the direct dependencies, i.e. if sample A depends on sample B which in turn depends on sample C, only sample B should be listed as a dependency to sample A.

CompactDirectSampleReferencesBox is structured as a list of entries for reconstructing the list of samples and their sample references, each entry defining either the sample identifier and the dependency list associated with a sample or a reference to a pattern of previously coded sample IDs and sample references in the reconstructed list of samples and their sample references.

The sample ID may be any identifier. It does not need to be unique, and the IDs used by the list of sample references refer to the last sample defined with the given ID in the CompactDirectSampleReferencesBox. A sample ID used in a dependency list but not present in the track or past track fragments indicate a broken dependency, i.e. that the sample cannot be decoded.

NOTE For video track, the sample ID could be the Picture Order Count (POC) of a sample. Broken dependencies typically happen when tuning a stream on a stream access point sample of type 3 (SAP 3 sample), e.g. in an open-GOP case.

The reconstruction process of the list of samples and their sample references shall produce the same result as the following model:

* initializing an empty list flat\_refs
* for each entry in the box,
  + if is\_ref is false, appending to the flat\_refs list an entry containing {nb\_refs, is\_abs, sample\_ID\_code, ref\_diff\_IDs}
  + otherwise (is\_ref is true), for each K ranging from 0 to num\_samples – 1 of the entry, appending to flat\_refs the entry flat\_refs[offset + K%pattern\_length]
* validating that the number of entries in flat\_refs is the same as the number of samples in the track or track fragment
* for each sample J in the track or track run:
  + assigning sampleID as follows:
    - if flat\_refs[J].is\_abs is true, setting the sample sampleID to flat\_refs[J].sample\_ID\_code and
    - otherwise, setting the sample sampleID to flat\_refs[J].sample\_ID\_code + sample[J-1].sampleID
  + assigning sample reference IDs by removing from sample[J].sampleID the value flat\_refs[J].ref\_diff\_IDs, i.e. referenceSampleID = sample[J].sampleID - flat\_refs[J].ref\_diff\_IDs.

#### Syntax

aligned(8) class CompactDirectSampleReferencesBox extends Box('cdrf')

{

unsigned int(8) flags;

if (flags & 2) bits = 32;

else if (flags & 1) bits = 16;

else bits = 8

if (flags & 8) entry\_bits=32;

else if (flags & 4) entry\_bits=16;

else entry\_bits=8

unsigned int(entry\_bits) nb\_entries;

for (i = 0;i < nb\_entries; i++) {

bit(1) is\_ref;

if (is\_ref) {

unsigned int(bits-1) offset;

unsigned int(bits) pattern\_length;

unsigned int(bits) num\_samples;

} else {

unsigned int(bits-1) nb\_refs;

bit(1) is\_abs;

signed int(bits-1) sample\_ID\_code;  
 signed int(bits) ref\_diff\_IDs[nb\_refs];

}

}

}

#### Semantics

nb\_entries indicates the number of entries in the loop for reconstructing the list of references.

is\_ref indicates, if set to 1, that the entry is a reference to a pattern of previously coded sample references in the reconstructed list of references. Otherwise, if set to 0, an explicit list of sample references follows. The first sample in a track or track fragment shall have an associated is\_ref value of 0.

offset indicates the start of the pattern in the reconstructed list of references, value 0 designating the list of sample references of the first sample in the track or track fragment.

pattern\_length indicates the number of samples in the pattern starting from offset

num\_samples indicates the number of samples described. If this value is greater than pattern\_length, the pattern is looped over until all samples indicated by the number of samples num\_samples are described. It is not necessarily the case that num\_samples is a multiple of pattern\_length; the last repeated pattern may be truncated.

nb\_refs indicates the number of direct sample references for this sample. If 0, the sample has no direct references (i.e. the sample is a sync sample).

NOTE The number of direct sample references nb\_refs can represent a number of sample references possibly used for processing the sample, and does not necessarily represent a number of sample references actually used in a codec reference list for processing the sample.

is\_abs indicates, if set to 1, that sample\_ID\_code is the absolute value of the sample identifier. If set to 0, it indicates that sample\_ID\_code is the difference between the sample identifier and the preceding sample identifier in the reconstructed list of references. The first sample in a track or track fragment shall have an associated is\_abs value of 1.

sample\_ID\_code indicates the difference or absolute value of the sample identifier coded as specified by is\_abs flag.

ref\_diff\_IDs indicates the sample identifier of a direct sample reference coded as a difference between the identifier of the sample being described by this entry and the identifier of the referenced sample, i.e. *sampleID - referenceSampleID*.

When present in a SampleTableBox (respectively a TrackFragmentBox), the number of samples described by CompactDirectSampleReferencesBoxshall be equal to the number of samples present in the track (respectively in the track fragment).

# Generic sub-picture track grouping extensions

Issue:[*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/53*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/53)

*And updated proposal in http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/152*

The purpose of this proposal is to allow describing 2D spatial relationship between multiple video bitstreams that relate to a same source content (characterized by a source\_id) (for instance, multiple videos representing subparts of a large panorama). This is currently defined in OMAF, but the concept is generic-enough to apply to ISOBMFF.

## Updated semantics of track\_group\_type

The semantics of track\_group\_type of the TrackReferenceBox is changed from

track\_group\_type indicates the grouping\_type and shall be set to one of the following values, or a value registered, or a value from a derived specification or registration:

'msrc' indicates that this track belongs to a multi-source presentation. Specified in 8.3.4.4.1.

'ster' indicates that this track is either the left or right view of a stereo pair suitable for playback on a stereoscopic display. Specified in 8.3.4.4.2.

The pair of track\_group\_id and track\_group\_type identifies a track group within the file. The tracks that contain a particular TrackGroupTypeBox having the same value of track\_group\_id and track\_group\_type belong to the same track group.

to

track\_group\_type indicates the grouping\_type and shall be set to one of the following values, or a value registered, or a value from a derived specification or registration:

'msrc' indicates that this track belongs to a multi-source presentation. Specified in clause 8.3.4.3 of ISO/IEC 14496-12.

'ster' indicates that this track is either the left or right view of a stereo pair suitable for playback on a stereoscopic display. Specified in clause 8.3.4.4.2 of ISO/IEC 14496-12.

'2dsr' indicates that this track belongs to a group of tracks with two dimensional spatial relationships (e.g. corresponding to spatial parts of a video source). Specified in clause (TBD) of this document.

The pair of track\_group\_id and track\_group\_type identifies a track group within the file. The tracks that contain a particular TrackGroupTypeBox having the same value of track\_group\_id and track\_group\_type belong to the same track group.

Add a new section in 12.1 Video media:

## 12.1.10 Two dimensional spatial relationships

### 12.1.10.1 Definition

A SpatialRelationship2DDescriptionBox TrackGroupTypeBox indicates that this track belongs to a group of tracks with 2D spatial relationships (e.g. corresponding to planar spatial parts of a video source). A SpatialRelationship2DDescriptionBox TrackGroupTypeBox with a given track\_group\_id implicitly defines a coordinate system with an arbitrary origin (0,0) and a maximum size defined by total\_width and total\_height; the x-axis is oriented from left to right and the y-axis from top to bottom. The tracks that have the same value of source\_id within a SpatialRelationship2DDescriptionBox TrackGroupTypeBox are mapped as being originated from the same source and their associated coordinate systems share the same origin (0,0) and the orientation of their axes. For example, a very high resolution video could have been split into sub-picture tracks. Each sub-picture track then conveys its position and sizes in the source video.

Tracks in the same track group shall declare the same source\_id, total\_width, and total\_height. Track groups with different track\_group\_id values and the same source\_id represent the same source content, possibly at different resolutions (i.e. with different values of total\_width or total\_height).

NOTE: A source can be represented by different such track groups (for instance when the same source is available at different resolutions). Each of these track groups is identified by its own identifier track\_group\_id. Since all of these track groups originate from the same source, they share the same source\_id.

There shall be one or more associated SpatialRelationship2DGroupEntry(s) in the associated track (this track possibly has a dynamic size and/or position). When every sample maps to the default indicated in the SampleGroupDescriptionBox in the MovieBox; that SampleGroupDescriptionBox can indicate a default sample group or indicate that all samples are unmapped, depending on its version by setting the static\_group\_description and static\_mapping flags.

### 12.1.10.2 Syntax

aligned(8) class SpatialRelationship2DSourceBox  
 extends FullBox('2dss', 0, 0) {  
 unsigned int(32) total\_width;  
 unsigned int(32) total\_height;  
 unsigned int(32) source\_id;  
}

aligned(8) class SpatialRelationship2DDescriptionBox extends TrackGroupTypeBox('2dsr') {  
 // track\_group\_id is inherited from TrackGroupTypeBox;  
 SpatialRelationship2DSourceBox(); // mandatory, must be first  
 // other optional boxes  
}

### 12.1.10.3 Semantics

total\_width specifies the maximum width in the coordinate system of the SpatialRelationship2DDescriptionBox track group. The value of total\_width shall be the same in all instances of SpatialRelationship2DDescriptionBox with the same value of track\_group\_id.

total\_height specifies the maximum height in the coordinate system of the SpatialRelationship2DDescriptionBox track group. The value of total\_height shall be the same in all instances of SpatialRelationship2DDescriptionBox with the same value of track\_group\_id.

source\_id parameter provides a unique identifier for the source. It implicitly defines a coordinate system associated to this source.

## 12.1.10.4 Spatial relationship 2D sample group

### Definition

The '2dsr' grouping\_type for sample grouping declares the positions and sizes of the samples from a sub-picture track in a spatial relationship track group. Version 1 of the SampleToGroupBox shall be used when grouping\_type is equal to '2dsr'. The value of grouping\_type\_parameter shall be equal to track\_group\_id of the corresponding spatial relationship track group.

### Syntax

class SpatialRelationship2DGroupEntry extends VisualSampleGroupEntry('2dsr') {  
 unsigned int(16) object\_x;  
 unsigned int(16) object\_y;  
 unsigned int(16) object\_width;  
 unsigned int(16) object\_height;   
 /\* Application specific extension here \*/  
 unsigned int(32) app\_specific\_parameters;  
}

### Semantics

object\_x specifies the horizontal position of the top-left corner of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The position value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 0 to total\_width − 1, inclusive, where total\_width is included in the corresponding SpatialRelationship2DDescriptionBox.

object\_y specifies the vertical position of the top-left corner of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The position value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 0 to total\_height − 1, inclusive, where total\_height is included in the corresponding SpatialRelationship2DDescriptionBox.

object\_width specifies the width of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The width value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 1 to total\_width, inclusive.

object\_height specifies the height of the samples in this group within the coordinate system specified by the corresponding spatial relationship track group. The height value is the value prior to applying the implicit resampling caused by the track width and height, if any, in the range of 1 to total\_height, inclusive.

app\_specific\_parameters is a parameter that provides an extension point to define codec specific parameters.

# Integrating new codecs

Issue: [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/95*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/95)

## Introduction

New codecs (audio, video, text, …) are regularly defined within MPEG or externally, and that need to be integrated into ISOBMFF. After 20 years of evolution, the best way to integrate a new codec is not necessarily the same way it was years ago. This section contains a discussion on this topic for the purpose of improving the documentation, for example Annex B.5 of ISOBMFF.

There are 2 high level questions one can ask when integrating a codec into ISOBMFF:

* Given a codec already specified, how to best integrate it into ISOBMFF? What tools should I use for this and that?
* Given a codec under development, how to structure its high level syntax such that it integrates well with ISOBMFF and what are the pitfalls to avoid?

These questions are discussed below.

## Integrating an existing codec

The following questions should be asked when new codecs are integrated in ISOBMFF:

* Which handler should be used? Should a new handler be defined?

Although it is not documented (yet?), the underlying assumptions of a track handler are the following:

* Codecs within a handler are interchangeable, in the sense that once track data is decoded, all codecs should produce the same type of output. In other words, theoretically within a track you can have multiple sample description entries each with a different codec for that handler. (But one should not do that!!!)
* Each handler comes with a base sample entry (vide -> VisualSampleEntry, soun -> AudioSampleEntry, …). If a base sample entry contains the necessary fields for a codec (sample rate or width/height), probably the codec should use the corresponding handler (if the base sample entry is not enough, extend it using the base sample entry as a parent class). Conversely, if there is no corresponding base sample entry (or if there is no such media type yet supported), a new handler should be created.
* How to design the codec-specific Sample Entry ?

**About Decoder initialization.** It has been assumed in the past that whatever is needed for decoder initialization (including opaque sequences of bytes) should be in the sample entry. Doing so should be considered carefully as it leads to the problematic dichotomy: live vs. ondemand (avc1 vs avc3, hvc1 vs. hev1, …). In live cases, all opaque sequences of bytes for the entire session are not necessarily known upfront, and creating a new sample entry on the fly is not (yet?) possible in ISOBMFF. Usually the concern with decoder initialization is initialization latency, but often latency is due to memory allocation which for example for video can be done by knowing width, heigh, and depth. MPEG welcomes feedback on how decoder initialization is done for various codecs. One can use this public issue https://github.com/MPEGGroup/FileFormat/issues/58.

**How to compute “codecs” parameter?** In segmented media, it should be possible to compute the “codecs” parameter without having to fetch any media segment, and only the initialization segment. This should be done primarily based on the information in the sample entry. Sample group information should not be used (because that would remove freedom to put sample group descriptions in the fragments).

**Consider content splicing complexity.** Content splicing (i.e. merging 2 single-track files with the same handler and codec into 1 single-track file) is a typical operation that can be performed with ISOBMFF files. It can always be done by using 2 sample description entries in the output track. However, using multiple sample entries makes implementation more complex (and the “codecs” parameter only describes the first one). A proposed rule is to minimize the amount of data in the sample description in order to maximize the opportunity for single sample description splicing. Processing sample entries when splicing should be as simple as "doing a binary comparison of the 2 input sample entries and if they differ produce 2 output sample entries". Alternatively, having specific merge rules such that “if the fields A, B, C differ between sample entries 1 and 2, create a sample entry with max (A1,A2), max(B1,B2) max(C1,C2)"

A good way to reduce the amount of information in the sample entry is to use **sample groups**. All samples of a track can be easily made part of a common sample group, for example using the default sample group in the ‘sgpd’ box without even using an ’sbgp’ box. This should be considered in particular instead of defining a new sample entry child box type (e.g. colr, pasp, dmix, …).

## Adjusting a codec under development for better integration

In order to get integrated simply and to produce less error-prone files, codecs should have some good properties as discussed here. Designers of high level syntaxes of new codecs should try to follow these recommendations:

* **Enable shallow parsing.** The basic operation of an ISOBMFF packager is to read an elementary stream and produce an ISOBMFF file. Therefore, it should be simple for the packager to identify what will constitute a sample. It should be able to parse the stream (without decoding it fully) and to identify elements that are needed by the packager.
  + Examples of stream types following this good practice are: NALU-based streams, OBU-based streams, TLV-based streams.
  + Example of stream types not following such approach: AAC (non-ADTS) streams.
* **Separate information with different update rates**. Streams typically have information that vary over time, but at different rates. For example, in a video stream, slice-level information will vary frequently and be different for each sample. Picture Parameter Set may vary at each sample but typically varies at GoP granularity or more. Sequence Parameter Sets vary even less, e.g. only a few times in the scope of a sample entry. Profile and level values typically don’t vary within a track. Designers of new codecs should structure data in a way that the packager can easily identify data that it needs and that varies at different rates to store them at the appropriate place (sample entry, sample group, sample). This can mean creating new units (NALU, OBU, Packet) specifically for each type of data.
* **Reduce elementary stream/container file redundancy.** In today’s designs, there is often redundancy between elementary stream information and file level information. For example, width and height are stored in the elementary stream but also in the visual sample entry. The same is true for a lot of information: color (VUI and colr box), HDR static metadata (SEI and ‘mdcv’/’clli’ boxes), sample rate, frame rate, etc. Often the information is replicated at the ISOBMFF level in a codec-agnostic way and to simplify the processing at the client side. But this is a source of error when only one level is modified and not the other one. Approaches to overcome this problem include:
  + Design data units that are codec-agnostic, leveraging CICP or equivalent whenever possible.
  + Consider the possibility of the packager replacing entire data units with ISOBMFF structures when storing in ISOBMFF and the demuxer restoring these data units from the ISOBMFF structures. For example, if a stream had a color (NAL/OB)unit, the packager could avoid storing that unit, replacing it with a ‘colr’ box, and upon reading the demuxer would recreate a (NAL/OB)unit from the ‘colr’, if needed or directly communicate the ‘colr’ box to the decoder/renderer.
* **Ease identification of encryptable data**. Often when encrypting streams, payload headers are kept in the clear. However, it is not always easy to determine where those headers end without having to parse the entire header. Codec designs should consider facilitating this (adding header length, using fixed headers, …).

# MovieFragmentHeaderBox update

Issue : [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/97*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/97)

MPEG is considering defining a new version of the ‘mfhd’ box as follows:

aligned(8) class MovieFragmentHeaderBox

extends FullBox('mfhd', version, flags){

if (version == 0) {

unsigned int(32) sequence\_number;

else if (version == 1) {

unsigned int(64) sequence\_number;

}

}

With the additional semantics:

When version 1 is used, the following ‘flags’ values have the following meaning:

0 : the creator makes no statements, promises, warranties about how sequence\_number is updated

1 : the sequence\_number in this moviefragment is larger than the sequence\_number in the preceding one

3 : the sequence\_number in this moviefragment is one greater than the sequence\_number in the preceding one

# Sample Run Sample Group

Issues :

[*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/98*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/98)

[*https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/286*](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/286)

## Introduction

The overhead of file format metadata, such as the TrackRunBox(es), could be significant especially when it comes to applications using multi-track approach. The overhead of TrackRunBox(es) has been clearly established and thoroughly studied in the context of OMAF tile based streaming [1-3]. For convenience, an example from [1] is copied below.

*“Imagine a 4K video being HEVC encoded with 50 tiles, each tile measuring 384x384 pixels. In such a tile, it is not uncommon for a NAL unit containing a P or B-slice to be as small as 40 bytes, and in areas with little motion sometimes even significantly less. In such a case, the 32-bit sample\_size already results in at least a 10% TrackRunBox overhead, and that’s assuming all of the other optional TrackRunBox fields have been disabled.”*

Conventionally, file writers operate by parsing the high-level syntax of a given input video bitstream and generate the file format metadata from the information of the bitstream. Consequently, under certain constraints, the information present in TrackRunBox(es) could instead be regenerated at the client side by parsing the high-level syntax of the received bitstreams.

More details are provided in Section 8.2, below.

## Discussion

For video bitstreams encapsulated as a track, the information of the TrackRunBox could be tightly packed at a coarser granularity and additional finer-level information can be concluded in the file reader based on the received MediaDataBox for a movie fragment. Figure 19.2.1 shows the conventional encapsulation of TrackRunBox where a encoded segment has m access units, with each access unit having either VCL NAL units or non-VCL units or both. Information about each access unit is encapsulated in a sample of the track as part of the TrackRunBox.

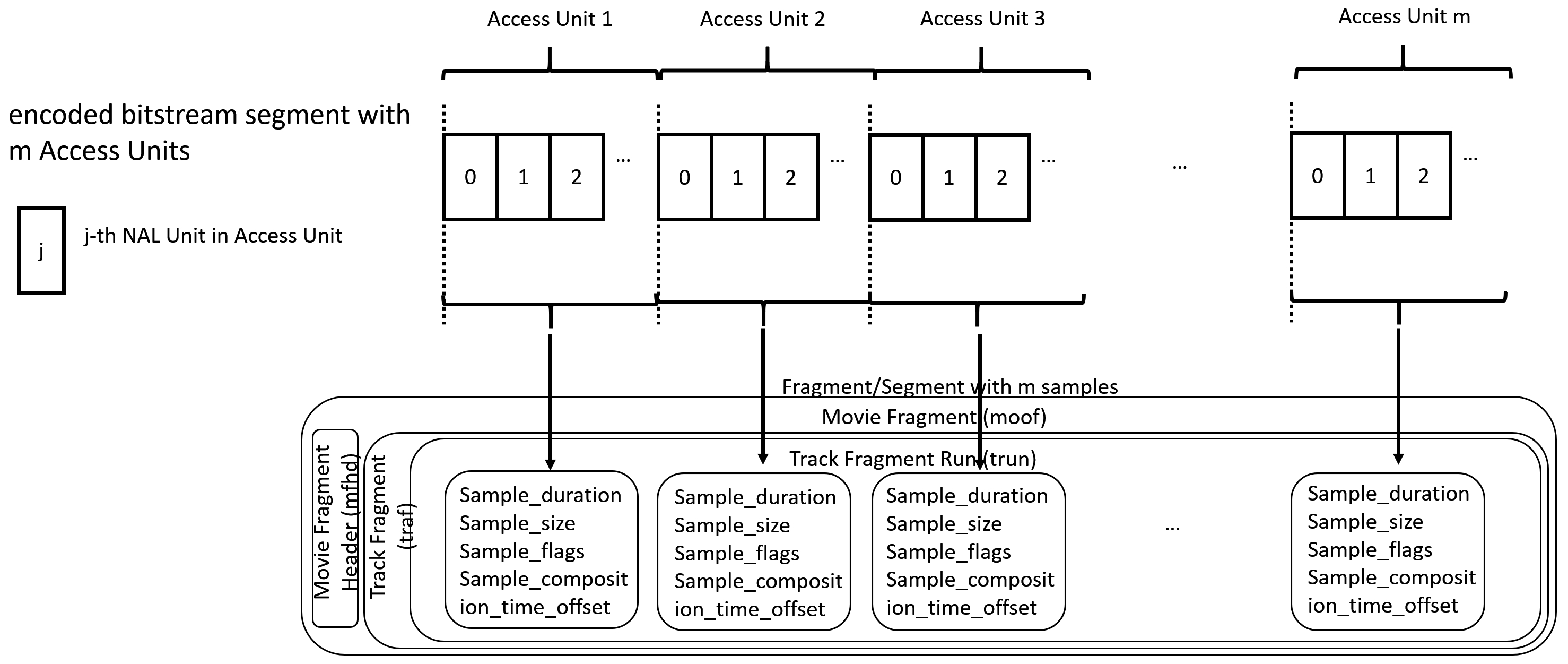


Figure 19.2.1: Conventional sample information in TrackRunBox

As shown in Figure 19.2.2, the sample-level information of the TrackRunBox (m samples) are tightly packed into a single sample information in the SampleRun sample group. Kindly note that the packing of m samples into a single sample in Figure 19.2.2 is only for demonstrating the usage of SampleRun sample group. The SampleRun sample group may contain information about n < m samples where more than one sample of the original TrackRunBox are tightly packed into the SampleRun sample group samples.

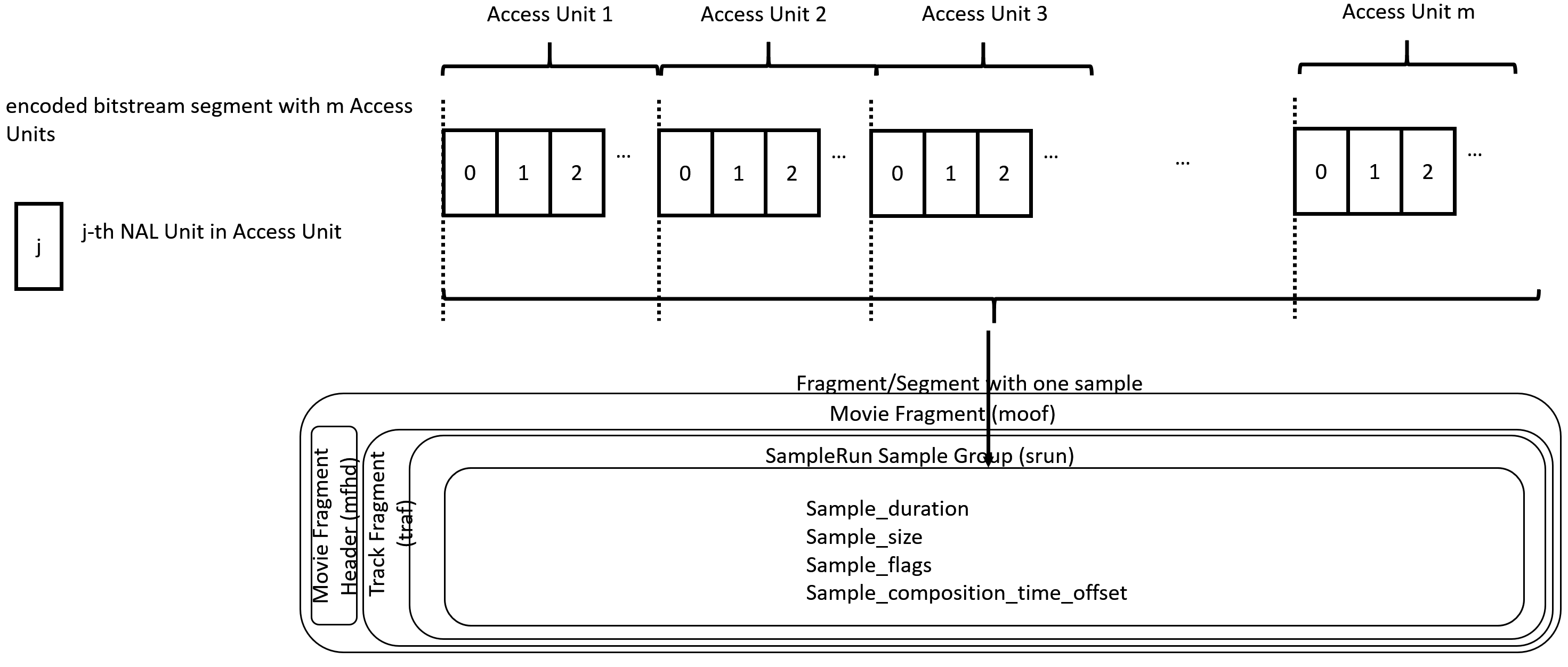


Figure 19.2.2: An example usage of SampleRun Sample group.

An approach for optimizing the delivery of MovieFragmentBox for video could be designed as follows.

* A file writer performs the following to pack a MovieFragmentBox:
  + Identifying GOP structures where pictures have the same pattern of duration, composition time offset and sample flags.
  + Generating a SampleRun sample group description entry for each identified GOP structure.
  + Merging the samples of each GOP into one sample in an 'spkt' transformed video track.
  + Mapping the samples in the 'spkt' transformed video track to the respective SampleRun sample group description entries in SampleToGroupBox(es).
* A player:
  + Performs the access unit boundary determination as specified in AVC, HEVC, VVC, or EVC for each sample in an 'spkt' transformed video track.
  + Generates the original TrackRunBox information in MovieFragmentBoxes based on the signalled SampleRun sample group and the sample sizes determined from the access unit boundaries.

## Proposal

It is proposed to support the new SampleRun sample group as follows.

### Definition

A sample run sample group may be present in a 'spkt' transformed video track. A sample run documents contiguous samples of the untransformed track where each access unit in the mapped sample of the 'spkt' transformed video track is its own sample.

Note: The sample run sample group is limited to be used in a 'spkt' transformed video track containing bitstreams which support picture boundary detection. For example, VVC, which supports signalling of picture header information and the client need not maintain a parsing context to detect picture boundaries.

### Syntax

class SampleRunEntry() extends SampleGroupDescriptionEntry('srun') {  
 unsigned int(24) flags;

if (flags & 0x000008)

unsigned int(32) default\_sample\_duration;

if (flags & 0x000010)

unsigned int(32) default\_sample\_size;

if (flags & 0x000020)

unsigned int(32) default\_sample\_flags;

unsigned int(32) sample\_count;

if (flags & 0x000004)

unsigned int(32) first\_sample\_flags;

{

if (flags & 0x000100)

unsigned int(32) sample\_duration;

if (flags & 0x000200)

unsigned int(32) sample\_size;

if (flags & 0x000400)

unsigned int(32) sample\_flags;

if (flags & 0x000800)

signed int(32) sample\_composition\_time\_offset;

}[ sample\_count ]  
}

### Semantics

flags is a map of flags

The following flags are allowed to be set in the flags:

0x000008 default-sample-duration-present

0x000010 default-sample-size-present

0x000020 default-sample-flags-present

0x000004 first-sample-flags-present; this overrides the default flags for the first sample only. This makes it possible to record a group of frames where the first is a key and the rest are difference frames, without supplying explicit flags for every sample. If this flag and field are used, sample-flags-present shall not be set.

0x000100 sample-duration-present: indicates that each sample has its own duration, otherwise the default is used.

0x000200 sample-size-present: each sample has its own size, otherwise the default is used.

0x000400 sample-flags-present; each sample has its own flags, otherwise the default is used.

0x000800 sample-composition-time-offsets-present; each sample has a composition time offset.

default\_sample\_duration: indicates the default duration of the samples in the sample run

default\_sample\_size: indicates the default size of the samples in the sample run

default\_sample\_flags: indicate the default flags values for the samples in the sample run.

sample\_count the number of samples in this sample run

first\_sample\_flags provides a set of flags for the first sample only of this sample run.

# Improvements to movie fragments and dynamic tracks

The File Format group contained in its mandate the following topic:

1. Study improvements to movie fragments especially when recording e.g. at gateways, and dynamic tracks.

The inputs below are considerations and proposals addressing this topic.

## Dependent Movie Fragments

Reference to discussions: <http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/148>

### Discussion

With low-latency delivery techniques for ISOBMFF, it is quite common that demultiplexers synchronize on the first fragment in a series (e.g., segment) but do not try to demultiplex the following fragments first, typically because the SAP frame is in the first fragment of the series. This means that a lot of information is duplicated in the following fragments due to ISOBMFF rules, but are identical with the information sent in the first segment used to synchronize. This applies to:

* sample group descriptions inserted in movie fragments
* MetaBox, UserDataBox
* Any possible extension of ISOBMFF allowing box injection in movie fragments that do not carry per-sample information

This is because ISOBMFF makes no difference between movie fragments in terms of random access for the parser, while higher delivery protocols impose different constraints on fragments (first fragment of a segment is a SAP1, must have a TFDT, etc…).

For example, a typical ‘seig’ sample group description inserted in each fragment to allow for key rolling will cost, assuming a single entry (one key) is used:

* 61 bytes if constant IV 128bits is used (cbcs)
* 46 for 128 bits IV (cenc)

At 100ms fragment duration, this leads to costs of 3.7 / 4.9 kbps.

In very low latency cases (one sample per fragment):

* for 25fps video : 9.2 / 12.2 kbps
* for AAC at 44100Hz: 16 / 21 kbps.

The signaling rate will obviously increase:

* when adding other sample group descriptions to be updated in the fragments or other boxes such as MetaBoxes
* when more complex sample descriptions are used, e.g. multiple keys per sample.

### Proposal

It is proposes to define a new version in the movie fragment header box:

aligned(8) class MovieFragmentHeaderBox extends FullBox('mfhd', version, 0){   
 unsigned int(32)sequence\_number;   
}

If version is not 0, any SampleGroupDescriptionBox, UserDataBox or MetaBox defined in the last movie fragment, or in the last TrackFragmentBox in the last movie fragment, whose MovieFragmentHeaderBox version is 0 also apply for this movie fragment, and there shall not be any SampleGroupDescriptionBox or MetaBox defined for this movie fragment.

Note that this new version of a movie fragment can still be parsed, in terms of box structure and sample information, independently of the previous fragment, however its interpretation usually re-use information from a previous movie fragment.

### Comments received at MPEG#140

* We already have a new version in the TuC (see Section 7), should we use flags instead?
* Should we pre-declare in the ’mvex’? or ’moov’?
* The use of 'dmof' instead of 'moof' as an option should be studied, so that old readers don't get puzzled when they see something that is erroneous in their opinion.
* we may not always have sample group or encryption for true low latency
* meta and udta should not be in the fragments, they are not mandatory at all therefore rotating keys based on the segments, maybe better to use the manifest for key rotation instead? (but would assume that a manifest is always present).

## On dynamic tracks in fragments

This topic comes from the discussions related to the input contribution m61140:

<http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/147>

The following comments were received at MPEG#140:

- Relying only on brands for parsers to correctly understand the features can be dangerous, and a possible rename of movie fragments boxes should be considered, for example ‘dmof’ instead of reusing ‘moof’.

- Some of the proposed functionalities can be achieved using external signaling such as DASH MPD.

### Discussion

#### Signaling cost

As noted in m58085, the proposed signaling of TrackBox in movie fragments is 4kbps on average for a 1 sec duration segment. Striping the TrackBox from non-essential box (empty sample table, data references) gives around 3.1 kbps for 1s segments.

It is proposed hereafter a further reduction of the track box for inclusion in movie fragments, with an achieved overhead of roughly 1.3 kbps for 1s segments.

#### Change tracking

In use cases such as MPEG-2 TS to fMP4 gateways, the proposed dynamic tracks could be further improved by signaling (un)changed configurations in the fragments, usually called “carousel” in broadcast terminology.

It is noted that sample can be signaled as repeated samples using dependency flags (MPEG-4 systems), but we lack support for such signaling for non-sample data:

* MetaBox
* UserDataBox in track fragment
* Sample Group Description

Obviously relying on the box version is a bad choice:

* Some boxes ('udta') do not have versions
* Other boxes have a version field but already use it (sample group description)
* It is not the intended purpose of box version, which is to indicate variation in the binary syntax and not in the payload.

One approach to deal with this problem is to perform a comparison of past and current boxes, typically through a hashing function. While this allows detecting identical configurations, it has several drawbacks:

* It is costly in client resources
* It does not allow for signaling a repeated configuration with slight variations (in the meta-data for example) that do not require reparsing of the data

If we want dynamic tracks or sample description changes in movie fragments (regardless of the method), it could be useful to introduce some change detection mechanism so we can properly identify repeated information across fragments.

#### Track Removal

In MPEG-2 TS, media streams can be removed due to a PMT update. We noticed that track fragments with duration-is-empty flag set could be used to signal no samples apply to this track. It has however some drawbacks:

* The flag cannot be used if an edit list is present for this track in the movie box,
* It cannot reliably be used as a hint that the track is no longer present; this implies that
  + resource optimization (closing decoders/buffers/etc.) cannot be fully performed.
  + Remixing to e.g. MPEG-2 TS might trigger unused PIDs in the PMT (i.e. PID declared but no packet for this PID), with unpredictable result at the demuxer side
* Moreover, support for this flag is not very good, typically triggering rebuffering in some MSE implementations (because ignored)

If we want a proper/unambiguous signaling of track removal, we need a dedicated box/field to indicated tracks no longer present.

#### File Concatenation, splicing or time-aligned track addition

MPEG-2 TS and some other delivery formats can embed splicing information, allowing a media pipeline to identify temporary service reconfigurations.

While investigating support for such signaling in ISOBMFF along with dynamic tracks, we faced an old but never resolved topic in ISOBMFF: how can we simply build files out of a collection of files, whether as a sequence (‘append a presentation to another one’), or as a set of additional tracks (‘add these tracks, time-aligned’).

In fragmented mode, addition of time-aligned tracks requires editing of the 'moov' to inject the track, and then direct concatenation can work, assuming track IDs do not conflict and timestamps origin is 0 for each “track file”. This obviously will result in one media track being stored after all other ones, but the result is a valid ISOBMFF.

Extending a fragmented presentation with another one is more problematic:

* trackIDs may need to be rewritten: if not the same in each presentation, some tracks may have to be added to the initial 'moov'
* decoder configuration may need to be updated
* timing needs to be rebuilt:
  + alignment of timestamps at boundaries
  + introduction of “gaps” in the timeline to deal with AV sync
* Signaling codec priming (e.g. AAC) is even more complex: a dedicated edit list needs to be created to remove the priming period at the beginning of the second stream.

When concatenating files to splice new content (A1->B->A2), we faced the following issues:

* If the splice is a content replacement,
  + The timing of the injected content B must be rewritten
  + There is no guarantee that the injected content B ends up exactly at the right position, in which case the timing of A2 may need to be rewritten
* If the splice is a content injection (timeline is extended),
  + The timing of the injected content B must be rewritten
  + The timing of A2 must be rewritten

We therefore thought that it would be interesting to design dynamic tracks such that file concatenation is possible without any rewrite. We designed the proposal to allow:

* “early-splicing”, where the source timeline is spliced but contiguous and the only need is to signal the splice points (e.g. MPEG-2 TS to fMP4 gateway)

“late-splicing”, where splicing is done after the fMP4 muxing stage (non-contiguous timelines)

### Proposal

#### Design

We reworked the proposal from m58085 to provide a single method for declaring new tracks and declaring new sample description(s) in an existing track, based on the previous observations.

The proposal defines a DynamicMovieBox in a movie fragment, containing

* Indication on source (timeline/splicing) and configuration changes compared to previous DynamicMovieBox
* Zero or one UserData box, Zero or one Meta box
* Zero or more DynamicTrackBox, each containing
  + a specific DynamicTrackHeaderBox containing
    - track setup (width/height/delay…)
    - indications on modification changes compared to previous DynamicTrackBox
  + some common boxes found in TrackBox.

##### Dynamic movie box

###### Definition

Box Type: 'dymv'  
Container: MovieFragmentBox  
Mandatory: No  
Quantity: Zero or one

A DynamicMovieBox completely or partially overrides the MovieBox setup (track list, user data and meta) for the current fragment.

Each DynamicMovieBox has an associated source\_id, which indicates how the movie fragment extends the initial MovieBox:

* A source\_id equal to 0 indicates that the MovieBox is modified by the movie fragment,
* A source\_id different from 0 indicates that the MovieBox is ignored (i.e. considered not present)

A change of source\_id between two consecutive movie fragments N and N-1 in a single bytes sequence (file, remote resource, etc.) indicates that tracks in N shall be considered as new tracks and tracks in N-1 shall no longer be considered present in the file. In this case, there is no guarantee that the timeline is contiguous between fragments N and N-1. How file readers handle such discontinuities is out of scope of this specification and usually driven by the processing pipeline capabilities. Implementations should however avoid introducing long playback gaps at source\_id change points.

When dynamic tracks are used, the first track fragment of each track in the parent movie fragment shall have a TrackFragmentBaseMediaDecodeTimeBox.

If two consecutive movie fragments N and N-1 have the same value for source\_id, the timeline of all tracks active in both fragments is contiguous, i.e. the constraints on TrackFragmentBaseMediaDecodeTimeBox of each track shall be respected: for a track fragment with the same trackID, the first TrackFragmentBaseMediaDecodeTimeBox in movie fragment N is equal to or greater than the first TrackFragmentBaseMediaDecodeTimeBox in movie fragment N-1 plus the sum of the sample durations in movie fragment N-1.

If several tracks need to be inserted or replaced, these tracks may be declared

* all in a single DynamicMovieBox (i.e. a single movie fragment),
* each one in its own DynamicMovieBox (i.e. one movie fragment per new track) or
* a mix of both approaches.

If a track from the MovieBox is not listed either for update or removal in a DynamicMovieBox with source\_id value of 0, it is valid, but there could be no track fragment for this track in the movie fragment, as is the case with regular movie fragments.

A DynamicMovieBox may contain zero or more DynamicTrackBox.

NOTE Usage of DynamicMovieBox with source\_id different from 0 and zero DynamicTrackBox can be used to force a discontinuity between two movie fragments. Usage of DynamicMovieBox with source\_id equal to 0 and zero DynamicTrackBox can be used to update MetaBox or UserDataBox.

Tracks declared or modified in a DynamicMovieBox may have the same configuration for several consecutive movie fragments. source\_flags allow a file parser to detect that a DynamicMovieBox is a repetition of the previous DynamicMovieBox with the same values for source\_id and bundle\_id. For a same value of source\_id, if multiple dynamic tracks or movie-related metadata (user data, meta) are modified or declared in more than one DynamicMovieBox, then each of these DynamicMovieBoxes shall use a different bundle\_id.

Presence of dynamic tracks in movie fragments shall be indicated using the brand ‘dytk’ in the ExtendedTypeBox or by using the brand ‘isod’ or higher in the FileTypeBox. The MovieBox is not mandatory when using dynamic tracks, but in that case the first movie fragment loaded shall have a FileTypeBox or ExtendedTypeBox indicating support for dynamic track.

EDITOR’S NOTE: we could make FileTypeBox optional in this case but have SegmentTypeBox mandatory.

For DynamicMovieBox the following flags are defined:

* 0x000001 source-info-present if set, indicates that source information is present; if not set, source\_id, bundle\_id and source\_flags take the value 0.
* 0x000002 in-splice if set, indicates that the tracks described in the DynamicMovieBox correspond to a content splice period and will soon move back to previous configuration. By monitoring this flag and the source\_id field, the processing media pipeline can be optimized if desired (e.g. avoid unloading/reloading decoder resources for instance). This flag shall not be set if source\_id is 0.

EDITOR’S NOTE: we could also use a flag in 'styp' to signal this, to simplify edition of files during concatenation

###### Syntax

aligned(8) class DynamicMovieBox extends FullBox('dymv', version=0, flags){   
 if (flags & 1) {  
 unsigned int(32) source\_id;  
 unsigned int(32) bundle\_id;   
 unsigned int(24) source\_flags;  
 } else {  
 source\_id=0;  
 bundle\_id=0;  
 source\_flags=0;  
 }   
 DynamicTrackBox track; // optional: zero or more  
 UserDataBox user\_data; // optional: zero or one   
 MetaBox meta; // optional: zero or one   
}

###### Semantics

source\_id identifies the origin of the fragment. The value 0 indicates that the source is the MovieBox andthe DynamicMovieBox modifies the MovieBox. Other values identify another source than the MovieBox and all tracks, UserDataBox, MetaBox and any other properties defined in the MovieBox shall be ignored.

bundle\_id provides an identifier for tracking partial configuration changes for a given source\_id.

NOTE: The bundle\_id is typically needed when two or more dynamic tracks are inserted in initial movie or during a splice period, each in their own movie fragment. This allows a file reader to detect that the changes advertised for a given bundle have already been processed in a preceding movie fragment.

source\_flags identify the modifications declared in this DynamicMovieBox compared to the previous DynamicMovieBox with the same value of source\_id and bundle\_id. The following flags are defined:

* 0x000001 if set, indicates that one or more tracks configurations have changed
* 0x000002 if set, indicates that the global (MovieBox-level) user data has changed
* 0x000004 if set, indicates that the global (MovieBox-level) meta box has changed
* 0x800000: if set, indicates that the modifications are functionally equivalent to the previous DynamicMovieBox with the same source\_id and bundle\_id. When this flag is set, a file reader may safely skip processing the DynamicMovieBox if a previously parsed DynamicMovieBox has the same source\_id and bundle\_id. Otherwise (this is the first DynamicMovieBox parsed with this source\_id and bundle\_id values), the flag may be set but shall be ignored (i.e. considered as not set) by file readers.

When source\_flags is not set (either explicitly or per the above rule) or has the value 0, the box shall not be skipped. In this case, there is no information regarding modifications of child boxes compared to previous DynamicMovieBox; and the entire content of the box must be re-evaluated.

When source\_flags is not set to 0x000001, any DynamicTrackBox present in this DynamicMovieBox shall have the 0x800000 modification\_flags set.

##### Dynamic track box

###### Definition

Box Type: 'dytk'  
Container: DynamicMovieBox  
Mandatory: No  
Quantity: Zero or more

A DynamicTrackBox declares a new track or modifies an existing track for the duration of the parent movie fragment.

Tracks declared by a DynamicTrackBox, and for which

* the associated source\_id is not 0,
* or there are no tracks with a matching track\_ID in the MovieBox

implicitly declare a TrackExtendsBox with the value default\_sample\_description\_index set to 1 and the values default\_sample\_duration, default\_sample\_size, default\_sample\_flags set to 0.

NOTE This implies that default values will likely need to be set in the TrackFragmentHeaderBox; if multiple track fragments are used for a dynamic track within one movie fragment, the default values may need to be re-coded for each track fragment.

###### Syntax

aligned(8) class DynamicTrackBox extends Box('dytk'){  
 DynamicTrackHeaderBox dyn\_tkhd;//mandatory, must be first  
 SampleDescriptionBox stsd;//conditionally mandatory  
 Box minf\_header\_info; //optional, one of vmhd, smhd, sthd, hmhd…  
 DataInformationBox data\_info; //optional  
 UserDataBox udta; //optional  
 MetaBox meta; //optional  
 TrackReferenceBox tref; //optional  
 TrackGroupBox trgr; //optional  
}

###### Semantics

data\_info if present, gives the source(s) of the samples’ data for this dynamic track. If not present, the samples’ data is present in the container.

minf\_header\_info if present, gives the media-specific header box usually found in the MediaInformationBox of a track with the same handler\_type as this dynamic track. Derived specification may mandate its presence.

Other boxes contained in the DynamicTrackBox (except DynamicTrackHeaderBox) have unchanged semantics. When present, they replace their counterpart boxes in the TrackBox (and children) of the MovieBox.

##### Dynamic track header box

###### Definition

Box Type: 'dtkh'  
Container: DynamicTrackBox  
Mandatory: Yes  
Quantity: One

A DynamicTrackBox can be used to disable an existing track from the MovieBox, override the definition of an existing track from the MovieBox or define a completely new track.

NOTE The DynamicTrackBox is a compaction of TrackHeaderBox, EditListBox and MediaHeaderBox in a single container, in order to keep the track signalling overhead low.

The following flags are defined for a dynamic track header box:

0x000001 dyn\_tk\_ignore\_track indicates, when set, that the track declared in the MovieBox or in a previous DynamicMovieBox with the same source\_id as the parent DynamicMovieBox should be ignored (treated as if not present) until a next movie fragment is received for this track.

Flag values 0x000002, 0x000004, 0x000008, 0x000010, 0x000020 and 0x000040 are used for box parsing. Other values are reserved.

If flag dyn\_tk\_ignore\_track is not set in DynamicTrackHeaderBox, the parent DynamicTrackBox overrides an existing track or declares a new track. In this case:

* If stsd is present in the parent DynamicTrackBox: if there is a track with same ID in the MovieBox, overwrite it with current track, otherwise add the new track to the presentation

NOTE1: derived specification can mandate that the handler type / timescale / width/height remain the same in this case

NOTE2: This is typically used to update a sample description for a track

* If stsd is not present in the parent DynamicTrackBox : there shall be a track with same ID in the MovieBox and all fields in DynamicTrackHeaderBox shall match their counterpart fields in the track/handler/media header boxes declared in the MovieBox. This is used to update UserDataBox, MetaBox, TrackReferenceBox, TrackGroupBox of the track. The sample descriptions of the track remain unchanged in this case.

EDITOR’S NOTE: we could also use a dedicated flag for the case where no 'stsd' is present, in order to avoid re-listing things that shall match what is in the 'moof'’s track

###### Syntax

aligned(8) class DynamicTrackHeaderBox extends FullBox('dtkh', version=0, flags){  
 unsigned int(32) track\_ID;  
 if (!(flags & 1)) {  
 if (flags & 2)   
 unsigned int(24) modification\_flags;  
 unsigned int(32) handler\_type;  
 unsigned int(32) media\_timescale;  
 if (flags & 4)   
 signed int(32) delay;  
 if (flags & 8)   
 unsigned int(24) track\_flags; //as in TrackHeaderBox  
 bit(1) lang\_3cc;  
 if (lang\_3cc) {  
 unsigned int(5)[3] language; // ISO-639-2/T language code  
 } else {  
 bit(7) reserved;  
 utf8string extended\_language

}  
 int(16) alternate\_group;  
 if (flags&16) {  
 unsigned int(32) width;  
 unsigned int(32) height;  
 int(16) layer;  
 if (flags&32) {  
 int(32)[9]matrix ;  
 }  
 }  
 if (flags&64) {  
 int(16)volume;  
 }  
 }  
}

###### Semantics

track\_ID indicates the ID for the track declaration

modification\_flags identify the modifications in the parent DynamicTrackBox compared to the previous DynamicTrackBox with the same value of track\_ID and the same value of source\_id in the parent DynamicMovieBox. These flags may be used by the reader to optimize processing of consecutive track fragments with the same track\_id and source\_id. The following flags are defined:

* 0x000001: the track configuration has changed (changes in one or more fields other than modification\_flags in DynamicTrackHeaderBox, or changes in the associated media header box (‘vmhd’, ‘smhd’, etc.) or changes in the associated data information box)
* 0x000002: the media configuration has changed (new sample description, ‘stsd’)
* 0x000004: the track UserDataBox has changed
* 0x000008: the track MetaBox has changed
* 0x000010: the track TrackReferenceBox has changed
* 0x000020: the track TrackGroupBox has changed
* 0x800000: the track modifications are functionally equivalent to the previous DynamicTrackBox with the same track\_id and source\_id. This flag may be set but shall be ignored (i.e. considered as not set) by file readers when the previous DynamicMovieBox has a different source\_id or when this is the first DynamicTrackBox parsed for this track\_id and source\_id.

When modification\_flags is not set (either explicitly or per the above rule) or has the value 0, there is no information available on possible changes of the track compared to previous DynamicTrackBox with the same track\_id and source\_id and the entire content of the DynamicTrackBox must be re-evaluated.

handler\_type same as handler\_type in HandlerBox

media\_timescale same as timescale in MediaHeaderBox

delay indicates media delay of the track, in media\_timescale. The presentation time of any sample in the track is the sum of the composition time of the sample and this value. A negative presentation time indicates that (part of) the sample data shall not be presented (media skip). If not coded, the value 0 is used.

track\_flags same as flags in TrackHeaderBox. If this field is not coded:

* If source\_id of the dynamic track is 0 and there is a matching track with the same track\_ID in the MovieBox, the flag values from the TrackHeaderBox are used,
* Otherwise, this field value is inferred to be 0x000003

language same as language in MediaHeaderBox

extended\_language same as in ExtendedLanguageBox

alternate\_group same as in TrackHeaderBox

width same as in TrackHeaderBox. If not coded, the media width, after pixel aspect ratio and clean aperture applied, is used

height same as in TrackHeaderBox If not coded, the media height, after pixel aspect ratio and clean aperture applied, is used

layer same as in TrackHeaderBox. If not coded, the layer is 0

matrix same as in TrackHeaderBox. If not coded, the identity matrix is used

volume same as in TrackHeaderBox. If not coded, full volume (1.0) is used

### Examples

### Sample description update

When a sample description changes for fragments N to K, this can be expressed as:

* Fragment N->K :
  + 1 DynamicMovieBox with source\_id=0, (bundle\_id= any), source\_flags=0x800000 (or 0x800001 as a hint).
  + 1 DynamicTrackBox with one stsd
  + DynamicTrackHeaderBox with track\_ID of the updated track and no modification\_flags or
    - modification\_flags=0x800002 if no track layout change
    - modification\_flags=0x800003 if track layout change
* Fragment K+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Track addition

When **a new track appears** for fragments N to K, this can be expressed as:

* Fragment N->K :
  + 1 DynamicMovieBox with source\_id=0 (bundle\_id= any), source\_flags=0x800000 (or 0x800001 as a hint).1 DynamicTrackBox with one stsd
  + DynamicTrackHeaderBox with new track\_ID (different from all previous tracks due to source\_id=0) with
    - modification\_flags=0x800000
* Fragment K+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Splicing AV contained in a single movie fragment

When splicing AV content with T tracks contained in a single movie fragment for fragments N to K, this can be expressed as:

* Fragment N->K :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id>0, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (any track\_ID) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800000
* Fragment K+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Splicing AV contained in separate movie fragments

When splicing AV content contained in separate movie fragments for fragments Ni to Ki, i representing each media, this can be expressed as:

* Fragment Ni->Ki :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id>0, bundle\_id=***i***, source\_flags=0x800000 (or 0x800001 as a hint)
  + 1 DynamicTrackBox (any track\_ID) with one stsd
    - DynamicTrackHeaderBox with modification\_flags=0x800000
* Fragment Ki+1-> :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Splicing with configuration changes in a single movie fragment

When splicing AV content contained in a single movie fragment for fragments N to K with codec configuration change happening from C1 to C2 (N < C1 < C2 < K), this can be expressed as:

* Fragment N->C1 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (any track\_IDs called *splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800000
* Fragment C1 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x000001
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x000002
* Fragment C1+1->C2 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800002
* Fragment C2+1 :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x000001
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags= 0x000002
* Fragment C2+2 -> K :
  + 1 DynamicMovieBox with flags=3 (source-id-present and in-splice), source\_id=1, bundle\_id=0, source\_flags=0x800000 (or 0x800001 as a hint)
  + T DynamicTrackBox (*splice\_track\_IDs*) with one stsd each
    - DynamicTrackHeaderBox with modification\_flags=0x800002
* Fragment K+1 :
  + 0 DynamicMovieBox (if same config from moov can be used)

### Efficiency

With this proposal, signaling a new track in simple cases costs STSD plus:

* 12 bytes for DynamicTrackHeaderBox header (cost of the FullBox)
* 17 bytes for audio (no volume), 27 bytes for video with size/layer and without track matrix (63 with matrix)
* 8 bytes for DynamicTrackBox header (cost of the Box)
* 12 bytes for DynamicMovieBox header

Overhead for 1s fragments is 0.392 kbps for audio, 0.47 kbps for video (0.76 kbps if full visual matrix is specified) hence roughly 1.3 kbps for full signaling (i.e. track and sample descriptions (stsd)) versus 4+ kbps for TrackBox injection as proposed in m58085.

Signaling removal of a track costs 12+4+8+12 bytes, 0.288 kbps

### Usage with dependent movie fragments (defined in section9.1)

The proposed definition for “dependent movie fragments” can be updated when combined with dynamic movie fragment as follows:

If version is not 0, any SampleGroupDescriptionBox, DynamicMovieBox or MetaBox defined in the last movie fragment, or in the last TrackFragment in the last movie fragment, whose MovieFragmentHeaderBox version is 0 also apply for this movie fragment, and there shall not be any SampleGroupDescriptionBox, DynamicMovieBox or MetaBox defined for this movie fragment.

# Undetermined mdat size

Contribution [m65338](https://dms.mpeg.expert/doc_end_user/documents/144_Hannover/wg11/m65338-v1-m65338.zip), gitlab <https://mpeg.expert/software/MPEG/Systems/FileFormat/isobmff/-/issues/245>

Use case: write (capture mode, mdat first) in non-fragmented mode on a non-seekable media.

X Data Termination Indication Box

X.1 Definition

Box Type: 'dtib'  
Container: file  
Mandatory: No  
Quantity: Zero or one

The DataTerminationIndicationBox provides the position of the first byte immediately following the last byte of a Media Data Box with undetermined size. The position is indicated from the end of the file.

The DataTerminationIndicationBox shall only be present when a media data box uses a size of 0. When present, it shall be at the end of the file with no other box following it.

A reader can locate this fixed-size box by offsetting from 12 bytes from the end of file. If these 12 bytes correspond to a DataTerminationIndicationBox, reader can then read the value of the data termination offset. From a byte position corresponding to (*fileSize - data\_termination\_offset -12)*, reader may identify zero or more top-level boxes between this byte position and the DataTerminationIndicationBox.

When this box is found, replacing size of the 0-sized top-level box to the indicated size, without using large-size escape mechanism if the indicated size is greater than 32 bits, shall result in a compliant ISOBMFF file.

X.2 Syntax

class DataTerminationIndicationBox() extends Box('dtib') {  
 unsigned int(32) data\_termination\_offset;  
}

X.3 Semantics

data\_termination\_offset indicates the number of bytes between the last byte of the preceding data box with size=0 and the first byte of this box. It is computed as follows: position of the 1st byte of ‘dtib’ - (1 + position of last byte of media data box)

# MetaBox compatibility with QT

Contribution m65602, gitlab <https://mpeg.expert/software/MPEG/Systems/FileFormat/isobmff/-/issues/257>

Proposal:

At the end of 8.11.1 add the following

NOTE: other specifications based on the box structure defined in this document use a Box instead of a FullBox for ‘meta’. Readers supporting several of these specifications need to be careful when parsing this box.

Consider rewriting the syntax from:

aligned(8) class MetaBox (handler\_type)

extends FullBox('meta', version = 0, 0) {

to

aligned(8) class MetaBox (handler\_type)

extends Box('meta') {

bit(32) reserved = 0;

with a note in semantics:

NOTE: future version of this specification will only use value 0, 2, 4, 6 for the reserved field

# Update to the Segment Index Box

## reference\_count as 32 bits

### Justification

Much of the professional content created today is longer than 18.2 hours, especially archives of CMAF content produced in live OTT setups or used as archiving for television content.

### Proposal

Alter the definition of the segment index box as follows (change highlighted)

|  |
| --- |
| aligned(8) class SegmentIndexBox extends FullBox('sidx', version, 0) {  unsigned int(32) reference\_ID;  unsigned int(32) timescale;  if (version==0) {  unsigned int(32) earliest\_presentation\_time;  unsigned int(32) first\_offset;  }  else {  unsigned int(64) earliest\_presentation\_time;  unsigned int(64) first\_offset;  }  if (version ==0) {  unsigned int(16) reserved = 0;  unsigned int(16) reference\_count  }  else {  unsigned int(32) reference\_count;  }   for(i=1; i <= reference\_count; i++)  {  bit (1) reference\_type;  unsigned int(31) referenced\_size;  unsigned int(32) subsegment\_duration;  bit(1) starts\_with\_SAP;  unsigned int(3) SAP\_type;  unsigned int(28) SAP\_delta\_time;  } } |

# Signaling poisoned data for AI training

## Rationale

AI models and inference engines used in computer vision tasks make use of images/video frames for training. However, in certain scenarios the content provider or the subject in the scene do not wish their content to be part of such AI applications (for example protecting the privacy of a person without blurring the face in photos/videos). One other reason could be copyright infringement, so that the content owners would not want their media to be used for training AI models.

Number of algorithms that allows to create ’poisoned data’ for making an image unusable for AI models, but at the same time viewable and consumable by an end user, is growing.

For example:

* https://towardsdatascience.com/how-nightshade-works-b1ae14ae76c3
* https://nightshade.cs.uchicago.edu/faq.html
* https://arxiv.org/abs/2310.13828

The algorithms can be used by content creators to protect their work from being used by the AI training systems and models. At the same time, it can be assumed that those content creators do not have bad intentions. However, the images with the additional ’poisoned data’ can be harmful for the AI models and significantly reduce their accuracy and performance. Especially when an AI model is used for a mission-critical task, such as detecting pedestrians on the road or detecting certain objects in an area, avoiding AI poisoned media content for AI training is a must.

It is required for an AI expert or system to know that the image/video dataset that he/she uses does not contain such AI-poisoned image/video samples.

## Proposed approach

One way to address the issue is to define a dedicated AIUsabilityBox box describing AI related information.

class AIUsabilityBox extends Box('aiuc') {  
 unsigned int(1) valid\_for\_training;  
 unsigned int(7) reserved;  
}

valid\_for\_training equal to 1 indicates that the item/track can be used for training AI models. When this value is 0, it indicates that an AI-poisoning algorithm might be applied on the item/track which makes it unusable (even harmful) for the AI training process.

Such box could be:

* an item property and can be attached to an image item it describes.
* contain on track level, e.g., as a child box of TrackBox

# Empty Track References

## Question #1: How many elements can be in track\_IDs[]?

The ISOBMFF specification is silent about the minimum number of elements of array track\_IDs[].

track\_IDsis an array of integers providing the track identifiers of the referenced tracks or track\_group\_id values of the referenced track groups. Each value track\_IDs[i], where i is a valid index to the track\_IDs[] array,is an integer that provides a reference from the containing track to the track with track\_ID equal to track\_IDs[i] or to the track group with both track\_group\_id equal to track\_IDs[i] and (flags & 1) of TrackGroupTypeBox equal to 1. When a track\_group\_id value is referenced, the track reference applies to each track of the referenced track group individually unless stated otherwise in the semantics of particular track reference types. The value 0 shall not be present. In the array there shall be no duplicated value; however, a track\_ID may appear in the array and also be a member of one or more track groups for which the track\_group\_IDs appear in the array. This means that in forming the list of tracks, after replacing track\_group\_IDs by the track\_IDs of the tracks in those groups, there might be duplicate track\_IDs. A track\_group\_ID shall not be used when the semantics of the reference requires that the reference be to a single track.

Regarding reference software, it seems the implementation does not specifically expect to have no element in the array. Here is for instance the serialization code of the box:

bytesToRead = s->size - s->bytesRead;

self->trackIDCount = bytesToRead / sizeof(u32);

self->trackIDs = (u32 \*)calloc(1, bytesToRead);

TESTMALLOC(self->trackIDs)

for(i = 0; i < self->trackIDCount; i++)

{

GET32(trackIDs[i]);

}

In the code above, calling calloc with 0 byte and using the return pointer may cause issue since this behaviour is implementation specific.

## Question #2: Value 0 as element of Track\_IDs

As specified in ISOBMFF, the value 0 is forbidden “The value 0 shall not be present”. Therefore, this value could not be used if a file should keep a Track Reference box, for reasons that is use case specific, but not have a specific value to write, for example the exact track to refer to is not known or in a different file.

## Question #3: When tracks are split in different files (e.g. CMAF)

First of all, the CMAF specification is silent on the usage of the Track Reference box (tref). Assuming that an auxiliary track is used in the context of CMAF, it follows that the value in the track IDs array of the Track Reference Type box is unclear, since only one track would be present in the file. If no track IDs is signaled, assuming this is allowed by ISOBMFF and not breaking parsers, the Track Reference box may be kept.

The current assumption is that the question of track references would be handled at higher level than the file, i.e. in the manifest file. At the very least, there should be some guidance on how to transform the track references to CMAF, especially if the Track Reference box should be removed, and if kept, what values to put in the Track\_IDs array.

# Sample auxiliary information

## Generalized approach

The Sample Auxiliary Information (SAI) defined in ISOBMFF provides the ability to store auxiliary information on a per-sample basis. Recently, many use cases have emerged which may use sample auxiliary information boxes(enabling multi-layer bitstream in single layer track, m67816 and carriage of Key-Length-Value (KLV)). The File Format group desires to study and explore the creation of a single, generalized approach to serve all applications utilizing sample auxiliary information.

In m68870, proposes a generic signalling of Sample Auxiliary Information configuration in the ISOBMFF based on the exploration requirements, the proposal in m70219 provides the following clarifications.

* Is the approach static for the file?

Response: the design allows the SAIBox to be present in the SampleTableBox which makes it static for a file. In the current proposal the SAIBox is allowed to be additionally present in the TrackFragmentBox. This should automatically allow for updating the config in a track fragment.

* Does the approach have specific data for the SAI scheme?

Response: SAI specific data is allowed to be present in the SAIDescriptionBox

* The design allows encryption SAI’s. However, if one uses ‘cenc’ for the protection of SAI, it may end up with SAIs for media protection and SAIs for SAI protection, both using the same aux type. How would that work?

Response: The aux\_info\_type\_parameter should be different in both cases, which would help the readers/parsers to differentiate them

* What is the purpose of the sai\_ID field?

Response: This is just an additional identifier for an SAI.

**Proposal**

A track may contain several streams of sample auxiliary information. A new SAIBox ('saib') is defined. The SAIBox may be present in SampleTableBox and TrackFragmentBox. The SAIBox documents information needed to process selected sample auxiliary information data. The proposed SAIBox does not require all the sample auxiliary information in a track to be described, in order to be backward compatible with the currently deployed SAI’s, such as encryption related SAI’s.

8.x.x Sample Auxiliary Information box

8.x.x.1 Definition

Box Type: 'saib'  
Container: SampleTableBox and TrackFragmentBox  
Mandatory: No  
Quantity: Zero or one

The SAIBox provides information about selected sample auxiliary information data. It may optionally occur, but if it does, it shall be used to process the data in the SAI.

8.x.x.2 Syntax

aligned(8) class SAIBox extends Box('saib')  
{  
}

The SAIBox may document information needed to process several sample auxiliary information data. A new SAIDescriptionBox is defined. The SAIDescriptionBox is present in SAIBox. The SAIDescriptionBox documents information needed to process single sample auxiliary information data.

8.x.x Sample Auxiliary Information Description box

8.x.x.1 Definition

Box Type: 'sade'  
Container: SAIBox  
Mandatory: Yes, within SAIBox  
Quantity: At least one

This is a container box for a single sample auxiliary information. The SAIBox consists of one or more SAIDescriptionBoxes. Each SAIDescriptionBox documents information needed to process single sample auxiliary information data. There shall be at least one SAIDescriptionBox within a SAIBox.

8.x.x.2 Syntax

aligned(8) class SAIDescriptionBox extends Box('sade')  
{  
}

A new SAIHeaderBox is defined. The SAIHeaderBox is present in SAIDescriptionBox. The SAIDescriptionBox specifies the characteristics of a single sample auxiliary information.

8.x.x Sample Auxiliary Information Header box

8.x.x.1 Definition

Box Type: 'saih'  
Container: SAIDescriptionBox  
Mandatory: Yes, in SAIDescriptionBox  
Quantity: Exactly one

This box specifies the characteristics of a single sample auxiliary information. Exactly one SAIHeaderBox is contained in a SAIDescriptionBox. The SAIHeaderBox shall be used to identify the format of the data in the SAI. The data in SAI may be protected and a content encoding method may have changed the format of the data in the SAI. If both content encoding and protection are indicated for an SAI, a reader should first un-protect the SAI, and then decode the SAI’s content encoding.

8.x.x.2 Syntax

aligned(8) class SAIHeaderBox extends FullBox('saih', version=0, flags)   
{  
 unsigned int(32) sai\_ID;

unsigned int(32) aux\_info\_type;

unsigned int(32) aux\_info\_type\_parameter;

unsigned int(1) sai\_content\_encoding\_present\_flag;

unsigned int(7) reserved = 0;

if (aux\_info\_type == 'mime') {

utf8string content\_type;

}

else if (aux\_info\_type == 'uri ') {

utf8string encoding\_uri\_type;

}

If(sai\_content\_encoding\_present\_flag) {

utf8string content\_encoding; //optional

}

}

8.x.x.3 Semantics

sai\_ID contains the ID of the SAI for which the following information is defined.

aux\_info\_type is an integer that identifies the type of the sample auxiliary information

aux\_info\_type\_parameter identifies the “stream” of sample auxiliary information having the same value of aux\_info\_type and associated to the same track. The semantics of aux\_info\_type\_parameter is determined by the value of aux\_info\_type.

sai\_content\_encoding\_present\_flag when set to 1 indicates that the sample auxiliary information is encoded with the content\_encoding method. sai\_content\_encoding\_present\_flag when set to 0 indicates that the sample auxiliary information is not encoded.

content\_type is the MIME type of the SAI. If the SAI is content encoded, then the content type refers to the SAI after content decoding.

encoding\_uri\_type is an absolute URI, that is used as a type indicator.

content\_encoding optionally indicates that the SAI is encoded and needs to be decoded before interpreted. The values are as defined for Content-Encoding for HTTP/1.1. Some possible values are “gzip”, “compress” and “deflate”. An empty string indicates no content encoding. Note that the SAI is stored after the content encoding has been applied.

When the sample auxiliary information is protected, allow the ProtectionSchemeInfoBox to be present within SAIDescriptionBox. When ProtectionSchemeInfoBox is present within SAIDescriptionBox, it indicates that the sample auxiliary information data is encrypted by the scheme\_type defined in ProtectionSchemeInfoBox.

If the sample auxiliary information described in SAIDescriptionBox is encrypted, the following sub-box is required to be present: ProtectionSchemeInfoBox. When ProtectionSchemeInfoBox is present within SAIDescriptionBox, it indicates that the sample auxiliary information data is encrypted by the scheme\_type defined in ProtectionSchemeInfoBox.

If the sample auxiliary information described in SAIDescriptionBox needs additional configuration data to decode/process the sample auxiliary information, then corresponding configuration sub-boxes may be present in SAIDescriptionBox.

NOTE CENC SAI uses the aux\_info\_type to indicate CENC and the aux\_info\_type\_parameter to indicate CENC SAI versionning (regular and multi-key for the time being). If both the sample and the SAI containing the layer are protected, it may end up with two CENC SAIs to store, but both have the same aux\_info\_type + aux\_info\_type\_parameter. We continue to explore this further as the proposal targets generic sample auxiliary information signalling and encryption of SAI's is to be supported

## Multi-layer bitstream in single layer track SAI

The file format group desired to have solutions to carry multi-layer bitstream in a single-layer track. The m70219 proposal utilizes the sample auxiliary information to carry multi-layer bitstream in a single-layer track, which is also backward compatible with single layer track 4cc.

**Proposal**

**4.xx Enhancement layer storage as sample auxiliary information**

**4.xx.1 General**

Subclause 4.xx specifies the storage of enhancement layer as sample auxiliary information. Consequently, a single-layer track may store enhancement layer(s) of the multi-layer bitstream in sample auxiliary information of the track.

**4.xx.2 Sample auxiliary information type**

The samples of multi-layer bitstream shall be stored in sample auxiliary information of the track using the SAIBox. The SAIBox shall include the information needed to process the multi-layer bitstream.

The ConfigurationBox which includes the required decoder configuration record of the multi-layer bitstream shall be contained in the SAIDescriptionBox ('sade') indicating that the sample auxiliary information contains multi-layer bitstream.

The aux\_info\_type in the SAIHeaderBoxBox is set equal to a sample entry type of the multi-layer bitstream, the sample auxiliary information is formatted according to the sample format indicated by the sample entry type and aux\_info\_type\_parameter shall be equal to 0 unless specified otherwise for the sample entry type.

If the samples of multi-layer bitstream is encrypted, the SAIDescriptionBox shall contain the ProtectionSchemeInfoBox which documents the encryption scheme used for the samples of multi-layer bitstream.

## Extension

ISOBMFF includes an ability to attach sample auxiliary information (SAI) to each sample in a track. SAI is a generalized capability, enabling the attachment of generic forms of information to samples in a track. Shortcomings of the current approach addressed in this contribution include:

1. A limit of 255 bytes of SAI for each sample.
2. No defined mechanism for specifically signaling an encoding method for the SAI. The default method for using the aux\_info\_type fourCC, enables signaling a type, but not a specific implementation instance of certain common encoding methods, such as SMPTE key-length-value.
3. No defined mechanism for including static information related to a specific set of SAI.

An example of encryption was discussed in a prior contribution as a justification for additional space. Other potential use cases involve pixel level metadata, where SAI can exceed the size of an uncompressed image. As a result, options for 16-bit and 32-bit size information are included in this proposal. Finally, as per the prior proposal, best practices for syntax definitions related to versioning are included in this proposal.

**Proposal**

aligned(8) class SampleAuxiliaryInformationSizesBox extends FullBox('saiz', version, flags)

{

if (flags & 1) {

unsigned int(32) aux\_info\_type;

unsigned int(32) aux\_info\_type\_parameter;

}

unsigned int sample\_size\_bits=0;

if (version==2) {

sample\_size\_bits = 32;

} else if (version==1) {

sample\_size\_bits = 16;

} else if (version==0) {

sample\_size\_bits = 8;

}

unsigned int(sample\_size\_bits ) default\_sample\_info\_size;

unsigned int(32) sample\_count;

if (default\_sample\_info\_size == 0) {

unsigned int(sample\_size\_bits) sample\_info\_size[sample\_count];

}

}

No changes on semantics

8.7.10 **Sample auxiliary information configuration box**

8.7.10.1 **Definition**

BoxType: 'saiC'  
Container: SampleTableBox or TrackFragmentBox  
Mandatory: No  
Quantity: Zero or More

The SampleAuxiliaryInformationConfigBox allows associating private information to all per-sample sample auxiliary information for a given tuple {aux\_info\_type, aux\_info\_type\_parameter}.

It allows for defining an encoding type, which when enabled, applies to the sample auxiliary information and any associated private information.

When SampleAuxiliaryInformationConfigBox is present in a track fragment, the private information applies to all sample auxiliary information for the identified type in that fragment; otherwise (not present in a fragment or file not fragmented), the private information of the SampleAuxiliaryInformationConfigBox present in the track SampleTableBox applies to the sample of the track fragment or of the track.

8.7.10.2 **Syntax**

aligned(8) class SampleAuxiliaryInformationConfigBox extends FullBox('saiC', version=0, flags=0)

{

unsigned int(32) aux\_info\_type;

unsigned int(32) aux\_info\_type\_parameter;

if ( flags & 1){

unsigned int(32) aux\_info\_enc\_type;

if (aux\_info\_enc\_type == 'mime') {

utf8string content\_type;

}

else if (aux\_info\_enc\_type == 'uri ') {

utf8string encoding\_uri\_type;

}

utf8string content\_encoding; //optional

}

bit(8) data[]; //till end of box

}

8.7.10.3 Semantics

(flags & 1) equal to 1 specifies the sample auxiliary information uses a ‘mime’ or ‘uri ‘ defined encoding method. (flags & 1) equal to 0 indicates the encoding of the sample auxiliary information is determined by association with the aux\_info\_type and aux\_info\_type\_parameter values.

aux\_info\_type and aux\_info\_type\_parameter are defined as in SampleAuxiliaryInformationSizesBox

aux\_info\_enc\_type is a 32-bit fourCC value defining a valid sample type indicator, either 'mime' or ‘uri ‘.

content\_type is the MIME type for the SAI. If the samples are content encoded (see below), then the content type refers to the SAI after content decoding.

encoding\_uri\_type is an absolute URI, indicating the method for encoding the SAI. If the samples are content encoded (see below), then the encoding URI type refers to the SAI after content decoding.

content\_encoding optionally indicates the SAI is encoded and needs to be decoded before interpreted. The values are as defined for Content-Encoding for HTTP/1.1. Some possible values are “gzip”, “compress” and “deflate”. An empty string indicates no content encoding. Note the SAI is stored after the content encoding has been applied.

data contains private information for sample auxiliary information associated with a specific aux\_info\_type and aux\_info\_type\_parameter.

## Configuration

ISOBMFF includes the ability to carry sample auxiliary information for each sample in a track. The existing method for signaling the type of information is based on two unregistered 4CC codes (aux\_info\_type and aux\_info\_type\_parameter). With a growing interest in using sample auxiliary information for a variety of use cases, it is recognized a more robust approach is needed to achieve desired levels of functionality and interoperability. The two issues listed above (m67816 and m67971) propose different methods for different use cases related to this topic. It is desired to study and explore the creation of a single, generalized approach for the configuration of sample auxiliary information to serve all use cases and applications utilizing SAI.

Related Issues:

1. **m67816 [NALuFF][TuC] Update to multiple layers in single-layer track design**

<https://git.mpeg.expert/MPEG/Systems/FileFormat/NALuFF/-/issues/194>

1. **m67971 [ISOBMFF] On methods for encoding sample auxiliary information**

<https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/289>

**Use cases**

1. Enabling multiple layers in a single-layer track in NALuFF.
2. Configuration and carriage of Key-Length-Value (KLV) encoded SAI (and other defined forms of encoding metadata)
3. Carriage of 2D grid metadata, mapped to the sample data. For instance, 1 or 8-bit cloud cover data in a geospatial image sequences, or pixel level metadata.
4. Carriage of numerically lossless compressed SAI.

**Explorations**

1. Explore/define a uniform solution for signaling configuration information (including encoding methods and layer configuration) for sample auxiliary information. Methods shall include registered methods, such as through a 4CC, as well as general methods based on MIME and URI techniques, such as found in the Metabox.
2. Explore/define methods for carrying and associating static information to all per-sample SAI for a given tuple {aux\_info\_type, aux\_info\_type\_parameter}.
3. Explore/define the location to store SAI configuration information
4. Explore/define the structure for how SAI configuration is to be carried
5. Explore/define options to handle configuration data associated with the SAI which is not static. For instance, using sample groups to update the configuration data.
6. Explore/define options for encrypting SAI payloads
7. Explore/define options for configuring the compression of SAI payloads (see 23001-17 Amd2)
8. Explore/define how to signal the presence of additional layers in the MIME type

# Allow for a track reference ID to be 0

## Original proposal

If an event-based media track is expected not to be synchronized with other tracks and to be ignored by the legacy players, part 14 of ISO/IEC 14496 (MP4FF) specifies to signal a sync track reference with the value 0 (as described in the contribution m62599). This definition could be reused for the above purpose and should be defined in ISO base media file format.

## References / Citations

14496-14, section 4.3:

If a track (stream) contains a track reference of type 'sync' whose value is 0…

This conflicts with the base spec that states in section 8.3.3.3:

The value 0 shall not be present.

## Consideration

We should fix the base text to state that the value 0 shall not be present except if it has a well-defined meaning for the indicated reference type.

# Generic sample extensions

## Overview

ISOBMFF has always lacked a way to store media samples as non-contiguous pieces of data within a single track, and this has usually been dealt with using multiple tracks (cf tiling, extractors, …). However, the multi-track approach is heavy, introduces complex track reference methods and implies many redundant information such as timing and composition offsets. There are many use cases where being able to store a single track while splitting the samples would be beneficial:

* Scalable layers in a single track
* Multi-codec support in a single track (AVC+LHVC, LCEVC, …)
* Tiled bitstreams

A non-contiguous splitting would allow a file reader to fetch only the sub-parts it requires for a given time in a single IO/network operation. Instead of using SAI, a design is proposed where the entire bitstream is described in the file in terms of samples, so that subsample, sample groups and sample auxiliary information stay the same, and allow a file reader to “thin” the file. The design can be summarized as follows

* A single track, no extractors-like constructs
* Keep the original design for base split-layer
* Add for each additional split-layer one or more decoder configurations if desired, in the sample description
* For non-fragmented cases
  + Add a hierarchy of sample sizes boxes to describe non-base split-layers
  + Use either split-layer chunks offsets reconstructed from lower split-layers offsets.
  + or same rules as sai for location: same sample to chunk grouping, and chunk offsets / enhancement sizes indicated per split-layer
* For fragmented case
  + Have an extension to track runs to describe split-layers sizes and offsets
  + Always use explicit data offset from moof

More details can be found in m70232

## Proposed extensions

Definitions:

* *split-layer:* a part of a sample. A split-layer may contain a single video enhancement layer or several ones, or multiple tiles for example
* *base split-layer* : *split-layer* corresponding to the sample data described by existing ISOBMFF structures (sample size, offset, parent chunk)

### Split Sample Configuration

In SampleTableBox, define at most one, present first in stbl

SplitSampleLayerConfigBox extends FullBox(‘sslc’, 0, 0){

unsigned int(32) num\_split\_layers;

}

num\_split\_layers gives the number of split layers used in this track, not including the base split layer.

### Split Sample Description

The base split-layer uses classical sample description.

In SampleEntry, allow at most one (mandatory if split layers are present) :

SplitSampleDescriptionsBox extends Box(‘sshd’){

SampleEntryBox boxes[] //any sample entry of the same base type as the sample entry, eg lhv1/lhe1…

}

The default reconstruction process is to aggregate the base split-layer sample data and each of the additional split-layers, in the order they are defined in the SplitSampleDescriptionsBox.

We propose to further describe possible reconstruction processes and backward-compatibility of the base split-layer in the sample entry of the track with the following:

SplitSamplesConfigurationBox extends FullBox(‘ssdC’, 0, 0)

{

bit(1) all\_split\_layers\_independent; // tiles, no coding deps

bit(2) single\_base\_compatible;

bit(1) no\_default\_reaggregation;

bit(1) other\_reaggregations;

bit(3) reserved;

if (other\_reaggregations) {

unsigned int(32) num\_reaggregations;

[

unsigned int(32) nb\_split\_layers\_reagg;

unsigned int(32) split\_layers\_index[nb\_split\_layers\_reagg];

unsigned int(32) num\_configurations;

unsigned int(32) split\_stsd\_idx[num\_configurations];

](num\_reaggregations)

}

}

all\_split\_layers\_independent indicates, if set to 1, that all split layers are independent and could be processed independently

single\_base\_compatible takes one of the following values:

* 0: the sample shall be reaggregated to be compliant to the associated sample entry type
* 1: the track without the additional split-layers is conformant to the sample entry type and does not have any empty samples introduced by the splitting process.
* 2: the track without the additional split-layers is conformant to the sample entry type only after removal of 0-size samples.
* 3: reserved

no\_default\_reaggregation indicates, if set to 1, that the default reaggregation is not possible. The default reaggregation consists in appending to the sample data the split-layer data in the order the layers are described in the SplitSampleDescriptionsBox

other\_reaggregations indicates, if set to 1, that other reaggregations than the default one are possible. This bit shall be set to one if no\_default\_reaggregation is set to 1.

num\_reaggregations indicates the number of additional possible reaggregations

nb\_split\_layers\_reagg indicates the number of split layers in a reaggregation

split\_layers\_index indicates the index of the split layer to aggregate, with the value 0 being the base sample data

num\_configurations gives the number of split sample entries used by this reconstruction

split\_stsd\_idx is the index of a sample entry used by this reaggregation, the value 0 meaning the base sample entry, the value 1 being the first entry in SplitSampleDescriptionsBox. This value shall not be greater than the number of entries in SplitSampleDescriptionsBox. Any codec specific data (parameter sets or equivalent for non video media types) present in this entry shall be forwarded to the decoder in the order of appearance in this list. The last entry shall indicate the highest codec requirements for the reaggregation, e.g. highest profile-tier-level in video coding.

### Split-layer storage for non-fragmented cases

In SampleTableBox, define at most one:

SplitSampleSizesBox extends Box (‘ssss`)

{

//as many as split-layers sslc.num\_split\_layers

//if no data in split sample, allow size=0

SampleSizeBox boxes[]; //and only of this type; if no data for a sample, sample\_size 0 is allowed

}

In SampleTableBox, define at most one:

SplitChunkOffsetBox extends Box(‘ssco’) {

//as many as split-layers sslc.num\_split\_layers

Box chunk\_offsets[]; //only stco or co64

}

The chunking shall be the same between all split-layers, i.e. a single SampleToChunk box is present.

The ChunkOffsetBox or ChunkLargeOffsetBox indicates the chunk size of the samples from the base split-layer only. It is allowed to have a chunk size of 0 if all data for the samples of the chunk are in split-layers.

TimeToSampleBox, CompositionToDecodeBox, SyncSampleBox and all other boxes are defined for the complete bitstream (include all AUs). When bytes are trimmed from the layers, any of these boxes referring to byte ranges (CENC SAI, subsamples) will have to be re-interpreted by the player (see section **Error! Reference source not found.**).

The SampleSizeBox defines the size of the samples in the base split-layer only. It is allowed to have a sample size of 0 if the data for the full sample is only present in other split-layers.

The SplitSampleSizesBox defines the size of the non-base split-layers in the order they are defined in the SplitSampleDescriptionsBox.It is allowed to have a sample size of 0 if there is no data for the given split-layer. Each of the SampleSizeBox included in SplitSampleSizesBox shall have the same sample count as the SampleSizeBox in the SampleTableBox.

If SplitChunkOffsetBox is absent, the chunk offset for split-layer N is equal to the chunk offset of split-layer N-1 (if 0, chunk offset of the base) plus the total size of the samples for the split-layer N-1 in the chunk.

Otherwise (SplitChunkOffsetBox is present), the offset for chunk K for split-layer N is given by the K-th entry of stco/co64 box at the N-th entry in this box, 1 being the first entry. There shall be as many entries in each of these chunk offsets boxes as there are entries in the chunk offset box for the base split-layer (present in the SampleTableBox).

### Split-layer storage for fragmented cases

In TrackFragmentBox add (at most one):

SplitSampleTrackRunBox extends FullBox(‘srun, 0, 0)

{

unsigned int(32) nb\_entries; //shall be the same as the number of split-layers sslc.num\_split\_layers

unsigned int(32) sample\_count; // shall be the same as the sample\_count from immediately preceding track run

for (i=0; i<nb\_entries; i++) {

signed int(32) data\_offset; // offset relative to containing moof!   
 {  
 unsigned int(32)sample\_size;  
 }[ sample\_count ]  
 }

}

Each SplitSampleTrackRunBox shall immediately follow the TrackFragmentBox containing the information for the base split-layer. Per construction, the number of split-layer samples per track run is the same as the number of associated base split-layer samples per track run.

### Backward compatibility

The above proposal assumes that any description of the sample (sample groups, subsample, sample auxiliary information) is done on the complete reaggregated sample, and if performing sample thinning, a file reader must edit these descriptions accordingly.

Backward compatibility is possible as long as none of these description structures include byte ranges from the additional split-layers. This works well in most cases, but is problematic in common encryption using subsamples: if all split-layers are protected, a reader will encounter a CENC SAI describing more bytes than are present in the sample. We would need to test implementations to understand how they behave in such cases.

One approach to this problem could be to use per split-layers CENC SAI with the same mechanism as above, i.e. an array of saiz (resp. saio) boxes for each split-layer SAI size (resp. SAI offset). This would be required only when backward compatibility is desired.

Another approach could be to use only a new CENC scheme (as one will probably be defined soon) which explicitly warns readers about such cases.

For use cases where only the base split-layer is encrypted, it would be interesting to allow a CENC SAI to describe less than the sample size, leaving unmapped bytes in clear. This would avoid adding many subsamples with no protected byte ranges.

Full sample encryption modes would not be recommended (to preserve access to split layers).

# Compressed MetaBox (MPEG #146, issue [#151](https://git.mpeg.expert/MPEG/Systems/FileFormat/HEIF/-/issues/151)))

## Abstract

During MPEG #146 the compression of MetaBox with the deflate algorithm for the slimHEIF design was proposed. More than 250 HEIF files for different HEIF usecases were compressed, resulting into an average compressed MetaBox size of approximately 213 bytes.

## Proposal

ISO/IEC 14496-12 supports compression of top-level boxes in subclause 8.19. The processing model when using compression of a top-level box is also well established.

One of the requirement of SlimHEIF design has been to reduce the overhead of HEIF file headers.

In this proposal a solution of compressing the file-level MetaBox with the deflate algorithm was explored. This proposal presents the result of compressing the file-level Metabox against the original file-level MetaBox.

For simulation the conformance files with file-level MetaBox [1][2] was used. The conformance files include different HEIF use cases and is not limited to small images. The file-level MetaBox was compressed using Python zlib library using default settings. The results indicate a saving of 30% on average over 250+ files, whereas in some cases the savings peaked at 65% and above. The average compressed MetaBox size was approximately 213 bytes.

The compression of MetaBox as one of the solutions for the design of SlimHEIF may be considered as it has the following advantages.

* already has a well established processing model
* significant bitrate savings compared to MetaBox
* likely minimal changes to specification text

## References

1. HEIF conformance files Nokia’s GitHub repository <https://github.com/nokiatech/heif_conformance/tree/master/conformance_files>
2. AVIF conformance files

<https://github.com/AOMediaCodec/av1-avif/tree/master/testFiles>

# On handling invalid depth bands (MPEG #150, issue[#350](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/350)))

## Abstract

When using stereo-based depth acquisition systems, it results in the presence of invalid depth bands in the captured depth images. In MPEG 150, m72330 proposes a solution to facilitate the handling of invalid depth bands.

## Discussion

Invalid depth bands are frequently observed in depth images captured by stereo-based depth acquisition systems, due to non-overlapping fields-of-view of the colour cameras.

As an example, the Intel RealSense D400 series depth cameras are equipped with two identical colour cameras that are displaced by a certain distance (i.e., baseline). Therefore, there is always a part of the scene that falls in the field-of-view of one colour camera and not in the others. As a result, an invalid depth band will be present in captured depth images.

## Proposal

To efficiently identify and handle invalid depth bands present in a depth image, the invalid depth band box is proposed, with syntax elements indicating the side and the width of an invalid depth band in a depth image.

## Invalid depth band

### Definition

Box Type: 'indb'   
Container: DepthInformationBox  
Mandatory: No  
Quantity: One

The InvalidDepthBandBox is used to provide information related to the presence of an invalid depth band in a depth image, independently of the coding.

### Syntax

aligned(8) class InvalidDepthBandBox extends Box('indb') {  
 unsigned int(2) band\_side;

unsigned int(14) band\_length;  
}

### Semantics

**band\_side** specifies the side of the depth image in which the invalid depth band is located. When the band\_side value is equal to 0, the band is located at the left side of the depth image. Analogously, when the band\_side value is equal to 1, 2, and 3, the band is located at the top, right and bottom side, respectively.

**band\_length** specifies the length of the invalid depth band in the cropped depth image in pixel dimensions. When the band\_side value is equal to 0 or 2, the band\_length value indicates the width in pixel dimensions. When the band\_side value is equal to 1 or 3, the band\_length value indicates the height in pixel dimensions.

# On supporting high bit-depth depth maps (MPEG #150, issue[#354](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/354)))

## Abstract

Currently, depth data is assumed to be encoded using existing video or image codecs, with widespread support for 8-bit and 10-bit depth formats. However, it is observed that captured depth data is frequently acquisitioned in higher bit depths, such as 16, 24, or even 32 bits. This discrepancy creates a potential incompatibility between current coding standards and real-world depth data requirements.

To address this issue, m72359 proposes documenting the bit depth incompatibility between standard codecs and higher-precision depth data in the TuC, for further analysis and possible solutions.

## Proposal

## Google AR Core API

Google’s AR Core, supports capturing Depth Images and Raw Depth Images. Depth Image is a post-processing product of Raw Depth Image using its corresponding Confidence Image to match each pixel to a depth value. Both have the depth information for each pixel returned as an unsigned 16-bit with values representing millimeters.

## Intel RealSense Depth Cameras SDK

Intel’s RealSense is a series of depth cameras that rely on various technologies (infrared, stereo, LiDAR) to capture the geometry of a scene. Depending on the camera model their API offer the depth stream in 16bits for L515 and D400 series, to 32bits for the SR300.

It is evident that the need for supporting depth maps in a range of bit depths varying from 8 to 32 bits is essential in order to be able to avoid data loss. Considering that existing video encoders/decoders can handle 8/10bits there are some aspects that can be taken in account:

* Compatibility with existing 8/10bit encoding/decoding pipeline
* Accessing depth ranges of 8/10bits pp to avoid data loss
* Avoid creating artifacts by using multiple colour channels for encoding. Additionally, avoid losing information by using chroma subsampling.

# On Alpha, Depth and Gain map metadata (MPEG #150, issue [#355](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/355)))

## Abstract

Containers and codecs such as ISOBMFF, HEIF, AV1 etc. allow for multiple ways to signal the presence and location of alpha/transparency metadata, depth map metadata, and High Dynamic Range gain map metadata, but there are also gaps and missing use cases.

m72499 presents the current options and suggests future designs for images and videos.

## Proposal

**Auxiliary data in ISOBMFF files**

Auxiliary video data can be stored in two ways in an ISOBMFF container:

1. Separate Tracks or Items: Auxiliary data is encoded independently from the primary media data, then associated through ISOBMFF container-level signaling (e.g., using track or item reference types).

2. Layered Codecs in a Single Track or Item: Auxiliary data and primary data coexist within the same elementary stream, organized using layered codecs such as MV-HEVC.

More details can be found in m72499.

**Main missing features**

HEIF standardized way to signal depth metadata for HEIF still images and image sequences, although the depi box from ISO/IEC 23001-17 could be used (but the metadata it provides is limited)

Gain map signalling (presence, and possibly static and/or dynamic metadata) for HEIF image sequences

Alpha/depth signalling for a layer of a track (presence, and possibly static and/or dynamic metadata)

Gain map signalling for tracks and layers/subtracks (presence, and possibly static and/or dynamic metadata)

Signaling of multiview, 4:4:4 components in separate layers, & other future views/auxiliary data types?

**Requirements**

At a high level, we propose adding new sample grouping types as defined in ISOBMFF (ISO/IEC 14496-12:2022 8.9 "Sample group structures") for alpha metadata, depth metadata and gain map metadata. This approach works for both AVM in ISOBMFF movie files and HEIF image sequences, whether the data is in a separate track or if it's in a layer. The presence of a SampleToGroupBox and/or SampleGroupDescriptionBox of the relevant type signals the presence of this auxiliary data. If the timed metadata (SampleToGroupBox) is not present, it may be signalled within the media bitstream in a codec-specific way. In this case, the presence of a SampleGroupDescriptionBox still signals that this auxiliary data type is present and may provide some static metadata.

Alternatively, essential sample groups may be used to avoid duplicating information in sample groups and sample entry.

For signalling in HEIF still images, such as signalling of depth metadata, the same boxes as the SampleGroupDescriptionBox boxes can be reused as item properties.

A simpler alternative to new sample grouping types is to add new boxes to the VisualSampleEntry box. This could be suitable for metadata that is unlikely to change per frame, such as alpha metadata. The box would need the ability to signal that the metadata refers to frames of a specific layer of the video when using a layered codec.

Open question: how does this interact with the draft ISO/IEC 14496-12 8th edition AMD 3 which adds new boxes to the VisualSampleEntry box to signal static metadata: alpi (AlphaInformationBox), and depi (DepthInformationBox) for depth, and also suggests adding a new handler type for depth? Should it be changed or can both be combined?

# Inventory of technologies to store a sequence of image-encoded frames (MPEG #150, m72634)

## Introduction

m72634 presents the current status of an ongoing activity on gathering existing technologies to store sequences of image-encoded frames. The current version might be partial and only aims at being presented to gather comments and pointers for further search, that would lead to a complete list in the next revision.

## Background

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Standard | Timing information | Intra prediction | Sample formats | Handler type |
| HEIF [2] | Advisory | Yes | AVC, HEVC, VVC, EVC, JPEG, other | ‘pict’ |
| MIAF [3] | \*Yes/Advisory | Yes | AVC, HEVC | \*‘vide’/‘pict’ |
| Motion JPEG 2000 [4, 5] | ^Yes/Advisory | Yes | JPEG 2000 | ^‘vide’/‘pict’ |
| JPEG XR [6] | Advisory | Yes | JPEG XR | ‘pict’ |
| Motion JPEG XR [7] | ^Yes/Advisory | Yes | JPEG XR | ^‘vide’/‘pict’ |
| Uncompressed [8] | ^Yes/Advisory | Yes | UNCV, UNCI | ^‘vide’/‘pict’ |

\* Video tracks/Image tracks

^ ISOBMFF/HEIF

## HEIF

According to Clause 7.1:

“Image sequences are stored in the ISO base media file format in tracks. In order **to distinguish image sequences from video**, the **handler type** in the HandlerBox of the track is **'pict'** to indicate an image sequence track. In particular, in an image sequence track, **the timing is advisory**: it may be the ***timing at collection*** (e.g. of an image burst) or the ***suggested display timing*** (e.g. for a slide show). In all other respects, the definitions and requirements for a video track apply unless otherwise specified in this document.”

According to Clause 7.2.2:

“The handler type of an image sequence track shall be **'pict'**. When the syntax and semantics of features of the ISO base media file format are applied to a track with the 'pict' handler type, the specifications for a track with the 'vide' handler type apply, unless otherwise specified in this document. Specifically, when handler\_type is equal to 'pict', the **VisualSampleEntry** structure is used in the **SampleDescriptionBox** and the **VisualSampleGroupEntry** structure is used in the **SampleGroupDescriptionBox**. The sample entry shall be used as specified for storage in a track with the handler type 'vide'.”

## MIAF

According to Clause 6.2:

“Image items are suitable when neither timing nor coding dependency is required. If either timing is required or the coding scheme takes advantage of inter-picture prediction, **an image sequence track** or ***a video track*** is used. A key difference between image sequence tracks and video tracks is that, in image sequence tracks, the **timing is advisory**. Consequently, it is suggested that MIAF files contain image sequence tracks when timing is not meaningful or essential for playback — for example, images captured with exposure or focal bracketing.”

According to Clause 7.2.1.5:

“The file-level MetaBox shall be identified as carrying images by carrying in the associated HandlerBox the value required by ISO/IEC 23008-12 (which is **'pict'**).”

According to 7.2.2:

“ISO/IEC 23008-12 mandates this box for 'pict' tracks; it may also be present for **'vide'** tracks.”

Editor’s note: It is not clear whether carrying a video track requires a ‘vide’ handler type.

## Motion JPEG 2000

**Based on ITU-T 802 (Motion JPEG 2000 [4] – Encapsulation in ISOBMFF):**

According to clause 3.1:

According to clause 3 Definitions:

“Motion sequence: **A timed sequence** of JPEG 2000 images”

According to clause E.2:

“**Motion JPEG 2000 video is stored in video tracks, as documented in the ISO Base Media File Format**. This format defines the structure of video tracks in a format-independent fashion. The exact format of the video is declared by the sample description; the sample description for Motion JPEG 2000 is defined in Clause 6 of this Recommendation | International Standard.”

According to clause 6.1.2:

“class MJ2SampleEntry() extends VisualSampleEntry ('mjp2'){

JP2HeaderBox();

FieldCodingBox(); // optional

MJP2ProfileBox(); // optional

MJP2PrefixBox(); // optional

MJP2SubSamplingBox(); // optional

MJP2OriginalFormatBox(); // optional

}”

**Based on ITU-T 815 (Motion JPEG 2000 [5] – Encapsulation in HEIF):**

According to clause 7.2:

“**JPEG 2000 image sequence is an image sequence**, as defined in ISO/IEC 23008-12, that conforms to the following:

* each visual sample entry shall be a JPEG 2000 visual sample entry, as specified in 7.3;
* each sample shall consist of one Contiguous Codestream box as specified in Rec. ITU-T T.800 | ISO/IEC 15444-1; and
* each sample shall be a sync sample.”

According to clause 7.3.1:

“class J2KSampleEntry extends VisualSampleEntry('j2ki') {

J2KHeaderInfo j2kheader;

J2KCodestreamPrefix j2kprefix;

}”

## JPEG XR

**Based on ITU-T T.832 (JPEG XR [6] – Encapsulation in HEIF):**

According to clause F.1:

“This annex specifies a file format to encapsulate images, image collections, and **image sequences** coded using the JPEG XR format in the image file format specified in **ISO/IEC 23008-12**.”

According to clause F.7.2:

“For JPEG XR coded **image sequences**, the codecs parameter contains one or both of the 'mjxr' or 'mjxs' sample entry types, and the first value of the codecs parameter should represent a track having the **handler type 'pict'**; other list items represent other tracks.”

According to clause F.4.2:

“A sample entry of type **'mjxr'** shall be used for an image sequence track coded with JPEG XR, as defined in **Rec. ITU-T T.833** | ISO/IEC 29199-3, using the **MJXRSampleEntry** and sample format as specified in Rec. ITU-T T.833 | ISO/IEC 29199-3.”

## Motion JPEG XR

**Based on ITU-T T.833 (Motion JPEG XR [7] – Encapsulation in ISOBMFF):**

According to clause 1:

“This Recommendation | International Standard specifies the use of JPEG XR coding for timed sequences of images (Motion JPEG XR sequences) within files based on the **ISO base media file format**.”

According to clause 3.1:

“Motion JPEG XR sequence: **A timed sequence of JPEG XR images**”

According to clause 4.2:

“Except as otherwise specified herein (if any such differences exist), files conforming to this Recommendation | International Standard **shall conform to the file format specified in ISO/IEC 15444-12**.” [ISOBMFF]

According to clause 5.1.1:

“The format of a sample when the **sample entry** name is 'mjxr' is a CODED\_IMAGE() **as defined in Rec. ITU-T T.832** | ISO/IEC 29199-2, *without* the IMAGE\_HEADER(). Each image presented to a JPEG XR decoder is logically formed by appending the content of each sample to the content of the JPEG XR Header Box in its associated Visual Sample Entry. …”

According to clause A.2:

“At least one **video track** is present, using at least one **MJXRSampleEntry**.”

According to clause 5.1.2:

“class MJXRSampleEntry() extends VisualSampleEntry('mjxr'){

JPEGXRInfoBox();

JPEGXRHeaderBox();

JPEGXRProfileBox(); // optional

ColourInformationBox(); // optional

}”

## Uncompressed

According to clause 1:

“Uncompressed frames may be stored in **ISO base media** files as media samples of media tracks or as **image items** using a generic uncompressed video description defined in this document.”

According to clause 4.2:

“The **handler type** associated with the track is usually **'vide'**, **'auxv'** or **'pict'** but derived specifications may introduce new handler types.”

According to clause 4.2:

“Uncompressed video tracks compliant to this document are tracks compliant to **ISO/IEC 14496-12** that use a **VisualSampleEntry** with codingname equal to **'uncv'**, hereafter called uncompressed video sample entry.”

According to clause 4.3:

“An uncompressed image compliant to this document is an image item compliant to **ISO/IEC 23008-12** with the item\_type **'unci'**.”

# References

1. WG03 MPEG Systems, “WD of ISO/IEC 14496-12 8th edition AMD 3 Carriage of depth and alpha”, [N014055](https://dms.mpeg.expert/doc_end_user/documents/149_Geneva/wg11/MDS24755_WG03_N01455.zip), Geneva, CH, January 2025
2. ISO/IEC 23008-12:2022 “Information technology — High efficiency coding and media delivery in heterogeneous environments - Part 12: Image File Format (HEIF)”
3. ISO/IEC 23000-22:2025 “Information technology — Multimedia application format (MPEG-A)Part 22: Multi-image application format (MIAF)”
4. ITU-T T.802 “Information technology – JPEG 2000 image coding system: Motion JPEG 2000”
5. ITU-T T.815 “Information technology – JPEG 2000 image coding system – Encapsulation of JPEG 2000 images into ISO/IEC 23008-12”
6. ITU-T T.832 “Information technology – JPEG XR image coding system – Image coding specification”
7. ITU-T T.833 “Information technology – JPEG XR image coding system – Motion JPEG XR”
8. ISO/IEC 23001‑17 “Information technology — MPEG Systems technologies — Part17: Carriage of uncompressed video and images in ISO Base Media File Format”

# On multilayer mapping with ISOBMFF structures (MPEG #150, Issue [#357](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/357))

## Abstract

m72361 proposes the LayerConfigurationBox which provides information about ISOBMFF structures (boxes) when a multilayer bitstream is stored in a single track.

## Introduction

The file format group started a new exploration topic in MPEG 149 to study, how the existing Sample entry box structures apply to different layers when a multilayer bitstream is stored in a single track.

The ISO/IEC 14496-12 ISOBMFF specification currently documents the following boxes to be optionally present in the VisualSampleEntry of the track (Clause 12).

1. PixelAspectRatioBox ('pasp')
2. CleanApertureBox ('clap')
3. ColourInformationBox ('colr')
4. ContentLightLevelBox ('clli')
5. MasteringDisplayColourVolumeBox ('mdcv')
6. ContentColourVolumeBox ('cclv')
7. AmbientViewingEnvironmentBox ('amve')

The boxes stated above may have a corresponding SEI message in the video bitstream contained in the track. The video coding and SEI specifications enables provisions to map information in the SEI message to a specific layer of a multilayer bitstream (for e.g., using nuh\_layer\_id in the SEI NAL unit header). However, the ISOBMFF specification does not define any mapping information from the information in the box to the layers present in a multilayer bitstream.

## Proposal

### x.x Layer Configuration Box

### x.x.1 Definition

Box Types: 'layC'  
Container: VisualSampleEntry  
Mandatory: Optional  
Quantity: Exactly one

The LayerConfigurationBox contains information related to one or more layers of a multilayer bitstream present in the track. It documents the boxes associated with each layer of a multilayer bitstream present in the track.

A specific layer may be associated with one or more boxes with the same 4cc values (for e.g., ColourInformationBoxes), in such a case the processing defined for the corresponding box applies.

If information is supplied in both this box, and in the video bitstream, this box takes precedence, and over-rides the information in the bitstream. If boxes are absent both the VisualSampleEntry and in the LayerConfigurationBox then structures specific to the video codec should be examined. For maximum compatibility, the boxes defined in LayerConfigurationBox should follow, not precede, any boxes defined in or required by derived specifications.

**x.x.2 Syntax**

aligned(8) class LayerConfigurationBox  
 extends FullBox('layC', version = 0, flags = 0)   
{  
 unsigned int(16) num\_layers;   
 for(int i=0; i<num\_layers; i++){

unsigned int(6) layer\_id[i];

unsigned int(8) num\_associated\_boxes;

for(int j=0; j< num\_associated\_boxes; i++){

unsigned int(1) box\_inferred\_from\_other\_layer\_flag;

if(box\_inferred\_from\_other\_layer\_flag){

unsigned int (6) reference\_layer\_id;

unsigned int(32) box\_4cc;

}else{

Box boxes[]; // any of the Sample Entry Boxes

}

}

}

trailing\_bits();

}

#### x.x.3 Semantics

version and flags shall be set to 0.

num\_layers specifies the number of layers of a multilayer bitstream present in the track for which the configuration information is defined. The value of shall be one less than the number of layers in the multilayer bitstream present in the track. It is assumed that the information related to the base layer is present in the Sample entry of the track.

layer\_id[i] specifies the nuh\_layer\_id of the ith layer for which the associated boxes are defined.

num\_associated\_boxes specifies the number of boxes associated to the ith layer.

box\_inferred\_from\_other\_layer\_flag when set to 1 indicates that the box with 4cc equal to box\_4cc is inferred from the layer with nuh\_layer\_id equal to reference\_layer\_id

box\_inferred\_from\_other\_layer\_flag when set to 0 indicates that a box associated to the layer is present in the LayerConfigurationBox.

reference\_layer\_id specifies the layer\_id of the layer from which the box with 4cc box\_4cc is inferred. When the value of reference\_layer\_id is equal to 0, it refers to the boxes in the VisualSampleEntry.

box\_4cc indicates the 4cc value of the box which is inferred from other layers.

Two options are discussed to address similar use case of layered coding with multiple enhancements:

1. describing the change each layer provides (=node in a tree)
2. describe the achieved user experience (=leaves of a tree)

Both approaches have their pros and cons, including:

* Option 1 describes the actual coding layout.
* Option 2 provides an abstraction between coding and achieved experience.

The proposed LayerConfigurationBox fulfills option 1, while utilizing PreselectionGroupBoxes as utilized as basis in contribution **m72211** ( [#356](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/356" \o "m72211 [7.5][ISOBMFF] On codecs parameter string with ISOBMFF preselections) ) (and documented in DASH 6th Edition) follow the principle of option 2.

# On Generic Sample Extensions (MPEG #150, Issue [#358](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/358) [#360](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/360))

## Introduction

At MPEG#150, m72400 and m72362 provide solutions for carrying data from additional layers of a multilayer bitstream in a backward compatible single-track. While m72400 is an update to split-layer based solution proposed in m70232 (also in Section 17 above); m72400 proposes a sample extension solution to carry data from additional layers of a multilayer bitstream.

## Discussion

For illustration the difference between m72400 and m72362 are shown in the figures below.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Further discussion can be found in [#358](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/358).

## Proposal, m72400

3.1.XX

split-sample (multi-layer sample)

sample of a track carrying all the data from a multi-layer bitstream or multi-codec bitstream.

XX Split-samples

XX.1 Definition

A split-sample is a sample of a track that carries all the data of a multi-layer bitstream or multi-codec bitstream.

A split-sample can be described as multiple split-layers, a base split-layer and one or more additional split-layers.

A split-layer is a part of a sample data. A split-layer may represent a single video enhancement layer or several ones, or multiple tiles, or part of a sample coded with a different codec from another part of the sample for example.

The data corresponding to the split-layers of a split-sample can be stored either as a contiguous piece of data of a sample of a track or can be stored as non-contiguous pieces of data of a sample in a track.

When the split-layers of a split-sample are stored as non-contiguous pieces of data of a sample of the track:

* The offsets/sizes and timing of the part of sample data corresponding to the base split-layer are described by boxes directly declared in SampleTableBox or TrackFragmentBox (e.g. sample size, chunk offset, sample to chunk, time to sample).
* The offsets/sizes of the part of sample corresponding to an additional split-layer are described within a SplitLayerBox in the SampleTableBox or TrackFragmentBox.
* The original sample representing all the split-layers of a split-sample is obtained by reaggregating the data of the split-layers following the reconstruction process described by SplitSamplesConfigurationBox.

When the split-layers of a split-sample are stored as a contiguous piece of data of a sample of the track:

* The offsets/sizes and timing of the sample data corresponding to all the split-layers of the split-sample are described by boxes directly declared in SampleTableBox or TrackFragmentBox (e.g. sample size, chunk offset, sample to chunk, time to sample).

XX.2 Split sample layer configuration

XX.2.1 Definition

Box Type: 'sslc'

Container: SampleTableBox or TrackFragmentBox

Mandatory: Yes, if split-layers are present

Quantity: Zero or one (per container)

A SplitSampleLayerConfigBox, when present, indicates that samples of the track are composed of multiple layers described by split-layers and describes general configuration information about the split-layers of the track.

contiguous\_sample\_data\_idc indicates the sample data organization of split-layers:

* When it equals to 0, the data of each split-layer is contiguous within a sample and any SplitLayerBox in SampleTableBox or corresponding TrackFragmentBox shall not comprise SampleSizeBox (or CompactSampleSizeBox) and ChunkOffsetBox (or ChunkLargeOffsetBox).
* When it equals to 1, the data of split-layers of a split-sample is split into different contiguous chunks. Any SplitLayerBox in SampleTableBox or corresponding TrackFragmentBox shall comprise SampleSizeBox (or CompactSampleSizeBox) and shall not comprise ChunkOffsetBox (or ChunkLargeOffsetBox).
* When it equals to 2, the data of split-layers of a split-sample is split into different non-contiguous chunks. Any SplitLayerBox in SampleTableBox or corresponding TrackFragmentBox shall comprise both SampleSizeBox (or CompactSampleSizeBox) and ChunkOffsetBox (or ChunkLargeOffsetBox).

XX.2.2 Syntax

aligned(8) class SplitSampleLayerConfigBox

extends FullBox('sslc', 0, 0)

{

unsigned int(2) contiguous\_sample\_data\_idc;

unsigned int(1) SAI\_per\_split\_layer\_flag;

bit(5) reserved;

unsigned int(32) num\_split\_layers;

}

XX.2.3 Semantics

contiguous\_sample\_data\_idc indicates, if set to 0, that data of the split-layers for a split-sample are contiguous within a sample of the track. It indicates, if set to 1, that data of split-layers of a split-sample is split into different contiguous chunks. It indicates, if set to 2, that data of split-layers of a split-sample is split into different non-contiguous chunks. Value 3 is reserved. When not present, the default value for contiguous\_sample\_data\_idc is 0.

SAI\_per\_split\_layer\_flag indicates, if set to 0, that sample auxiliary information directly declared in the SampleTableBox is related or apply to the reaggregated sample, and if set to 1, that sample auxiliary information directly declared in the SampleTableBox is related or apply to the part of the sample corresponding to the base split-layer only.

num\_split\_layers gives the number of split-layers used in this track, not including the base split-layer.

XX.3 Split-sample sample entry

XX.3.1 Definition

Sample entries are defined to give detailed information about the coding type used, and optionally any initialization information needed for that coding and property information for each split-layer.

A sample entry describing the base split-layer is a sample entry defined in SampleDescriptionBox.

A sample entry describing an additional split-layer is a sample entry defined in a SplitSampleDescriptionsBox in the sample entry describing the base split-layer.

Box Type: 'sshd'

Container: SampleEntry box in SampleDescriptionBox

Mandatory: Yes, if split-layers are present

Quantity: Zero or one

A SplitSampleDescriptionsBox provides for each additional split-layer one or more decoder configurations if desired, in the sample description of the track. It is a container box for sample entries describing each split-layer, excluding the base split layer.

The 1-based index of the SampleEntry box in SplitSampleDescriptionsBox in declaration order indicates the layer identifier of the split-layer described by the SampleEntry. The value 0 is reserved and represents the base layer.

XX.3.2 Syntax

aligned(8) class SplitSampleDescriptionsBox extends Box('sshd') {

SampleEntry()[];

}

XX.3.3 Semantics

SampleEntry()[] is an array of SampleEntry boxes describing the codecs and configuration information, if needed, for decoding each split-layer, and the properties associated with the split-layer, excluding the base split-layer.

XX.4 Split-samples configuration box

XX.4.1 Definition

Box Type: 'ssdC'

Container: SampleEntry box in SampleDescriptionBox

Mandatory: No

Quantity: Zero or one

The SplitSamplesConfigurationBox provides parameters for defining the reconstruction process of a reaggregated or “full” sample from data of the base split-layer and additional split-layers.

The default reconstruction process is to aggregate the base split-layer sample data and each of the additional split-layers, in the order they are defined in the SplitSampleDescriptionsBox.

The SplitSamplesConfigurationBox allows describing alternative possible reconstruction processes, if any, and backward-compatibility of the base split-layer in the sample entry of the track.

XX.4.2 Syntax

SplitSamplesConfigurationBox extends FullBox('ssdC', 0, 0)

{

bit(1) all\_split\_layers\_independent;

bit(2) single\_base\_compatible;

bit(1) no\_default\_reaggregation;

bit(1) other\_reaggregations;

bit(3) reserved;

if (other\_reaggregations) {

unsigned int(32) num\_reaggregations;

[

unsigned int(32) nb\_split\_layers\_reagg;

unsigned int(32) split\_layers\_index[nb\_split\_layers\_reagg];

unsigned int(32) num\_configurations;

unsigned int(32) split\_stsd\_idx[num\_configurations];

](num\_reaggregations)

}

}

XX.4.3 Semantics

all\_split\_layers\_independent indicates, if set to 1, that all split-layers are independent and could be processed independently. If set to 0, layer dependencies is unknown. The default value is 0.

single\_base\_compatible takes one of the following values:

- 0: the sample shall be reaggregated to be compliant to the associated sample entry type

- 1: the track without the additional split-layers is conformant to the sample entry type and does not have any empty samples introduced by the splitting process.

- 2: the track without the additional split-layers is conformant to the sample entry type only after removal of 0-size samples.

- 3: reserved

The default value is 0.

no\_default\_reaggregation indicates, if set to 1, that the default reaggregation is not possible. The default reaggregation consists in appending to the sample data the split-layer data in the order the layers are described in the SplitSampleDescriptionsBox. The default value is 0.

other\_reaggregations indicates, if set to 1, that other reaggregations than the default one are possible. If set to 0, there is no information on other possible reaggregations than the default one. The default value is 0.

NOTE If no\_default\_reaggregation is set to 1 indicating that there is no default reaggregation of the split-layers in the reconstruction process and other\_reaggregations is set to 0 indicating that there is no other reaggregation rules in the reconstruction process, this indicates to the parser that the split-layers are not intended to be reaggregated in a “full” sample.

num\_reaggregations indicates the number of additional possible reaggregations. The default value is 0.

nb\_split\_layers\_reagg indicates the number of split-layers in a reaggregation. The default value is 0.

split\_layers\_index indicates the index of the split-layer to aggregate, with the value 0 being the base sample data. The default value is 0.

num\_configurations gives the number of split sample entries used by this reconstruction. The default value is 0.

split\_stsd\_idx is the index of a sample entry used by this reaggregation, the value 0 meaning the base sample entry, the value 1 being the first entry in SplitSampleDescriptionsBox. This value shall not be greater than the number of entries in SplitSampleDescriptionsBox. Any codec specific data (parameter sets or equivalent for non video media types) present in this entry shall be forwarded to the decoder in the order of appearance in this list. The last entry shall indicate the highest codec requirements for the reaggregation, e.g. highest profile-tier-level in video coding. The default value is 0.

XX.4 Split-layer box

XX.4.1 Definition

Box Type: 'sply'

Container: SampleTableBox or TrackFragmentBox

Mandatory: No

Quantity: Zero or more

The SplitLayerBox comprises boxes providing descriptive information related to a given split-layer other than the base split-layer. It may describe the offsets/sizes and additional descriptive information, e.g. sample auxiliary information or sample groups, of a part of the “full” samples for a split-layer other than the base split-layer.

The number of SplitLayerBox(es) in the SampleTableBox shall correspond to the parameter num\_split\_layers in the SplitSampleLayerConfigBox, when present.

The SplitLayerBox(es) shall be declared in SampleTableBox or TrackFragmentBox in same order as the corresponding SampleEntry box(es) in SplitSampleDescriptionsBox. The ith SampleEntry box in the SplitSampleDescriptionsBox is associated with the ith SplitLayerBox in declaration order.

For non-fragmented cases, the chunking shall be the same between all split-layers, i.e. a SampleToChunkBox is present only in the SampleTableBox.

The ChunkOffsetBox or (ChunkLargeOffsetBox) and the SampleSizeBox (or CompactSampleSizeBox) in the SampleTableBox provides the locations of the chunks and sizes of the data of the base split-layer only. It is allowed to have a chunk size of 0 if all data for the samples of the chunk are in split-layers.

TimeToSampleBox, CompositionToDecodeBox and SyncSampleBox in SampleTableBox are defined for the “full” samples (i.e. include all reaggregated samples).

The SplitLayerBox may comprise a SampleSizeBox (or CompactSampleSizeBox) and a ChunkOffsetBox (or ChunkLargeOffsetBox) describing the locations of the chunks and sizes of the data of the split-layer according to contiguous\_sample\_data\_idc in SplitSampleLayerConfigBox.

When present, it is allowed to have a sample size of 0 in the SampleSizeBox (or CompactSampleSizeBox) if there is no data for the given split-layer.

When present, the SampleSizeBox (or CompactSampleSizeBox) included in each SplitLayerBox shall have the same sample count as the SampleSizeBox (or CompactSampleSizeBox) in the SampleTableBox of the track.

If SampleSizeBox (or CompactSampleSizeBox) is present and ChunkOffsetBox (or ChunkLargeOffsetBox) is absent in a SplitLayerBox, the chunk offset for this split-layer is equal to the chunk offset of the preceding split-layer described by the preceding SplitLayerBox, if any in the SampleTableBox, or, by default, is equal to the chunk offset of the base split-layer plus the total size of the samples for this preceding split-layer in the chunk. Otherwise if SampleSizeBox (or CompactSampleSizeBox) is present and ChunkOffsetBox (or ChunkLargeOffsetBox) is present in a SplitLayerBox, there shall be as many entries as there are entries in the ChunkOffsetBox (or ChunkLargeOffsetBox) for the base split-layer (present in the SampleTableBox).

For fragmented case, the sample chunking is used as for the base split-layer.

The TrackFragmentBox comprises a SplitLayerBox for each split-layer other than the base split-layer.

For the split-layers other than the base split-layer, a simplified version of track runs, SplitLayerTrackRunBox, shall be used to describe split-layer sizes and optionally offsets.

XX.4.2 Syntax

aligned(8) class SplitLayerBox extends Box('sply') {

Box boxes[];

}

XX.4.3 Semantics

boxes are any box defined within the scope of a split-layer.

XX.5 Split-layer track run box

XX.5.1 Definition

Box Type: 'srun'

Container: SplitLayerBox

Mandatory: Yes

Quantity: One or more

The SplitLayerTrackRunBox describes the sizes and optionally offsets for the data in a split-layer.

In a track fragment box, there shall be as many SplitLayerTrackRunBox in each splitLayerBox as there are TrackRunBox in the track fragment box.

XX.5.2 Syntax

aligned(8) class SplitLayerTrackRunBox extends FullBox('srun', 0, 0) {

signed int(32) data\_offset;

{

unsigned int(32) sample\_size;

} [ sample\_count ]

}

XX.5.3 Semantics

sample\_count the number of samples being added in this run; also the number of rows in the following table (the rows can be empty). sample\_count is the sample\_count from the TrackRunBox describing the base split-layer in the same track fragment box.

data\_offset is added to the implicit or explicit base\_data\_offset established in the track fragment header.

XX.6 Sample auxiliary information per split-layer

Sample auxiliary information per split-layer is described by defining SampleAuxiliaryInformationSizesBox and SampleAuxiliaryInformationOffsetsBox in the SampleTableBox or TrackFragmentBox for the base split-layer and in a SplitLayerBox for an additional split-layer.

In subclause 8.7.8.1 and 8.7.9.1, replace the following:

Container: SampleTableBox or TrackFragmentBox

With

Container: SampleTableBox, TrackFragmentBox or SplitLayerBox

## Proposal, m72362

**XX Sample extension**

XX.1 Definition

Samples of a track can be extended with one or more sample extents to allow storing media as non-contiguous data in a track. A sample extent may be considered as part of track sample data when processing of sample extension is considered.

A sample extent may contain a single non-base layer or several ones, or multiple tiles, or part of a media coded with a different codec from another part of the sample for example.

The offsets/sizes related to sample extent is described within a SampleExtensionDescriptionBox in the SampleExtensionTableBox or TrackFragmentBox.

The aggregated media sample can be obtained by after combining the data of sample extents following the reconstruction process described by SampleExtensionDescriptionBox.

## XX.2 Sample Extension table

### XX.2.1 Sample table box

XX.2.1.1 Definition

Box Type: 'setb'  
Container: MediaInformationBox  
Mandatory: No  
Quantity: Exactly one

The sample extension table contains all the information related to one or more sample extents in a track. Using the table, it is possible to locate sample extents in time, determine their type (e.g. I-frame or not), and determine their size, container, and offset into that container.

The data referencing is used to retrieve the data associated with the sample extents described by this table.

The SampleExtensionTableBox, shall contain at least one SampleExtensionDescriptionBox and the SampleExtensionAggregateBox. Each SampleExtensionDescriptionBox includes the data needed to process a single sample extent. The number of sample extents to be proceed is determined by the number of SampleExtensionDescriptionBox(es) in the SampleExtensionTableBox.

XX.2.1.2 Syntax

aligned(8) class SampleExtensionTableBox extends Box('setb')  
{

}

### XX.2.2 Sample extension description box

XX.2.2.1 Definition

Box Types: 'sesd'  
Container: SampleExtensionTableBox or TrackFragmentBox  
Mandatory: Yes  
Quantity: Atleast one

When the sample extension description is present in the SampleExtensionTableBox it gives detailed information about the coding type used, and any initialization information needed for that coding for a sample extent and shall carry the Sample entry of the sample extent.

The SampleExtensionDescriptionsBox describes the offsets/sizes and additional descriptive information, e.g. sample auxiliary information, of a sample extent.

The number of SampleExtensionDescriptionsBox (es) in the SampleExtensionTableBox or TrackFragmentBox shall correspond to the number of sample extents to be processed in the track.

The SampleExtensionDescriptionsBox (es) shall be declared in SampleExtensionTableBox or TrackFragmentBox.

For non-fragmented cases, the chunking shall be the same between all sample extents, i.e. a SampleToChunkBox is present only in the SampleTableBox.

The ChunkOffsetBox or (ChunkLargeOffsetBox) and the SampleSizeBox (or CompactSampleSizeBox) in the SampleTableBox provides the locations of the chunks and sizes of the data of the track samples only. It is allowed to have a chunk size of 0 if all data for the samples of the chunk are in track samples.

TimeToSampleBox, CompositionToDecodeBox and SyncSampleBox in SampleTableBox are defined for the aggregated data (i.e. including the track samples and all sample extents).

The SampleExtensionDescriptionsBox comprises a SampleSizeBox (or CompactSampleSizeBox) and a ChunkOffsetBox (or ChunkLargeOffsetBox) describing the locations of the chunks and sizes of the data of the sample extent.

It is allowed to have a sample size of 0 in the SampleSizeBox (or CompactSampleSizeBox) if there is no data for the given sample extent.

The SampleSizeBox (or CompactSampleSizeBox) included in each SampleExtensionDescriptionsBox shall have the same sample count as the SampleSizeBox (or CompactSampleSizeBox) in the SampleTableBox of the track.

If ChunkOffsetBox (or ChunkLargeOffsetBox) is absent in a SampleExtensionDescriptionsBox, the chunk offset for this sample extent is equal to the chunk offset of the preceding sample extent described by the preceding SampleExtensionDescriptionsBox, if any in the SampleExtensionTableBox, or, by default, is equal to the chunk offset of the track samples plus the total size of the samples for this preceding sample extent in the chunk. Otherwise ChunkOffsetBox (or ChunkLargeOffsetBox) is present in a SampleExtensionDescriptionsBox, there shall be as many entries as there are entries in the ChunkOffsetBox (or ChunkLargeOffsetBox) for the sample extent (present in the SampleTableBox).

For fragmented case, the sample chunking is used as for the track samples.

The TrackFragmentBox comprises a SampleExtensionDescriptionsBox for each sample extent.

For the sample extents a simplified version of track runs, SampleExtentTrackRunBox, shall be used to describe sample extent sizes and optionally offsets.

XX.2.2.2 Syntax

aligned(8) class SampleExtensionDescriptionsBox extends Box('sesd') {

}

XX.2.3 Sample extension configuration

XX.2.3.1 Definition

Box Type: 'saec'

Container: SampleExtensionDescriptionsBox

Mandatory: No

Quantity: Zero or one (per container)

A SampleExtensionConfigBox, when present, indicates general configuration information about the sample extent of the track.

XX.2.2 Syntax

aligned(8) class SampleExtensionConfigBox

extends FullBox('saec', 0, 0)

{

unsigned int(1) SAI\_applies\_for\_sample\_extent\_flag;

bit(7) reserved;

}

XX.2.3 Semantics

SAI\_applies\_for\_sample\_extent\_flag indicates, if set to 0, that sample auxiliary information directly declared in the SampleTableBox is not related or do not apply to the sample extent, and if set to 1, that sample auxiliary information directly declared in the SampleTableBox is related or apply to the sample extent.

XX.2.4 Sample extent aggregate box

XX.2.4.1 Definition

Box Type: 'seag'

Container: SampleExtensionTableBox or TrackFragmentBox

Mandatory: Yes, if sample extents are present

Quantity: Zero or one

The SampleExtentAggregateBox provides parameters for defining the reconstruction process to aggregate data from the samples of the track and data from each of the sample extent.

The default reconstruction process is to aggregate the track sample data and each of the sample extent, in the order in which each SampleExtensionDescriptionsBox appears in the SampleExtensionTableBox or TrackFragmentBox.

XX.4.2 Syntax

SampleExtentAggregateBox extends FullBox('seag', 0, 0)

{

bit(1) all\_sample\_extents\_independent;

bit(1) no\_default\_sample\_extent\_aggregation;

bit(1) other\_sample\_extent\_aggregations;

bit(5) reserved;

if (other\_sample\_extent\_aggregations) {

unsigned int(32 num\_sample\_extent\_aggregations;

[

unsigned int(32) nb\_sample\_extents;

unsigned int(32) sample\_extent\_index[nb\_sample\_extents];

unsigned int(32) num\_configurations;

unsigned int(32) sample\_extent\_sesd\_idx[num\_configurations];

](num\_sample\_extent\_aggregations)

}

}

XX.4.3 Semantics

all\_sample\_extents\_independent indicates, if set to 1, that all sample extents are independent and could be processed independently

no\_default\_sample\_extent\_aggregation indicates, if set to 1, that the default sample extent aggregation is not possible. The default aggregation consists in appending to the track sample data the sample extent data in the order in which each SampleExtensionDescriptionsBox appears in the SampleExtensionTableBox or TrackFragmentBox.

other\_sample\_extent\_aggregations indicates, if set to 1, that other sample extent aggregations than the default one are possible.

NOTE If no\_default\_sample\_extent\_aggregation is set to 1 indicating that there is no default aggregation of the sample extents in the reconstruction process and other\_sample\_extent\_aggregations is set to 0 indicating that there is no other aggregation rules in the reconstruction process, this indicates to the parser that the sample extents are not intended to be aggregated.

num\_sample\_extent\_aggregations indicates the number of additional possible sample extent aggregations.

nb\_sample\_extents indicates the number of sample extents in a aggregation.

sample\_extent\_index indicates the index of the sample extent to aggregate, with the value 0 being the track sample data.

num\_configurations gives the number of sample extent entries used by this reconstruction.

sample\_extent\_sesd\_idx is the index of a sample entry used by this aggregation, the value 0 meaning the track sample entry, the value 1 being the entry in the first SampleExtensionDescriptionsBox. This value shall not be greater than the number of SampleExtensionDescriptionsBox(es). Any codec specific data (parameter sets or equivalent for non-video media types) present in this entry shall be forwarded to the decoder in the order of appearance in this list. The last entry shall indicate the highest codec requirements for the aggregation, e.g. highest profile-tier-level in video coding.

XX.2.5 Sample extent track run box

XX.2.5.1 Definition

Box Type: 'seru'

Container: SampleExtensionDescriptionsBox

Mandatory: Yes

Quantity: One or more

The SampleExtentTrackRunBox describes the sizes and optionally offsets for the data in a sample extent.

In a track fragment box, there shall be as many SampleExtentTrackRunBox in each SampleExtensionDescriptionsBox as there are TrackRunBox in the track fragment box.

XX.2.5.2 Syntax

aligned(8) class SampleExtentTrackRunBox extends FullBox('seru', 0, 0) {

signed int(32) data\_offset;

{

unsigned int(32) se\_sample\_size;

} [ se\_sample\_count ]

}

XX.2.5.3 Semantics

se\_sample\_count the number of samples of a sample extent being added in this run; also, the number of rows in the following table (the rows can be empty). se\_sample\_count is the sample\_count from the TrackRunBox describing the track samples in the same track fragment box.

data\_offset is added to the implicit or explicit base\_data\_offset established in the track fragment header.

XX.2.6 Sample auxiliary information per sample extent

Sample auxiliary information for each sample extent is described by defining SAIBox in the SampleExtensionDescriptionsBox.

# On cross-track synchronisation (MPEG #150, Issue [#359](https://git.mpeg.expert/MPEG/Systems/FileFormat/isobmff/-/issues/359))

## Introduction

At MPEG#149, contribution m71353 proposed means to describe cross-track random access information. This proposal received comments suggesting to check CMAF, DASH and V3C regarding this aspect. This contribution addresses these comments and describes a simplified proposal.

## Discussion

## CMAF / DASH Cross-track Random Access

ISOBMFF Amendment 2 describes tools for enhanced CMAF and DASH integration. In particular, a new SwitchingGroupBox is proposed. This new entity group typically comprises a list of switchable tracks. In particular, if the time\_aligned\_flag equals 1, the tracks have some time alignment characteristics, which are further defined by DASH and CMAF and identified by the structural brand and/or the media profile brand fields.

While not published for the moment, some documents may therefore describe that if time\_aligned\_flag equals 1, all the tracks of the group are not only time-aligned, but also have time-aligned sync samples.

## Carriage of V3C

According to V3C Carriage in ISOBMFF specification, tracks belonging to the same coding V3C sequence are time-aligned, but their sync samples are not necessarily time-aligned. In the case where they are time-aligned, the sync samples of the V3C atlas track can be considered as the random-access points for the V3C content, and random access can be done by only referencing the sync sample information of the V3C atlas track.

## Conclusion

In specific cases, the presence of cross-track random access points may be inferred from other structures signalled in ISOBMFF files created according to CMAF, DASH or Carriage of V3C data. However, no specification provides generic means to signal that different tracks have cross-track random access points. Consequently, we propose in the next section a new entity group aiming at describing such information. This entity group is a simplified version of the one initially proposed in m71353. In particular, it is no more proposed to explicitly signal cross-track random access points by indicating corresponding sample numbers.

Thanks to this new entity group, the intent of content creator in terms of cross-track random access can easily be signaled.

## Proposal

In Section 8.15.3.1 of ISOBMFF, add a new grouping\_type value as follows:

'ctra': The tracks mapped to this grouping are associated with cross-track random access information, as specified in 8.15.4.2.

Add section 8.15.4.2 Cross-track Random Access Box

**8.15.4.2 Cross-track Random Access Group Box**

**8.15.4.2.1 Definition**

Box Type: 'ctra'

Container: GroupsListBox in a MetaBox on movie level

Mandatory: No

Quantity: Zero or one

This box describes cross-track random access information for a group of tracks.

Each entity\_id value of a cross-track random access entity group shall be equal to a track\_ID.

**8.15.4.2.2 Syntax**

aligned(8) class CrossTrackRandomAccessGroupBox

extends EntityToGroupBox('ctra', version=0, flags){

if (flags & period\_present\_flag)

unsigned int(32) presentation\_time\_ra\_period;

}

**8.15.4.2.3 Semantics**

(period\_present\_flag & 1) equal to 1 indicates that an indication about random access period is present for the tracks of this group. When equal to 0, there is no indication about random access period.

presentation\_time\_ra\_period when present is an integer indicating a period at which random access points **is possible** in each track. This period is expressed based on the timescale property in containing MovieHeaderBox from a starting offset time equal to 0. When not present, there is no timing indication about the random access points. When for a given track, presentation\_time\_ra\_period does not translate to an integer considering the track’s timescale, the closest sample to the translated presentation\_time\_ra\_period is selected as random access point for this track. The closest sample is the sample with the smallest difference between the timestamp obtained from presentation\_time\_ra\_period and its presentation time