PDF Association Cheat Sheet – PDF Basics

PDF Terminology

ISO 32000-2:2020

- ISO — Standards Organization(s)
- 32000 — Unique Document Number
- -2 — Part Number (optional)
- 2020 — Year of Publication

PDF 2.0	ISO 32000	Latest core PDF specification that applies to all PDF files. Fully vendor-neutral specification of every non-obsoleted PDF feature since PDF 1.0.
PDF/A	ISO 19005	**A**rchival. PDF for the long-term preservation of static page appearance.
PDF/X	ISO 15930	e**X**change. PDF for graphic arts and professional printing workflows including packaging and labelling.
PDF/UA	ISO 14289	**U**niversal **A**ccessibility. PDF supporting accessibility and assistive technology such as screen readers for those with vision impairments.
PDF/R	ISO 23504	**R**aster. Simplified PDF that uses banded images and that is easy to create by low-resource embedded devices such as consumer flatbed scanners. ISO version of **PDF/raster**.
PDF/VT	ISO 16612	**V**ariable/**T**ransactional for high-speed variable data (VDP) and transactional printing in the graphic arts and professional printing industries. Builds on **PDF/X**.
PDF/VCR	ISO 16613	**V**ariable **C**ontent **R**eplacement. Templatized **PDF/X** supporting late-stage variable content merging, such as adding batch numbers on pharmaceutical packaging.
PDF/E	ISO 24517	**E**ngineering. PDF 1.6 based subset to support engineering-centric 3D PDF workflows such as aerospace and automotive engineering. Superseded by **PDF/A-4**.

PDF/raster
PDF/X-5pg
PDF/A-4f

A single PDF can conform to multiple subsets and conformance levels.

PDF subset
Uppercase letters = ISO
Lowercase = Industry

Subset Conformance Level(s)
(optional, lowercase letters)

Version of subset
(optional)

IANA Media Types

`application/pdf`	Registered Media Type for all PDF files. See RFC 8118.
`application/fdf`	Registered Media Tpe for FDF (Forms Data Field) files. See ISO 32000 for FDF file specification.
`application/xfdf`	Registered Media Tpe for XFDF (XML Forms Data Field) files. See ISO 19444-1 for XFDF file format specification.

Key PDF features by version

PDF 1.0	1993	Pure text. Only RGB. Resolution independent graphics.
PDF 1.1	1996	Device-independent color. Article threads. External links. Security. Multimedia. Actions.
PDF 1.2	1996	Prepress features: CMYK, spot color, halftoning, overprinting, OPI (Open Prepress Interface). Flate compression. New types of annotations. Interactive forms.
PDF 1.3	2000	DeviceN color. 2-byte CID fonts. Smooth shadings. More types of annotations. Large media. Page labels. Digital signatures. JavaScript. Alternate images. Masked images.
PDF 1.4	2001	Transparency and blend modes. Improved security. More prepress features. JBIG2 images.
PDF 1.5	2003	JPEG 2000. Layers (optional content). Tagged PDF. Object streams and cross reference streams for better compression.
PDF 1.6	2004	OpenType. Ultra-large media. Watermarking. Visibility expressions. AES encryption. Interactive 3D with U3D. Measurement properties.
PDF 1.7	2006	Portable collections (packages). 3D enhancements. Redaction annotations. Standardized as ISO 32000-1:2008.
PDF 2.0	2017 2020	ISO 32000-2. UTF-8 strings. 256-bit AES-CBC encryption. Unicode passwords. Black point compensation. Rich media annotations. PAdES. PRC for 3D. Geospatial features. Document parts. Associated Files. Metadata streams. Deprecation of older encryption and other legacy features, including XFA.
PDF 2.0 Extensions		256-bit AES-GCM encryption. Extensions to hash algorithms, elliptical curves, integrity protection via MACs. STEP and glTF for 3D. Clarification on PDF 1.7 and PDF 2.0 namespace inclusion.

Common terms for PDF features

Bookmarks	Outlines which use Actions or Destinations.
Comments	Markup annotations.
Compression	Filters on streams. Object streams. Cross reference streams.
Fast Web View	Linearization.
Files	Embedded Files (incl. media) and File Attachment annotations.
Forms	Widget annotations and Fields. Also referred to as **AcroForm**.
Hyperlinks	Link annotations, URI actions. Actions and Destinations.
JavaScript (JS)	ECMAScript for PDF (ISO 21757). ECMAScript Actions.
Layers	Optional Content (OC), Visibility Expressions. Marked Content.
Multimedia	3D, Movie, Screen, and RichMedia annotations with Actions.
Page size	The page **MediaBox**.
Portfolios	Also called Collections or Packages. A collection of embedded files.
Properties	Document Information dictionary and XMP Metadata streams.
Scanned PDF	Images of content such as from scanner or camera. Often has OCR-ed invisible text on top of images allowing text selection
Security	Encryption, Crypt filters, and Digital Signatures.
Tags	Tagged PDF, including Marked Content and Logical Structure.

Glossary

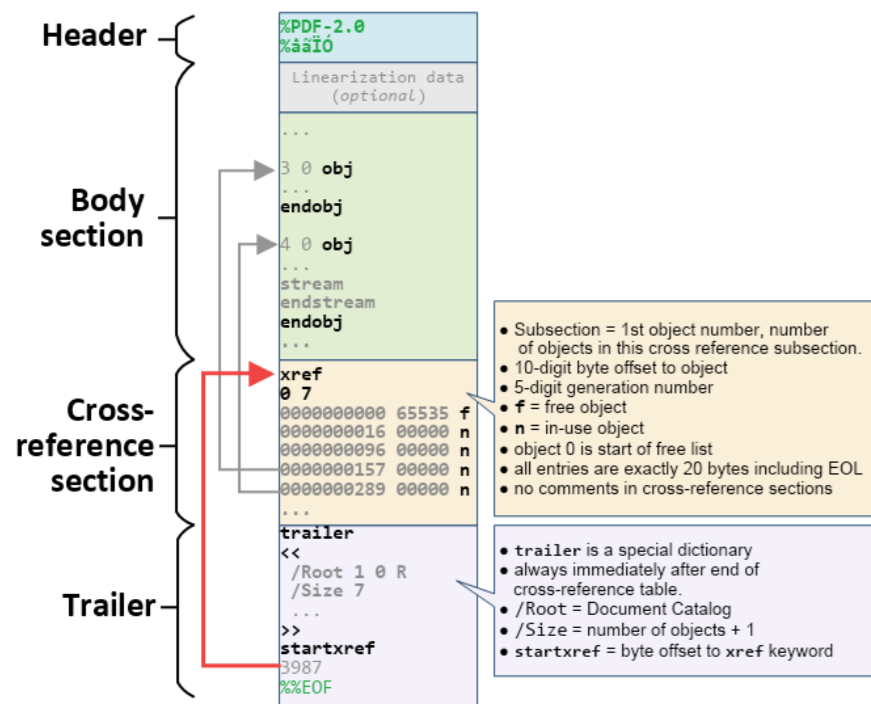|	
Action	PDF feature enabling automatic behaviours triggered by a user interaction or event (e.g., change to a different page when a bookmark is clicked, follow a URL hyperlink, etc.).
AT	**A**ssistive **T**echnology. Associated with PDF/UA and Tagged PDF.
BBox	Bounding Box. A common key name.
Conformance level	Represented by letter designators after a PDF subset acronym, (e.g., PDF/A-1b, PDF/X-5pg, PDF/VT-2s). Each Conformance Level has its own specialized set of rules and requirements.
COS	**C**arousel **O**bject **S**yntax. The syntax used by PDF and FDF files. "Carousel" was the codename for Acrobat 1.0.
Cross-reference stream	(*PDF 1.5 and later only*) Cross-reference information stored in a stream instead of a standard cross-reference `xref` table. Trailer dictionary entries are in cross reference stream dictionary.
Destination	An object defining a view of a document, comprising a page, the location of the document window on that page, and zoom factor.
Direct object	PDF object that occurs inline where it is defined and does not have its own object identifier (object number and generation number pair).
FDF	**F**orms **D**ata **F**ield file format to store interactive form data
Hybrid-reference PDF	(*PDF 1.5 and later only*). PDF containing objects referenced by conventional cross-reference tables in addition to objects in object streams referenced by cross-reference streams.
Linearized PDF	Commonly referred to as Fast Web View.
`obj`	Object abbreviation. A reserved PDF keyword.
Object stream	(*PDF 1.5 and later only*). A stream in which indirect objects may be stored, as an alternative to being stored in PDF body sections.
OCG	**O**ptional **C**ontent **G**roup. A selectable "layer" of page content.
Owner Password	Password with full (owner) access, including ability to change passwords and access permissions of the PDF document.
PAdES	**P**DF **Ad**vanced **E**lectronic **S**ignatures. ETSI standard EN 319 142.
Page Label	Optional descriptive text for referring to pages that can be shown on-screen (e.g., i, ii, iii, …, Chapter 1, Chapter 2, etc). This contrasts with the zero-based integer page index used internally in PDF files.
`startxref`	Reserved PDF keyword that occurs just before the `%%EOF` end-of-file comment marker along with the byte offset to the cross-reference data for the PDF file (expressed as an integer in ASCII).
`trailer`	The trailer dictionary is required in every PDF and defines special objects (e.g., largest object number, the Document Catalog root). Also keyword.
User Password	Password with restricted access permissions (as set by an author).
Widget	A subtype of PDF annotation used with interactive forms that represent the GUI "widgets" through which data entry is done.
XFA	**X**ML **F**orms **A**rchitecture. Proprietary XML-based specifications supporting dynamic forms. Deprecated in PDF 2.0.
XFDF	XML-based version of FDF defined by ISO 19444-1.
XMP	e**X**tensible **M**etadata **P**latform. XML-based metadata standard (ISO 16684) used by many file formats. Required by PDF subsets and PDF 2.0.
`xref`	Reserved PDF keyword that indicates the start of a standard cross-reference table. Often shorthand for "cross reference table".

PDF association

PDF Association Cheat Sheet – Basics

Lexical Rules

End-of-Line (EOL) Sequences

`0x0D`	Carriage Return (CR) only
`0x0A`	Line Feed (LF) only
`0x0D 0x0A`	Carriage Return (CR) followed by Line Feed (LF) only

White space (Table 1)

`0x00`	Null byte
`0x09`	Horizontal Tab (HT)
`0x0C`	Form Feed (FF)
`0x20`	SPACE
`EOL`	Any End-of-Line sequence (*see above*)
`%` …	PDF comments (starting from `%` to EOL) are treated as single white space

Token Delimiter symbols

`(`	Literal string start token
`)`	Literal string end token
`<, <<`	Hex string start token / dictionary start (`<<`) token
`>, >>`	Hex string end token / dictionary end (`>>`) token
`[`	Array start token
`]`	Array end token
`/`	PDF name
`%`	Comment to end-of-line (*outside of a string or inside a content stream*)
`{`	*Only* in Type 4 PostScript calculator functions
`}`	*Only* in Type 4 PostScript calculator functions

File Structure

(when not using cross-reference streams (PDF 1.5))

Header
```
%PDF-2.0
%áâÏÓ
```
```
Linearization data
      (optional)
```

Body section
```
...
3 0 obj
...
endobj
...
4 0 obj
...
stream
endstream
endobj
...
```

Cross-reference section
```
xref
0 7
0000000000 65535 f
0000000016 00000 n
0000000096 00000 n
0000000157 00000 n
0000000289 00000 n
```

Trailer
```
trailer
<<
  /Root 1 0 R
  /Size 7
  ...
>>
startxref
3987
%%EOF
```

- Subsection = 1st object number, number of objects in this cross reference subsection.
- 10-digit byte offset to object
- 5-digit generation number
- **f** = free object
- **n** = in-use object
- object 0 is start of free list
- all entries are exactly 20 bytes including EOL
- no comments in cross-reference sections

- **trailer** is a special dictionary
- always immediately after end of cross-reference table.
- /Root = Document Catalog
- /Size = number of objects + 1
- **startxref** = byte offset to **xref** keyword

Each **Incremental Update** appends an additional **Body**, **Cross-reference** and **Trailer** section to the PDF file, allowing edits and changes without rewriting the full PDF. Link to previous PDF state is via `/Prev` entry in the trailer dictionary to previous **xref**.
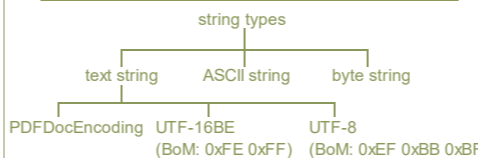
PDF Basics

Reserved Keywords (case sensitive)

`endobj, endstream, f, false, n, null, obj, R, startxref, stream, trailer, true, xref`

Objects

|		
Boolean	`true` `false`	• Case sensitive keywords.

Integer	`0`
	`123`
	`+045`
	`-67890`
		• Signed decimal integer.
		• No hexadecimal or octal integers.

Real Number	`1.23`
	`-45.6`
	`+7.8`
	`-.9`
	`0.`
		• Signed decimal floating-point numbers.
		• No exponential or scientific formats.
		• Integers can be used for real numbers.

String	`(literal string)`
	`(balanced () ok)`
	`(unbalanced \()`
	`(line \`
	` break)`
	`(line \nbreak)`
	`(octal \234 code)`
	
	`<hex-string>`
	`<48656c6c6F>`
	`<41424> % 0 added`
		• Encrypted PDFs encrypt string objects.
		• Unicode strings with byte order markers.
		• Backslash escape sequences for literal strings:

Escape sequences (for literal strings):

Sequence	Meaning
`\n`	LF (0x0A)
`\r`	CR (0x0D)
`\t`	Horizontal Tab (0x09)
`\b`	Backspace (0x08)
`\f`	Formfeed (0x0C)
`\)`	Left parenthesis
`\)`	Right parenthesis
`\\`	Backslash
`\ddd`	Octal code. 1-3 digits

string types:

text string	ASCII string	byte string
PDFDocEncoding	UTF-16BE (BoM: 0xFE 0xFF)	UTF-8 (BoM: 0xEF 0xBB 0xBF)

Name	`/CaseSensitive12` `/HashSign#23`	• Start with `/` SOLIDUS (0x2F). • Can use **#** followed by 2 hex digits.

Array	`[/AName true null`
	` -1.23 10 0 R … ]`
	`[] % an empty array`
		• One dimensional ordered collection with zero or more elements.
		• Array elements can be any type of object.

Dictionary	`<<`
	`  /KeyName value`
	`  ...`
	`>>`
		• Associative unordered table containing **key/value pairs** known as an **entry**.
		• Keys must be unique direct name objects.
		• If value is `null` then same as if key does not exist.

Stream	`10 0 obj`
	`<< /Length int`
	`  ...`
	`>>`
	`stream`
	`…stream data…`
	`endstream`
	`endobj`
		• Contains zero or more stream data bytes.
		• Always need a stream dictionary.
		• Stream data can be compressed and/or encrypted using **Filters**.
		• Always an indirect object.
		• Cannot be in object streams.
		• Encrypted PDFs encrypt stream data.

Null object	`null`	• Case sensitive reserved keyword.
Indirect Reference	`10 0 R`	• Object number then generation number. • Method to refer to another object.

Document Structure
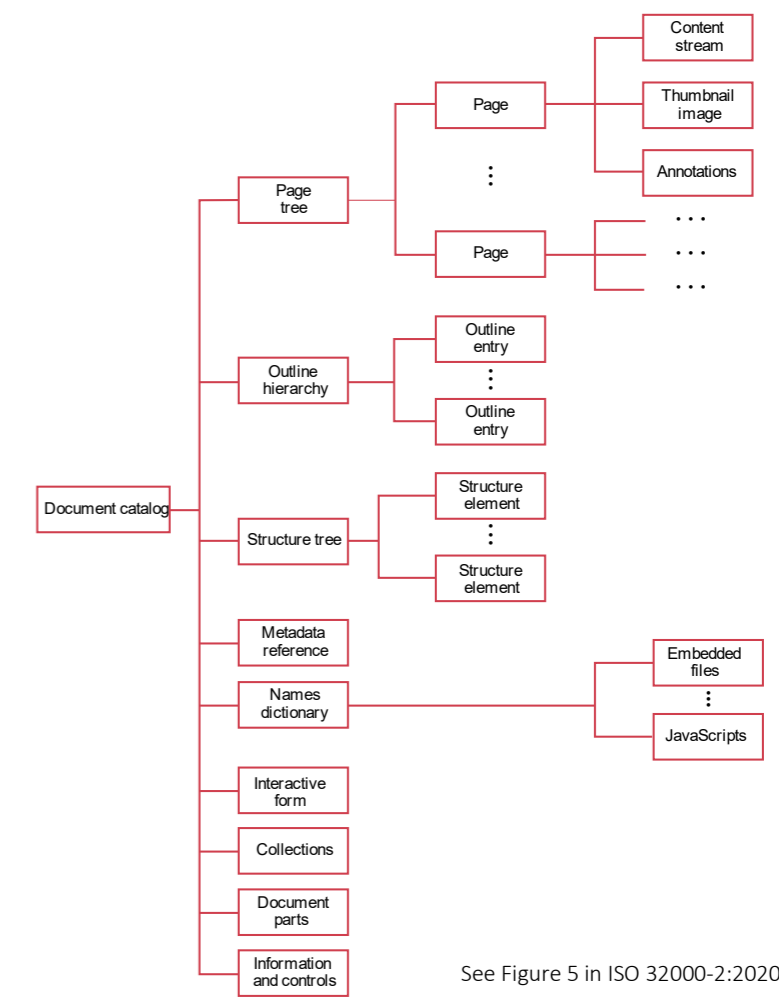
See Figure 5 in ISO 32000-2:2020

Painting Graphics
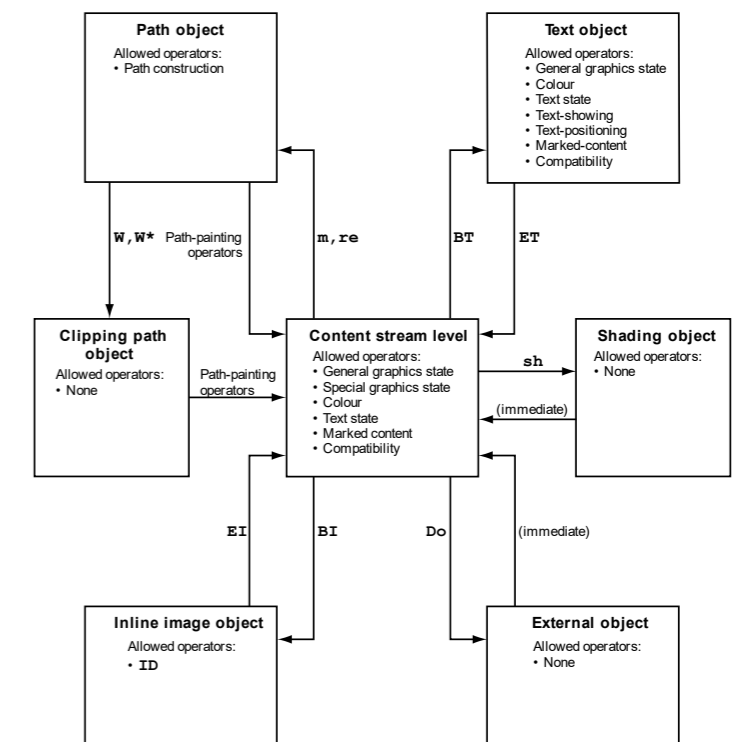
See Figure 9 in ISO 32000-2:2020