Text

Description automatically generatedISO/IEC JTC 1/SC 29/WG 03 N1469

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

**Document type:** Output Document

**Title:** Defect report of ISO/IEC 14496-12

**Status:** Approved

**Date of document:** 2025-01-24

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

**No. of pages:** 12 (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 N1469**

**January 2025, Kemer, TR**

|  |  |
| --- | --- |
| **Title** | **Defect report of ISO/IEC 14496-12** |
| **Source** | **WG 03, MPEG Systems** |
| **Status** | **Approved** |
| **Serial Number** | **24879** |

**Contents**

[1 Introduction 3](#_Toc188602599)

[2 Defects under consideration 3](#_Toc188602600)

[2.1 Segment index (sidx) 3](#_Toc188602601)

[2.2 General editing 3](#_Toc188602602)

[2.2.1 Presentation terms 3](#_Toc188602603)

[2.3 Extended language tag 6](#_Toc188602604)

[2.3.1 Observation 6](#_Toc188602605)

[2.3.2 Request from initial contribution 6](#_Toc188602606)

[2.3.3 On 2-letter codes 7](#_Toc188602607)

[2.3.4 On subtag usage 8](#_Toc188602608)

[2.4 Missing Semantics 10](#_Toc188602609)

[~~2.4.1~~ ~~Observation~~ 10](#_Toc188602610)

[~~2.4.2~~ ~~Consideration~~ 10](#_Toc188602611)

[~~2.4.3~~ ~~Proposal~~ 10](#_Toc188602612)

[2.4.4 Further identified issues w.r.t. semantics 10](#_Toc188602613)

[2.5 Missing track reference types 11](#_Toc188602614)

# Introduction

Comments on specific defects, and proposals for how to address them, can also be made in between meetings in Github, see <<https://github.com/MPEGGroup/FileFormat>> under the issue label ISOBMFF.

# Defects under consideration

## Segment index (sidx)

[*https://github.com/MPEGGroup/FileFormat/issues/4*](https://github.com/MPEGGroup/FileFormat/issues/4)

Note: The [GitHub issue #4](https://github.com/MPEGGroup/FileFormat/issues/4) already contains suggested solutions and should be considered in the design of the solution.

In the file containing the SegmentIndexBox, the anchor point for a SegmentIndexBox is the first byte after that box.

subsegment\_duration: when the reference is to SegmentIndexBox, this field carries the sum of the subsegment\_duration fields in that box; when the reference is to a subsegment, this field carries the difference between the earliest presentation time of any access unit of the reference stream in the next subsegment (or the first subsegment of the next segment, if this is the last subsegment of the segment, or the end presentation time of the reference stream if this is the last subsegment of the stream) and the earliest presentation time of any access unit of the reference stream in the referenced subsegment

**Comment (informative)**

anchor point for a SegmentIndexBox is the first byte after that box, does this imply that sidx comes before the media it indexes? In this case, sidx may also be no usable if the duration of a sample is unknown at the time of packaging it.

## General editing

### Presentation terms

*https://github.com/MPEGGroup/FileFormat/issues/8*

The term "presentation" is used throughout the spec (~180 times) with different meanings. This is confusing. We propose to clarify it when possible and to replace it in other cases.

* "presentation"

When used standalone, it usually means "rendering" or "a set of related media" as in the introduction:

"The ISO Base Media File Format is designed to contain timed media information for a presentation in a flexible, extensible format that facilitates interchange, management, editing, and presentation of the media. This presentation may be ‘local’ to the system containing the presentation, or may be via a network or other stream delivery mechanism."

It is currently defined as follows:

"one or more motion sequences, possibly combined with audio"

This definition is outdated.

We suggest replacing the definition with the simple:

"set of related media data "

We also suggest rephrasing the introduction which has too many 'presentation'.

* "presentation time" (60 occurrences)

There is no formal term in the definition section but the semantics of the 'tfra' box (8.8.10.3) says:

"The presentation time is the composition time of a sample, as adjusted by any edit list."

We suggest moving this text as a term definition in the definition clause.

We find also that:

"presentation times are in the movie timeline" (8.6.13.1 Segment Index Box definition)

which seems consistent with our understanding and the definition above. We suggest moving that text as a note in the definition clause.

But, we find in Annex A, A.4:

" The exact presentation time (its time-stamp) of a sample is defined by summing the durations of the preceding samples."

which is confusing "time stamp" and "time" and more importantly "decoding time stamp" and "presentation time".

We suggest fixing that sentence as follows, by replacing:

"The exact presentation time (its time-stamp) of a sample is defined by summing the durations of the preceding samples."

with:

"The exact decoding time stamp of a sample is defined by summing the durations of the preceding samples."

Sometimes the term "movie presentation time" is used. We suggest removing "movie" (or always using it) as the "presentation time" is indeed in a "movie time".

The terms are "earliest presentation time" and "end presentation time" are used but don't seem ambiguous as they do consider the movie timeline (i.e. with edit list).

The different sections about RTP use the term of "presentation time stamp" with a different meaning:

* "presentation time-stamp" (RTP Packet Entry Format, 9.1.3.2)
* "presentation timestamp" (H.3.2 Compensation for unequal starting for position of received RTP streams)
* In H.4.4 and H.5.3 (duplicate text) "A presentation time on a timeline of the receiver clock is derived for each sample. If RTCP reception hint tracks are in use, the presentation time is the composition time of the sample on the movie timeline, also including clock drift correction as described in step 3 above. If RTCP reception hint tracks are not in use, the presentation time is directly the composition time of the sample on the movie timeline."

This is wrong and should be fixed.

* "presentation order" (9 occurrences)

The term is not defined. When it is used, it is not related to the "presentation" (i.e. including edit list), it rather means: "order in which samples are with increasing composition times". Sometimes the term "output order" is used. Some other times the term "composition order" is used.

We suggest defining the term "composition order" (or "output order") as above and to use it consistently.

Similarly, we suggest defining "decoding order" and using it consistently (versus "decode order").

* "presentation file" (4 occurrences)

In 6.1.1, it is used to introduce the notion of external media data not stored in a main ISOBMFF file:

" A presentation may be contained in several files. One file contains the metadata for the whole presentation, and is formatted to this specification. This file may also contain all the media data, whereupon the presentation is self-contained. The other files, if used, are not required to be formatted to this specification; they are used to contain media data, and may also contain unused media data, or other information. This specification concerns the structure of the presentation file only. The format of the media-data files is constrained by this specification only in that the media-data in the media files must be capable of description by the metadata defined here."

Similarly, in 11.2, it says:

" The main file containing the metadata may use other files to contain media-data. These other files may contain header declarations from a variety of standards, including this one.

If such a secondary file has a metadata declaration set in it, that metadata is not part of the overall presentation. This allows small presentation files to be aggregated into a larger overall presentation by building new metadata and referencing the media-data, rather than copying it."

But in 6.1.2, the COR mixes this term with the notion of segment:

"A presentation file logically includes all its segments."

We believe this should be fixed. We suggest:

* removing the notion of presentation file
* and rewriting 6.1.1, 6.1.2 with the following text:

"When represented according to the format defined in this part of the standard, a presentation may be stored in a single file or in multiple files, or it may even be delivered without the bytes being written in a file, for instance when streamed over a network and consumed on the fly.

When split over multiple files, two different splitting options exist.

In one option, one file contains the metadata for the whole presentation, and is formatted to this specification. The other files are not required to be formatted to this specification. They are used to contain media data, and may also contain unused media data, or other information. The format of these other files is constrained by this specification only in that the media data in them must be capable of description by the metadata defined in this specification.

These other files may be ISO files, image files, or other formats. Only the media data itself, such as JPEG 2000 images, is stored in these other files; all timing and framing (position and size) information is in the ISO base media file, so the ancillary files are essentially free-format.

If an ISO file contains hint tracks, the media tracks that reference the media data from which the hints were built shall remain in the file, even if the data within them is not directly referenced by the hint tracks; after deleting all hint tracks, the entire un-hinted presentation shall remain. Note that the media tracks may, however, refer to external files for their media data.

In a second option, the media data is distributed over multiple files conformant to this specification. A first file contains some metadata valid for the whole presentation and possibly some media data and some metadata valid for a first part of the presentation. It also describes that additional files may be present. These additional files describe media and metadata for successive parts of the presentation.

In more complex scenarios, the two options could be combined.

In this specification, some boxes (called top-level boxes) are indicated as being at ‘file’ level, with the notation “Container: File”. This file corresponds to the single file when no other files are used; or when multiple files are used, to the virtual file formed by the concatenation of file containing the metadata for the first part of the presentation, with the other ISOBMFF compliant files in presentation order.

* "presentation metadata wrapper" (1 occurrence)

In "6.1.2 Object Structure", it says:

"The sequence of objects in the file shall contain exactly one presentation metadata wrapper (the MovieBox)."

We suggest replacing it with:

"The sequence of objects in the file shall contain exactly one MovieBox."

* "presentation information" (5 times)

The sentences using this term can easily be removed or the term replaced by "media information".

## Extended language tag

**Note: No consensus yet!**

[*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/169*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/169)

### Observation

From 14496-12:

The ExtendedLanguageBox represents media language information, and shall contain a code in conformance with IETF BCP 47. It is an optional peer of the MediaHeaderBox, and shall occur after the MediaHeaderBox.

However, issues and interop problems were reported on applying the rules from BCP 47.

### Request from initial contribution

At the 9th meeting of WG03 (MPEG#140) in Mainz, Germany, the DASH BoG decided to add further recommendations into their DASH WD to clarify the usage of BCP-47 compliant language codes.

These recommendations include:

1. a clarification that BCP-47 requires that 2-character codes are to be used whenever possible
2. a recommendation that other subtags except the primary language subtag should not be used unless essential for disambiguation

Since ISO 14496-12 and its “extended language tag” Box (‘elng’) carries BCP-47 compliant language tags, the ISOBMFF BoG should consider whether the ISOBMFF spec can benefit from the clarifications considered for the DASH specification.

It is recommended to align the usage of language codes in 14496-12, section 8.4.6 with revised text on the @lang attribute in ISO/IEC 23009-1:2022, including any future enhancements.

### On 2-letter codes

#### Initial Proposal

*Add the following text at the end of section 8.4.6:*

NOTE 1 Per IETF BCP 47, 2-character codes are to be used whenever possible, i. e. 3-character codes are not to be used when there is an equivalent 2-character code.

#### Discussion

More details on the underlying problem and a raised question whether this is redundant were provided:

The violation of BCP-47 in regard the use of 3-char codes when 2-char codes are defined, is clearly forbidden. Nevertheless it is rampant in practice, causing serious interop issues today. Implementors are simply not paying attention to this BCP-47 provision or applying the BCP-47 equivalence algorithm. This Note highlight is necessary and critical. We can emphasize it individually in ATSC, DVB, CTA, IETF,…… but why not clarify this informative Note broadly which causes no conflict/problem at all, and is a benefit? The fact that it is technically redundant is not helpful.

Although considered redundant, it was agreed to add Note 1 from 2.3.3.1 as it focusses readers to a requirement in BCP-47 often overlooked in applications.

Later, a comment raised the following concern and proposal:

As you may previously have discussed, because the spec already contains explicit references to the appropriate ISO well to draw from for other occurrences of language codes, such as the explicit reference to ISO 639-2/T in connection with the language field of the MediaHeaderBox, I suppose it's possible to assume that those references apply to other uses of language codes when no corresponding statement is made.

To correct such assumptions, perhaps going beyond a reiteration of material in BCP-47 by calling out the distinction between 'mdhd' and 'elng' would be even more helpful – see 2.3.3.3.

#### Alternative proposal

*Instead of Note 1, the following text could be added in proximity to the existing language regarding compatibility between indications of language in 'mdhd' and 'elng', which already includes the illustrative example of "eng" and "en-UK":*

When a two-character code is defined in ISO 639-1 for the same language as a three-character coded defined in ISO 629-2, whereas the MediaHeaderBox requires the use of the packed 3-character code, the 'elng' box, in conformance with IETF BCP 47, instead requires the use of the 2-character code.

### On subtag usage

#### Initial Proposal

*Add the following text at the end of section 8.4.6:*

~~NOTE 2~~ Other subtags except the primary language subtag, e.g. region subtags such as used in “es-US”, should not be used unless essential in language disambiguation.

*[Ed.Note: There is no consensus yet on this general provision based on the comment in* [*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/169#note\_69880*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/169#note_69880)*]*

#### Discussion

Commentators see Note 2 as ambiguous and unenforceable, because it doesn’t specify what is and what is not “essential in language disambiguation”. For example, if we have only one English language audio track that’s present in a collection of alternative tracks, and it happened to be produced by an Australian movie production company with speaking parts assigned to Australian actors, is the region code “AU” essential in its language tag or is it not, given that the language code alone is sufficient to distinguish this track from the others in the collection? Would a region code become eligible for use only later if a new English language track is produced by dubbing in dialog spoken by American actors or actors using American accents? Would we have to go back and re-author the original audio track at that point to add the region code “AU”?

Perhaps the second recommendation was added to DASH in an attempt to reduce the complexity of parsing BCP 47 language tags and matching them to language preferences. The likeliest outcome of giving implementers of the spec permission to leave the job only partly done, we think, is to decrease interoperability rather than to increase it.

The group initially concluded but later revised the decision to add the second note not as a note but as a general provision (because of the should) with an editor's note indicating that there is no consensus yet on this general provision based on the comment in [#169 (comment 69880)](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/169#note_69880).

Later, the following comments and questions were raised:

* The proposed simplification of suppressing the use of region post-fixes when an Adaptation Set has only one such language was not actually proposed by either the initial proponent or ATSC. Nevertheless others felt this was an important restriction to which we agree. Other SDOs were compelling on this, so we agree it is good advice to simplify which causes no harm and remains conformant to BCP-47. The BCP-47 requirement to implement the defined comparison algorithm is good, buy why unnecessarily complicate the decoder processing to no useful end? That is, when the presentation has only one, e.g. English, track why put the burden on the decoder to sort it out when there is really only one answer? It's just a “should”.
* The issue here was that it is not at all clear what is "essential in language disambiguation" and what is not. Which makes this statement ambiguous and unenforceable. Earlier, an example was provided  based on English audio tracks spoken in Australian vs American English which I mentioned in [our response](https://urldefense.com/v3/__http:/mpeg.expert/software/MPEG/Systems/FileFormat/isobmff/-/issues/169*note_69880__;Iw!!HOHtwYw!E5fU4ZqGVtoG5bxH9bDX3uFKOTrUvZWAVZtdXww3S-js2fdWlNmxw040FtZSrIZrZK-xpIUodJZEfJp5uSm8xA$).

Another point was also related to the complexity, for that part we would “trade” complexity for interoperability if we would give such flexibility to spec implementers. Also modern operating systems often provide APIs for handling of BCP 47.

* Regarding "essential in language disambiguation" – if you have American English and Australian English sound tracks, by all means you should state the region. If you have English, German, and Chinese the region is not that important as a person would select a language, not a regional dialect.

Similarly, script may matter for subtitles in some languages using multiple scripts (e.g., Chinese), but is irrelevant in others (such as English).

* Yes, but the issue I'm trying to highlight is that on Day 1 of content availability one might have American English only, but on Day 101 one may need to add Australian English without knowledge prior to Day 101 that such a need would eventuate. In our experience preparing content for worldwide distribution is an incremental process, with contributions from multiple actors with distinct language expertise.

In other words, there may be no specific moment at which one can know exactly what one will eventually have or need. As additions are made, anticipated or unanticipated, it's highly desirable for existing tracks that have already been vetted can be carried over for continued use in an expanded collection.

I am not suggesting that subtags that are always redundant should be included, such the use of the Latin writing system tag for English. I'm suggesting only that subtags that might eventually be required for disambiguation should not be made ineligible for inclusion by whatever specification text is eventually adopted, if any.

* I get this, and translation got easier and cheaper. If additional translations are anticipated, use en-AU by all means.

With that said, how many movies have you seen that offer various regional dialect? I have seen none (I would have appreciated one for "Trainspotting"…)

Moreover, I have never seen a track addition to an existing asset in its window. I would expect a re-validation and re-pitch at the minimum.

TLDR: in case you expect a future need to disambiguate, use the regional tags. Otherwise – it's an unnecessary complication.

* My feeling is that more attention needs to be paid to use cases in which tracks of the same media type can be accumulated incrementally. In particular, I think it's important to avoid cases in which the addition of new tracks to previously authored tracks, whether that addition takes the form of muxing the expanded collection to a new file or the introduction of a declaration of an fMP4 stream to an existing DASH or HLS manifest, requires previously existing tracks to be revised. Omitting region or writing system subtags until incremental accumulation of tracks makes them necessary would require such revision. Even if implementations are smart enough to recognize that something previously left out must be added, a prior omission of a region or writing system subtag introduces a requirement of an independent authoritative source of the newly required information. This would very likely be problematic. It's far better, I think, to write to each track as it's authored all of the language-related information that may be needed.

Should the use of region or writing system subtags remain the best practice in cases in which incrementally addition of tracks is possible — and to be clear that's the position I'm advocating, along with the supplementary observation that further addition of tracks to an existing collection is *always* possible — then I fail to perceive value in the proposed simplification

* In other words, we can leave the use of the 'elng' box up to the discretion of the content author, and I agree that that is appropriate. I can also agree that it's a good idea to keep them as simple as possible.

But I'll reiterate my last point: if the use of region or writing system subtags is ever a necessary complication, and I think we've now established that it is, then no pure simplification that can be achieved among interoperable implementations. Instead, all that can be achieved by assuming exclusive use of simple BCP 47 language tags is a deficit in interoperability.

How many movies have I seen that offer multiple tracks declaring the same primary language with distinct regional subtags? If we include movies that comprise collections of fMP4 streams: as many as I might care to look at. For example, "es-ES" and "es-419", "pt-PT" and "pt-BR", and so forth. Because we don't happen to employ a single uniform file for worldwide distribution via progressive download, I see it less often for movies comprising tracks stored in the same file, but it nevertheless does occur. I believe I've seen some TV shows distributed in Europe with "nl-NL" and "nl-BE", for example.

## Missing Semantics

*[Ed.Note: Clauses 2.4.1 to 2.4.3 are already fixed with 8th edition. These clauses are kept here for better understanding of the remaining clause 2.4.4.]*

[*http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/168*](http://mpegx.int-evry.fr/software/MPEG/Systems/FileFormat/isobmff/-/issues/168)

### ~~Observation~~

~~ISO/IEC 14496-12, 7th Edition, section 12.2.7 lacks a semantic definition for one element in the related syntax.~~

### ~~Consideration~~

~~Although a definition for the semantics can be interpolated from the semantics of other elements listed in the same clause, in this case a simple reference to ISO/IEC 23003-4 is not sufficient as this element’s counterpart in 23003-4 is a variable length field. Consequently, a mapping between the variable length field in 23003-4 and the fixed-length field in 14496-12 needs to be specified.~~

### ~~Proposal~~

*~~Add the following semantic definition to section 12.2.7.3:~~*

~~method\_value is the measured value as specified by the method\_definition, quantified as defined in~~ **~~Error! Reference source not found.~~**~~; all others are reserved. If the definition in~~ **~~Error! Reference source not found.~~** ~~utilizes less than 8 bits, the value shall be LSB-aligned with unused bits being set to ‘0’.~~

~~EXAMPLE: If methodDefinition==8 and methodValue==0x1 (”large room, X curve monitor”) then method\_value would be written as 0x1 (uimbsf).~~

### Further identified issues w.r.t. semantics

*[Ed.Note: This is a first quick analysis on this, there are likely more issues.]*

#### general comments

* Semantics for box versions are not used in consistent manner through the spec. Could probably be removed completely after clarifying in the syntax or even better to add the maximum version to the definition.
* Constant fields (fixed values elements) are not consistent, sometimes they appear in semantics sometimes are omitted. e.g.: 8.2.2.3

#### Specific observations

|  |  |  |
| --- | --- | --- |
| **Section** | | **Observation** |
| 8.1.4.3 | Identified media data box / Semantics | Semantics for data missing |
| 8.13.2 | FD item information box | Citation: “The semantics of the boxes are described where the boxes are documented.” |
| 12.1.5 | Colour information | missing element name for ICC\_profile type ICC\_profile type is not defined. |
| 12.1.6 | Content light level | missing semantics |
| 12.1.7 | Mastering display colour volume | missing semantics |
| 12.1.8 | Content colour volume | missing semantics |
| 12.1.9 | Ambient viewing environment | missing semantics |
| 12.2.3.2/3 | Audio media / Sample Entry / Syntax / Semantics | This is a bit of a mess |
| 12.3.3.3 | Metadata media / Sample Entry / Semantics | missing some of the semantics |
| 12.6.3.3 | Subtitle media / Sample Entry / Semantics | mentions Semantics element which are not define in the syntax |

## Missing track reference types

*https://github.com/MPEGGroup/FileFormat/issues/45*

Table I.3 in the AMD1 of 23003-4 refers to ISOBMFF for the definitions of track reference types 'adda' and 'adrc' as shown below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Track reference for additional audio track | *adda* | n/a | 14496-12 | 6.4.5 | a | a |
| Track reference for DRC metadata track | *adrc* | n/a | 14496-12 | 6.1.1 |  | b |

However, ISOBMFF does not define those track reference types in clauses 6.x.x. In ISOBMFF specification 'adda' is mentioned in clause 12.2.6 and it should be considered to add 'adda' entry to section 8.3.3.3 to keep all track references defined in Part 12 in one place. In addition to that a definition of 'adrc' should be clarified between ISOBMFF and 23003-4.