PDF/A Competence Center

TechNote 0003: Metadata in PDF/A-1

PDF/A-1 imposes certain requirements and restrictions on document metadata in a compliant PDF/A-1 file where certain information is required to be encoded as XMP metadata, while corresponding entries in the document information dictionary for a document may be present but must match information present in the document's XMP-encoded metadata.

Note **Due to an oversight before the publication of ISO 19005-1 [1] the provisions regarding XMP metadata and document information entries were not worded correctly. This is addressed by Technical Corrigendum 1 [2] which has been released by ISO in April 2007. The explanations offered in this TechNote are based on the PDF/A-1 publication (ISO 19005-1:2005) as amended by Technical Corrigendum 1.**

1 Metadata in PDF 1.4

1.1 Document Information Entries

In early PDF versions a document information dictionary (denoted by the **Info** entry in the Trailer dictionary of a PDF file) was intended to carry information about the PDF. This dictionary is not required by in PDF 1.4, but Adobe Acrobat seems to always create the document information dictionary if it's not present, whenever a PDF is saved.

PDF 1.4 specifies the following entries in the document information dictionary:

```
Title, Author, Subject, Keywords,
Creator, Producer, CreationDate, ModDate, Trapped
```

However, it neither strictly regulates whether and how these entries are to be used, nor prohibits the presence of other entries in the document information dictionary. Syntactically it's even possible to store arbitrary data structures – data types other than string, or even dictionaries and arrays – inside the Info dictionary, although the PDF 1.4 reference advises against storing private content or structural information in it.

It is important to understand that the PDF data type **text string**, which is used for most document information entries, is specified such that it either contains text encoded using **PDFDocEncoding**, or as Unicode. If it is encoded in Unicode (more precisely: big-endian UTF-16) the first two bytes must be the Unicode byte order mark U+FEFF, and the remainder of the string consists of Unicode character codes according to the UTF-16 format.

1.2 Document XMP Metadata

Beginning with PDF 1.4 an optional entry named **Metadata** in the Catalog (or root) object was introduced. This entry is a stream object where the stream contains metadata encoded in the XMP format which has been introduced by Adobe in 2001 with the release of Acrobat 5.0. The idea is that the XMP encoded document **Metadata** in the **Catalog** object will obsolete metadata stored in the **Info** dictionary.

The details of XMP metadata are described in the XMP specification. XMP is built on the Resource Description Framework (RDF), which in turn is based on XML.

1.3 Component-Level XMP Metadata

Besides the document-level **Metadata** entry in the **Catalog** object, metadata entries are also allowed for components in a PDF file, such as pages, form and image XObjects, embedded font dictionaries, or ICC stream dictionaries.

2 Document XMP Metadata in PDF/A-1

Note **PDF/A-1 is based on specific versions of the PDF Reference (PDF 1.4) as well as the XMP specification (January 2004 version). In the meantime newer specifications for PDF and XMP have been published. Nevertheless, in order to determine the conformance of a PDF/A-1 file the exact versions of the specifications have to be used which are referenced in PDF/A-1.**

PDF/A-1 requires the presence of the **Metadata** entry in the **Catalog** object. It must at least contain PDF/A version and conformance level identification. While PDF/A-1 does not require the presence of any other information in the document's XMP Metadata, it does define how such information must be stored if present: either by making use of a predefined XMP schema or by including a custom (extension) schema.

Component-level metadata is not required in PDF/A-1. However, PDF/A-1 recommends that embedded font streams should include a Metadata entry.

2.1 Filtering

PDF/A-1 requires that all XMP metadata streams must be included without any filter (compression filters like FlateDecode, or ASCII filters like ASCIIHexDecode) applied to them to make retrieval of the metadata easier.

2.2 PDF/A Identification Schema

The only mandatory XMP entries are those which indicate that the file is a PDF/A-1 file and its conformance level. The table below summarizes the PDF/A identification schema. See PDF/A Competence Center TechNote 0001 [8] for an in-depth discussion of namespaces in XMP. The namespace prefix is required.

Schema name and description	namespace URI	required namespace prefix
PDF/A identification schema	http://www.aiim.org/pdfa/ns/id/[1]	pdfaid

1. This namespace URI is incorrectly described in ISO 19005-1, and has been corrected in [2].

2.3 Predefined XMP Schemas

The table below summarizes the schemas described in the XMP specification. These schemas together form the set of predefined schemas for PDF/A-1. The table lists preferred namespace prefixes. However, the prefixes can freely be chosen.

Note **The PDF eXtension schema with the preferred namespace prefix pdfx (not related to the PDF/X standard) is not a predefined schema in XMP. It must therefore be treated as an extension schema. (Adobe applications use this schema for non-standard document information entries).**

Schema name and description	namespace URI	preferred namespace prefix
Dublin Core schema	http://purl.org/dc/elements/1.1/	dc
XMP Basic schema	http://ns.adobe.com/xap/1.0/	xmp
XMP Rights Management schema	http://ns.adobe.com/xap/1.0/rights/	xmpRights
XMP Media Management schema	http://ns.adobe.com/xap/1.0/mm/	xmpMM
XMP Basic Job Ticket schema	http://ns.adobe.com/xap/1.0/bj	xmpBJ
XMP Paged-Text schema	http://ns.adobe.com/xap/1.0/t/pg/	xmpTPg
Adobe PDF schema	http://ns.adobe.com/pdf/1.3/	pdf
Photoshop schema	http://ns.adobe.com/photoshop/1.0/	photoshop
EXIF schema for TIFF properties	http://ns.adobe.com/tiff/1.0/	tiff
EXIF schema for EXIF-specific properties	http://ns.adobe.com/exif/1.0/	exif

2.4 Extension Schemas

Any schema outside the set of predefined schemas is called an extension schema, and must be included in the XMP metadata entry by means of the extension schema container schema. PDF/A-1 defines how such extension schemas are to be specified. The need for an extension schema may arise where industry- or customer-specific metadata needs are not or not well covered by the predefined schemas. The table below summarizes the PDF/A extension schema container schema. The namespace prefix is required.

schema name and description	namespace URI	required namespace prefix
PDF/A extension schema container schema[1]	http://www.aiim.org/pdfa/ns/extension/	pdfaExtension
PDF/A field type schema	http://www.aiim.org/pdfa/ns/field#[1]	pdfaField
PDF/A property value type	http://www.aiim.org/pdfa/ns/property#[1]	pdfaProperty
PDF/A schema value type	http://www.aiim.org/pdfa/ns/schema#[2]	pdfaSchema
PDF/A ValueType value type	http://www.aiim.org/pdfa/ns/type#[1]	pdfaType

1. The description of this schema is missing in ISO 19005-1, and has been added in [2].

2. This namespace URI is incorrectly described in ISO 19005-1, and has been corrected in [2].

3 Document Information Entries in PDF/A-1

PDF/A-1 does not require a conforming document to contain any entries in the document information dictionary at all. Nevertheless, whenever those Info entries specified in the PDF 1.4 reference (except for the **Trapped** entry) are present, there must be an equivalent entry in the document's Metadata, and both must match according to the provisions of PDF/A-1.

3.1 Encoding Issues

In XMP the default encoding for text is UTF-8, while other Unicode encodings are allowed. The June 2005 version of the XMP specification requires that the encoding for XMP must always be UTF-8. While this is not a mandatory requirement in PDF/A-1 (which references the January 2004 version of the XMP specification), it is recommended to use UTF-8 for improved forward compatibility.

3.2 Data Type Mapping and Equivalence

The requirements in the next section use the concept of equivalence between PDF objects and XMP entries. The definition of equivalence depends on the respective data types:

- For properties that map from the PDF data type text string to one of the XMP types Text, ProperName, or AgentName, value equivalence shall be on a character-by-character basis, independent of encoding, comparing the numeric ISO/IEC 10646-1 code points for the characters.
- For properties that map between the PDF data type **date** and the XMP type **Date**, defined by Date and Time Formats [5], value equivalence shall be on a component-by-component basis, relative to Coordinated Universal Time (UTC), i.e. correcting for local time zone offset.
  The XMP type **Date** supports six levels of granularity with increasing accuracy, while the PDF data type **date** supports only a single level of granularity (which includes seconds, but not fractions of a second. It is therefore recommended to always include the time in XMP properties of this type (not only the date).

3.3 Requirements for Document Information Entries

Title

If the **Title** entry (of PDF data type text string) is present in the document information dictionary, the document's Metadata must contain an equivalent entry

dc:title["x-default"] of XMP data type Text where the prefix **dc** refers to the Dublin Core schema (the prefix is actually arbitrary).

Author

If the **Author** entry (of PDF data type text string) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **dc:creator** of XMP data type seq ProperName where **dc** refers to the Dublin Core schema (the prefix is actually arbitrary). While **dc:creator** is defined as a seq type in XMP, in PDF/A-1 this sequence must contain exactly one entry.

Subject

If the **Subject** entry (of PDF data type text string) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **dc:description["x-default"]** of XMP data type Text where **dc** refers to the Dublin Core schema (the prefix is actually arbitrary).

Keywords

If the **Keywords** entry (of PDF data type text string) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **pdf:Keywords** of XMP data type Text where **pdf** refers to the Adobe PDF schema (the prefix is actually arbitrary).

Creator

If the **Creator** entry (of PDF data type text string) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **xmp:CreatorTool** of XMP data type AgentName where **xmp** refers to the XMP Basic schema (the prefix is actually arbitrary).

Producer

If the **Producer** entry (of PDF data type text string) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **pdf:Producer** of XMP data type AgentName where **pdf** refers to the Adobe PDF schema (the prefix is actually arbitrary).

CreationDate

If the **CreationDate** entry (of PDF data type date) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **xmp:CreateDate** of XMP data type Date where **xmp** refers to the Basic XMP schema (the prefix is actually arbitrary).

ModDate

If the **ModDate** entry (of PDF data type date) is present in the document information dictionary, the document's Metadata must contain an equivalent entry **xmp:ModifyDate** of XMP data type Date where **xmp** refers to the Basic XMP schema (the prefix is actually arbitrary).

Other Entries

If entries other than those listed above are present in a document information dictionary, PDF/A-1 does not impose any provisions on these. This implies that no provisions are imposed on the **Trapped** entry although it is specified in PDF 1.4.

3.4 Summary of Document Information Entry Requirements

The table below summarizes PDF/A-1 requirements for document information entries. For each standard entry in the document information dictionary the corresponding XMP property is listed in which the entry must be mirrored. Preferred namespace prefixes are listed in the table. However, the prefixes can freely be chosen.

Document info entries	corresponding XMP property	XMP data type
Title	dc:title["x-default"][1]	Text[2]
Author	dc:creator	seq ProperName[2]
Subject	dc:description["x-default"][2]	Text[2]
Keywords	pdf:Keywords	Text
Creator	xmp:CreatorTool	AgentName[2]
Producer	pdf:Producer	AgentName[2]
CreationDate	xmp:CreateDate	Date
ModDate	xmp:ModifyDate	Date
Other entries (including Trapped)	shall not be embedded using any predefined XMP schema property	n/a

1. The qualifier "x-default" is missing in ISO 19005-1 as well as in Technical Corrigendum 1 [2]. However, it is required for mapping one of the language alternatives to the Text type. Another reason for using x-default is that Adobe Acrobat 8 and older seem to ignore non-default language alternatives.

2. This entry is incorrectly described in ISO 19005-1, and has been corrected in Technical Corrigendum 1 [2].

Bibliography

[1] ISO 19005-1: Document management — Electronic document file format for long-term preservation (PDF/A-1) — Part 1: Use of PDF 1.4 (PDF/A-1)
www.iso.ch

[2] ISO 19005-1: Document management — Electronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1). Technical Corrigendum 1
www.iso.ch

[3] PDF Reference: Adobe Portable Document Format, Version 1.4, Adobe Systems Incorporated – 3rd ed. (ISBN 0-201-75839-3).
www.aiim.org/documents/standards/PDFreference.pdf

[4] XMP Specification, January 2004, Adobe Systems Incorporated.
www.aiim.org/documents/standards/xmpspecification.pdf

[5] Date and Time Formats
www.w3.org/TR/NOTE-datetime

[6] ISO/IEC 10646-1, Information technology — Universal Multiple-Octet Coded Character Set (UCS)
www.iso.ch

[7] The Unicode Standard, Unicode Consortium
www.unicode.org

[8] TechNote 0001: PDF/A-1 and Namespaces, PDF/A Competence Center
www.pdfa.org/doku.php?id=pdfa:en:techdoc

Copyright and Usage

Status of this Document

2007-03-06 First released version

2007-04-16 Re-released without any changes in content to fix problems in the PDF document's XMP metadata caused by two bugs in Acrobat Distiller 8

2008-03-14 Update:

- Updated formatting and added references to the entries in the bibliography
- Integrated proper references to Technical Corrigendum 1
- Section 3.2: recommends to always include time in XMP dates
- Section 3.4: clarified type mappings for info entries Title and Subject; removed table footnote 3
- Bibliography: added reference to W3C Date and Time specification