Text

Description automatically generatedISO/IEC JTC 1/SC 29/WG 03 N0692

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

**Document type:** Output Document

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

**Status:** Approved

**Date of document:** 2022-08-25

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

**No. of pages:** 11 (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 N0692**

**October 2022, Virtual**

|  |  |
| --- | --- |
| **Title** | **Defect report for ISO/IEC 14496-12** |
| **Source** | **WG 03, MPEG Systems** |
| **Status** | **Approved** |
| **Serial Number** | **21971** |

**Contents**

[1 Introduction 3](#_Toc96427028)

[2 Defects under consideration 3](#_Toc96427029)

[2.1 What is mandatory in a file? 3](#_Toc96427030)

[2.2 Segment index (sidx) 3](#_Toc96427034)

[2.3 Subsegment Index and Level Assignment (ssix and leva) 3](#_Toc96427035)

[2.4 General editing 3](#_Toc96427076)

[2.4.1 Presentation terms 3](#_Toc96427077)

[2.5 Brands 6](#_Toc96427078)

[2.5.1 Structural brands analysis 6](#_Toc96427079)

[2.5.2 Issues 8](#_Toc96427080)

[2.5.3 Recommendations 10](#_Toc96427081)

# 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>*

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".

## Brands

### Structural brands analysis

The following diagram presents an overview of the brands defined in ISOBMFF. The upper part indicates the brand 4CC, ‘+’ indicates that support for a box was added as required in the given brand, ‘~’ indicates some semantics changes around an existing box, and the bottom sentences indicate additional support.

Graphical user interface, application

Description automatically generated

The full list of supported boxes in ‘isom’ is omitted.

A few notes on the figure and the derived specs:

* 14496-14 defines ‘mp41’ and ‘mp42’ but without relationships to the ‘isoX’ brands
* 14496-15 defines brands which are specific to layered HEVC (‘hvce’ and ‘hvci’) and omitted here.
* HEIF defines 2 structural brands
  + ‘mif1’ intersects ‘isoa’, ’iso7’, and ‘iso2’, but not the other ‘isoX’ brands (intersection not represented on the picture) for the following reasons:
    - The ‘meta’ box (defined in ISOBMFF and supported from ‘iso2’) has to be present and supported, but given that the ‘moov’ box is not necessarily present and is not required to be parsed, ‘iso2’ is not a subset of ‘mif1’.
    - A parser has to support ‘iloc’ v2, ‘iinf’ v1, ‘inf2’ v3, ‘iref’ v1 which are only permitted in ‘iso7’, but not all the features of ‘iso7’ are required to be supported.
    - A parser has to support ‘iprp’ only defined in ‘isoa’ but all the features of ‘iso8’ are not required to be supported.
  + ‘msf1’ requires full support for ‘iso8’, but adds required support for ‘pict’ tracks (and the ‘ccst’ box) and edit list repetition so it is a strict superset of ‘iso8’
* MIAF defines on structural brand ‘miaf’
  + A reader is required to support ‘mif1’ and ‘msf1’, so a mix of ‘iso8’ + some tools from ‘isoa’. It is difficult to represent it in the figure.
* CMAF defines 2 structural brands
  + ‘cmfc’ and ‘cmf2’
  + It leaves the choice to the writer to write an ‘isoX’ brand, but clearly if used, a file should not declare less than ‘iso6’ given that ‘tfdt’ must be in the file and has to be processed. It should be ‘iso8’ when subtitles tracks are used because of the presence of ‘sthd’ and ‘iso9’ if ‘elng’ is used.
  + However, not all tools of ‘iso6’, ‘iso8’ or ‘iso9’ have to be supported by readers so ‘cmfc’ and ‘cmf2’ are not supersets of ‘isoX’ brands and the precise intersection is not represented in the figure.
* MPEG-7 and MPEG-21 brands are not represented
* OMAF brands are not represented here

### Issues

While reviewing the brands, we found the following aspects that should be clarified:

#### Missing boxes

Even with all these brands defined, support for some boxes is not mentioned in any brands, e.g.:

* imda, imdt, snim
* mfra, mfro
* leva
* csgp
* kind
* strk
* …

We suggest cross-checking that no box is missing from the isoX tables and adding a section to explicitly list the boxes that are not required in any of the current brands.

#### Versions and flags

There is no specific mention of the version support for the following boxes

* ‘mvhd’ v0, v1
* ‘tkhd’ v0, v1
* ‘mdhd’ v0 v1
* ‘elst’ v0, v1
* ‘subs’ v0, v1
* ‘saio’ v0, v1
* ‘mehd’ v0, v1
* ‘tfra’ v0, v1
* ‘tfdt’ v0 v1
* ‘assp' v0 v1
* ‘sbgp’ v0 v1
* ‘sgpd’ v0 v1 v2
* ‘sidx’ v0 v1
* ‘prft' v0 v1

Similarly, support for specific flag values that affect the parsing of the following boxes is not indicated

* ‘saiz’
* ‘saio’
* ‘trun’
* ‘schm’

We suggest clarifying the definitions of brands to indicate the versions of the boxes (as done in HEIF) and the values of the flags that have to be supported for each brand.

#### Sample groups support

The definition of ‘iso3’ says:

“Within the sample groups, support for rate share information (grouping type ‘rash’) is required.”

The definition of ‘iso6’ says:

“Within the sample groups, support for random access point information (grouping type ‘rap ’) is required.”

It is unclear what requiring support for a sample group means (given that a file can always be processed ignoring the sample groups), and how this can be verified in a conformance program.

#### Hint track support

The definition of ‘iso3’ says:

“File delivery hint tracks (sample entry ‘fdp ’) must be recognized.”

It is not clear what it means given that hint tracks can be ignored in general. Section E.1 says:

“In general, readers are required to implement all features documented for a brand unless one of the following applies:

c) the context in which the product operates means that some structures are not relevant; for example, hint track structures are only relevant to products preparing content for, or performing, file delivery (such as streaming) for the protocol in the hint track.”

It is suggested to rephrase the requirement in ‘iso3’ along the lines of Annex E.1.

#### 32 bits in meta

‘Iso7’ indicates:

“Support for 32-bit item\_ID and item\_count values in MetaBox”

It is unclear that it applies to boxes in the MetaBox hierarchy. Also some boxes are not covered by ‘item\_ID’ and ‘item\_count’ like ‘iinf’. We suggest listing explicitly the versions of the boxes that are required to be supported: iloc v2, pitm v2, infe v3, iref v1, iinf v1

#### Recognize tracks?

‘iso7’ indicates

“Support for the following is required under this brand:

* Recognizing incomplete tracks.”

What is the meaning of “recognizing” in this case?

### Recommendations

We recommend updating the ISOBMFF specification (and possibly its derived specifications) to address the issues identified in this document.

We also recommend defining brands more precisely in ISOBMFF, following the way it is defined in HEIF as follows:

* File requirement:
  + What box must be present, which version, which flags (the ‘claim’ part of the brand definition).
    - It should be made clear that a file using one of these mandatory boxes with a version or flag value that is not explicitly permitted is invalid.
  + Other boxes may be present (file is still valid) and if present, can be ignored (the ‘permission’ part of the brand definition)
  + We should create invalid files against these requirements
* Reader requirement
  + ISOBMFF is loose wrt to reader requirements. We should limit to:
    - List the boxes that a reader shall be able to parse (interpretation of the box data is left non normative)