PDF 2.0 Application Note 003: Clarification on locations for object metadata streams

PDF Association
Friedenstr. 2A · 16321 Bernau bei Berlin · Germany
www.pdfa.org
E-mail: copyright@pdfa.org
Web: www.pdfa.org

Published in Germany and the United States of America

Table of Contents

Preface

In most cases metadata is all about interoperability: It is only useful if creator and processor agree on syntax and semantics. In the context of PDF this also means that creator and processor must agree on where metadata is present in the often-complex internal PDF structure.

Metadata locations are well defined for XMP metadata on the PDF document level and for some PDF objects, but not for all objects where metadata can occur (e.g. pages, page level objects and resources such as fonts or ICC profiles). However, in almost all cases sufficient guidance can be derived from the PDF specification, albeit sometimes in an indirect way.

The principal authors of this document are Boris Doubrov (Dual Lab) and Dietrich von Seggern (callas software).

Scope

This document summarizes exactly where PDF creators should add XMP metadata to PDF page level objects or resources, and thus where PDF processors should search for it, for all common use cases.

References

ISO 32000-2: 2020 (PDF 2.0), https://www.iso.org/standard/75839.html

Explicit mention of Metadata streams in PDF 2.0

PDF 2.0 defines a generic **Metadata** key that may be present in any dictionary (including stream dictionaries) and whose contents is a metadata stream in XMP syntax. PDF 2.0 explicitly specifies this entry for certain types of objects correlated with typical use cases for associating metadata with PDF objects. Table 1 lists all such cases in ISO 32000-2.

Table 1: Explicit **Metadata** usage in PDF 2.0

Object type	ISO 32000-2 reference	Description
Catalog	Table 29	(*Optional; PDF 1.4; shall be an indirect reference*) A metadata stream that shall contain metadata for the document.
Page	Table 31	(*Optional; PDF 1.4*) A metadata stream that shall contain metadata for the page.
ICC Profile stream	Table 65	(*Optional; PDF 1.4*) A metadata stream that shall contain metadata for the colour space.
Image XObject	Table 87	*(Optional; PDF 1.4)* A metadata stream containing metadata for the image.

Object type	ISO 32000-2 reference	Description
Form XObject	Table 93	(*Optional; PDF 1.4*) A metadata stream containing metadata for the form XObject.
Document Thread	Table 162	(*Optional; PDF 2.0; shall be an indirect reference*) A metadata stream containing information about the thread, such as its title, author, and creation date.
Document Part (DPart)	Table 409	(*Optional; PDF 2.0; shall be an indirect reference*) A metadata stream that shall contain metadata for this document part.

Notes on metadata locations

Document level metadata

Document level metadata is referenced by the **Metadata** key in the document catalog dictionary (first row in Table 1). Even though this key is formally optional, an entry is mandatory in all PDF subset specifications (e.g. PDF/A, PDF/X, PDF/E and PDF/VT). As the Info dictionary is deprecated in PDF 2.0, the document catalog dictionary is also the only recommended place in PDF 2.0 documents for specifying the document title, author, producer and other data previously stored in the Info dictionary.

Image XObjects

The JPXDecode filter (ISO 32000-2, 7.4.9) recommends the use of Image metadata for ensuring interoperability of relevant JPX format information such as layering, composition instructions, simple animation.

Document Parts and Document Parts Metadata

The XMP metadata stream referenced by the **Metadata** key in the Document Part dictionary should not be confused with Document Part Metadata (ISO 32000-2, 14.12.4.2), referenced from the same Document Part dictionary via a **DPM** key.

The main purpose of the DPart tree structure is to specify metadata for pages or page ranges using DPM entries. Usually all DPM entries in a PDF file will use the same vocabulary to communicate information about document parts to a downstream production workflow. DPM could, for example, associate addresses for a mass mailing with specific pages for use as input to the mailing system.

XMP metadata may use the same DPart tree to associate pages with metadata, but the metadata will then usually be less structured and the DPart structure might not cover all pages in a PDF file.

Implicit mention of Metadata in PDF 2.0

As noted in Table 1, an optional "Metadata" entry is defined for several dictionaries apart from the document catalog dictionary. Indeed, the specification actually goes much further; ISO 32000-2, 14.3.2 "Metadata streams" states that:

> *"…any PDF stream or dictionary may have metadata attached to it as long as the stream or dictionary represents an actual information resource."*

Accordingly, PDF 2.0 processors must be prepared to encounter metadata almost anywhere in PDF files.

In order for metadata to be useful a common understanding between creator and processor regarding the metadata's location is required. However, since some objects in PDF distribute related information over several interlinked entries the appropriate location is not always clear. ISO 32000-2, 14.3.2 provides some guidance for these cases:

> *"When there is ambiguity about exactly which stream or dictionary may bear the Metadata entry, the metadata shall be attached as closely as possible to the object that actually stores the data resource described."*

In alignment with this advice PDF 2.0 identifies objects that may be expected to include embedded metadata streams even while their PDF 2.0 definitions lack the optional "Metadata" entry noted in Table 1. These objects are:

- Embedded fonts (14.3.2 "Metadata streams", see Note 1)
- Tiling patterns (14.3.2 "Metadata streams", see Note 1)
- Shading patterns (14.3.2 "Metadata streams", see Note 1)
- Marked content (14.3.2 "Metadata streams", see paragraph 5)
- Structure elements (14.8.4.3 "Document level structure types", see paragraph 5)

The following clauses clarify what "the metadata shall be attached as closely as possible to the object that actually stores the data resource" means as applied to each of these object types.

Embedded fonts

For embedded font streams PDF 2.0 is pretty clear in the (non-normative) Note 1 in 14.3.2 "Metadata streams":

> *"…metadata for fonts needs to be attached to font file streams rather than to font dictionaries."*

This text means that any metadata associated with an embedded font program must be present in a Metadata entry in the corresponding embedded font stream dictionary (as opposed to other PDF entries for the same font, e.g. the Font Descriptor dictionary).

Embedded font stream dictionaries are defined in ISO 32000-2, Table 125 which may be understood as having an additional row as indicated in Table 2.

Table 2: Metadata for font streams

Metadata	stream	(*Optional; shall be an indirect reference*) A metadata stream that shall contain metadata for the font program (see 14.3.2, "Metadata streams").

Tiling patterns (Type 1 pattern dictionaries)

For tiling patterns PDF 2.0 is pretty clear in the (non-normative) Note 1 in 14.3.2 "Metadata streams":

> *"…metadata describing a tiling pattern needs to be attached to the pattern stream's dictionary."*

Since a tiling pattern only uses a single stream object as defined in ISO 32000-2, Table 74 this table may be understood as having an additional row as indicated in Table 3.

Table 3: Metadata for tiling patterns

Metadata	stream	(*Optional; shall be an indirect reference*) A metadata stream that shall contain metadata for the tiling pattern (see 14.3.2, "Metadata streams").

Shading patterns (Type 2 pattern dictionaries)

For shading patterns PDF 2.0 is pretty clear in the (non-normative) Note 1 in ISO 32000-2, 14.3.2 "Metadata streams":

> *"…a shading needs to have metadata attached to the shading dictionary rather than to the shading pattern dictionary that refers to it."*

Several types of shading dictionaries exist in PDF 2.0; the entries common to all types are defined in ISO 32000-2, Table 77. This table may be understood as to have an additional row as indicated in Table 4.

Table 4: Metadata for shading patterns

Metadata	stream	(*Optional; shall be an indirect reference*) A metadata stream that shall contain metadata for the shading pattern (see 14.3.2, "Metadata streams").

Marked content

PDF 2.0 specifies in ISO 32000-2, 14.3.2. “Metadata streams” as follows:

> *“Object level metadata can also be associated with marked content within a content stream, by including a metadata stream in the property list dictionary for this marked content. Because a stream dictionary is always an indirect object, a property list containing a metadata stream cannot be encoded inline in the content stream, but needs to be encoded as a named resource (see 14.6.2, “Property lists”).”*

According to the Note in ISO 32000-2, 14.6.2, property lists in marked content are…

> *“used by several PDF features, including optional content … , tagged PDF … and Associated Files …”*

Applying PDF 2.0’s general requirement for object level metadata; that it “shall be attached as closely as possible to the object that actually stores the data resource” (as cited above from 14.3.2), the Note should be understood as meaning that:

- Metadata for optional content (layers) should be located in the respective optional content group dictionary (ISO 32000-2, Table 96) in order to be available for all marked content sequences associated with that group.
- Metadata for structure elements should be located in the corresponding structure element dictionary (see Structure elements).
- Metadata for associated files should be located in the respective file specification dictionary (ISO 32000-2, Table 43) or - for embedded file streams - in the respective embedded file stream dictionary (ISO 32000-2, Table 44).

Remaining use cases for object level metadata in the context of marked content are association with content sequences that are either not tagged at all or include portions of content that are not included in the tagging structure. For this use case the content must be marked as described in ISO 32000-2, 14.6 “Marked content” using DP or BDC operators:

```
/SomeTag /SomePropertyList BDC
```

The property is then mapped to the corresponding properties list dictionary via an entry in the Properties subdictionary of the current resource dictionary. (The metadata stream has to be an indirect object and therefore the properties list cannot be an inline object.) In this example the Properties subdictionary could be:

```
/Properties << / SomePropertyList 10 0 R >>
```

In the following example object 10 0 is the property list dictionary and may then reference the actual metadata stream object via a Metadata entry, e.g.

```
10 0 obj <<
/Metadata 11 0 R
>>
endobj
```

Structure elements

In ISO 32000-2, 14.8.4.3 “Document level structure types” PDF 2.0 clearly highlights the use of metadata for certain structure element types:

> *“An XMP metadata stream (see 14.3.2, “Metadata streams”) in a Document or DocumentFragment structure element may be used to include document metadata for a logical document nested inside a tagged PDF.”*

Due to the very open concept of PDF to allow metadata in every stream or dictionary “as long as the stream or dictionary represents an actual information resource” the fact that the specification only highlights document-level semantics for this purpose should not be understood as limiting metadata usage to document level structure types.

Although the application of metadata to document-like structure elements is obvious the application to lower-level elements is equally feasible. If a process requires the association of metadata to any structure element the same approach as specified for document level structure types should be used.

Structure element dictionaries are defined in Table 355. This table may be understood as to have an additional row as indicated in Table 5.

Table 5: Metadata for structure elements

Metadata	stream	(*Optional usually only used for document level structure elements, see 14.8.4.3 “Document level structure types”; shall be an indirect reference*) A metadata stream that shall contain metadata for the structure element (see 14.3.2, “Metadata streams”).

Other potential uses of Metadata streams

Embedded files

A **Metadata** entry in an Embedded File Specification dictionary (ISO 32000-2, Table 44) can be used to specify additional information about the embedded file. It can complement the information stored in the embedded file parameter dictionary and can be used as an extension of the **AFRelationship** entry to provide extra information specifying how the embedded file is related to the main document.

Type 3 fonts

Some PDF processors implementing HTML to PDF conversion may transform web fonts in SVG or WOFF format to Type 3 fonts in PDF. The metadata associated with such fonts should be stored under a **Metadata** entry directly in the Type 3 font dictionary (ISO 32000-2, Table 110). Note that this approach is different to the metadata associated with an embedded font program, as discussed above.

A typical example of metadata stored in WOFF format is given in https://www.w3.org/TR/2012/REC-WOFF-20121213/#Metadata. Note that this data must be converted to an equivalent XMP syntax for embedding into PDF.

Annotations

Markup, Stamp or other annotation types may store additional properties in the Metadata entry of the annotation dictionary (ISO 32000-2, Table 166). This metadata might include additional information on the person or process creating the annotations and other supplementary data.

Optional content

Any optional content group may include associated metadata. In this case the corresponding XMP metadata stream shall be referenced by the **Metadata** entry in the Optional Content Group dictionary (ISO 32000-2, Table 96).

An example of such Metadata - albeit not in XMP - is ISO 19593 “Processing Steps” that defines metadata for OCGs specifying post print production steps such as diecut lines. Such metadata structures are likewise located in the Optional Content Group dictionary.

Rich media

Rich media such as video and audio data as well as 3D models can be embedded into PDF via Annotations of type RichMedia. All digital assets associated with this annotation are referenced by the **Assets** entry in the RichMediaContent dictionary (ISO 32000-2, Table 341) and form the tree of the embedded file specifications. In this case each asset metadata shall be referenced by the **Metadata** entry in the corresponding file specification dictionary. See “Embedded files” case discussed above.

3D artwork

3D artwork can be embedded into a PDF document as an Annotation of type 3D. In this case the actual data stream of the 3D model (see 13.6.3 “3D Streams”) is located under the key 3DD in the annotation dictionary (see ISO 320000-2, Table 309). The associated metadata for this 3D artwork must be referenced by the Metadata entry in the corresponding 3D stream dictionary (ISO 32000-2, Table 311). Note that, contrary to RichMedia annotations, the 3D streams of 3D annotations are not Embedded Files. Thus, 3D is not a special case of embedded file metadata.