Sustainability of Digital Formats: Planning for Library of Congress Collections

PDF (Portable Document Format) Family

>> Back

Table of Contents

Format Description Properties

- ID: fdd000030
- Short name: PDF_family
- Content categories: text, still-image
- Format Category: file-format, encoding, family
- Other facets: unitary, binary, symbolic
- Last significant FDD update: 2023-08-29
- Draft status: Full

Identification and description

Full name	PDF (Portable Document Format) Family
Description	PDF (Portable Document Format), developed by Adobe Systems Incorporated, is described by Adobe as a general document representation language. PDF represents formatted, page-oriented documents. These documents may be structured or simple. They may contain text, images, graphics, and other multimedia content, such as video and audio. There is support for annotations, metadata, hypertext links, and bookmarks. Later versions provide additional functionalities, for example, to embed geospatial information within documents that represent maps or other geospatial images, such as satellite photographs. At the core of PDF is an imaging model derived from the PostScript page description language. See Postscript Format Family. This model enables the description of text and graphics in a device-independent and resolution-independent manner at a complete, precise and professional level. Unlike PostScript, which is a programming language, PDF is based on a structured binary file format that is optimized for high performance in interactive viewing.
Production phase	In general, a final-state format for delivery to end users.
Relationship to other formats	
Has subtype	PDF_1_3, PDF Versions 1.0-1.3
Has subtype	PDF_1_4, PDF Version 1.4
Has subtype	PDF_1_5, PDF, Version 1.5
Has subtype	PDF_1_6, PDF, Version 1.6
Has subtype	PDF_1_7, PDF, Version 1.7 (ISO 32000-1:2008)
Has subtype	PDF_1_7_ext03, PDF, Version 1.7, ExtensionLevel 3
Has subtype	PDF_1_7_ext05, PDF, Version 1.7, ExtensionLevel 5
Has subtype	PDF_2_0, PDF, Version 2.0, ISO 32000-2 (2017, 2020)
Has subtype	PDF/A_family, PDF for Long-term Preservation. As of November 2012, there are three chronological versions of PDF/A.
Has subtype	PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4
Has subtype	PDF/A-2, PDF/A-2 for Long-term Preservation, Use of ISO 32000-1 (PDF 1.7)

Has subtype	PDF/A-3, PDF/A-3 for Long-term Preservation, Use of ISO 32000-1 (PDF 1.7), with Embedded Files
Has subtype	PDF/A-4, PDF for Long-term Preservation, Use of ISO 32000-2 (PDF 2.0)
Has subtype	PDF/E-1, PDF Engineering Document Format, Use of PDF 1.6
Has subtype	PDF/UA-1, PDF/UA-1, PDF Enhancement for Accessibility, Use of ISO 32000-1
Has subtype	PDF/X, PDF for Prepress Graphics File Exchange
Has subtype	PDF/VT-1 (ISO 16612-2:2010 Graphic technology — Variable data exchange — Part 2: Using PDF/X-4 and PDF/X-5 [PDF/VT-1 and PDF/VT-2]). Not described here at this time.
Has subtype	PDF/VT-2 (ISO 16612-2:2010 Graphic technology — Variable data exchange — Part 2: Using PDF/X-4 and PDF/X-5 [PDF/VT-1 and PDF/VT-2]). Not described here at this time.
Has subtype	PDF/VT-3 (ISO 16612-3:2020 Graphic technology — Variable data exchange — Part 3: Using PDF/X-6 [PDF/VT-3]). Not described here at this time.
Has subtype	PDF/VCR-1 (ISO 16613-1:2017 Graphic technology — Variable content replacement — Part 1: Using PDF/X for variable content replacement (PDF/VCR-1). Not described here at this time.
Has subtype	PDF/R-1, For raster image transport and storage. Based on PDF 1.4-1.7 (ISO 32000-1)
Has subtype	PDF/R-1_enc, For raster image transport and storage. Encrypted, based on PDF 2.0 (ISO 32000-2)
Has subtype	GeoPDF_file, GeoPDF File Format (TerraGo)
May contain	PDF_geospatial, PDF, Geospatial encoding (Adobe). Supported by version 1.7 ExtensionLevel 3.
May contain	GeoPDF_OGC, GeoPDF Encoding (TerraGo 2.2), OGC Best Practice

Local use

LC experience or existing holdings	Used as service format, including for some scanned historical materials, primarily to support convenient downloading and printing. Acceptable format for copyright registration of text and image works; see https://www.copyright.gov/eco/help-file-types.html.
LC preference	The Library of Congress expresses preferences for formats for content for its collections through two venues: • The "Best Edition" specification from the U.S. Copyright Office in Circular 7b. Rev: 09/2017 of Circular 7b lists formats acceptable for mandatory deposit of Electronic Serials available only online, in order of preference. For page-oriented renditions, PDF/A (the PDF/A-1 format or later versions) appears first on the list. Other forms of PDF are acceptable, preferably with searchable text. • The Library of Congress Recommended Formats Statement (RFS) includes high quality PDFs, with features such as searchable text, embedded fonts, lossless compression, high resolution images, as a preferred format for textual works in digital form, electronic serials and accompanying image/text files for digital audio as well as an acceptable format for 2D and 3D Computer Aided Design vector images and other graphic images - digital. The RFS list does not distinguish between chronological versions of PDF.

Sustainability factors

Disclosure	Fully documented. Many members of the PDF family of formats were developed by Adobe Systems Incorporated, which made the specifications available openly and at no charge. Several members of the family have been adopted as ISO international standards, e.g,. PDF/X (ISO 15930), PDF/A (ISO 19005), and PDF version 1.7 (ISO 32000-1:2008). These standards are available for sale, primarily through national standards bodies and approved agents. Additional information about specifications and standardization is provided in the format descriptions for several of the subtypes. Since the approval by ISO of PDF 1.7 as ISO 32000-1:2008, maintenance of almost all PDF specifications has been under the auspices of working groups of ISO TC 171 SC 2, with the main PDF specification under WG8. From 2002 to 2016, AIIM (The Association for Information and Image Management) acted as secretariat and U.S. Technical Advisory Group (TAG) to ISO/TC 171 SC 2 (see AIIM | U.S. TAG to ISO/TC 171 from 2015). In 2017, the 3D PDF Consortium was approved by the American National Standards Institute (ANSI) as a standards developer and assumed the roles of secretariat and U.S. TAG Administrator for ISO/TC 171 SC 2 (see 3D PDF Consortium

	Approved by ANSI as US TAG Administrator for PDF ISO Standards, link via Internet Archive). In April 2020, ANSI accredited the PDF Association (through its U.S. subsidiary PDF Association, Inc.) as the new U.S. TAG Administrator for ISO/TC 171 SC 2. See PDF Association to Serve as ANSI-Accredited US Technical Advisory Group Administrator for ISO TC 171 SC 2. Through ANSI, the PDF Association has been acting as secretariat for ISO TC 171 SC 2 since early 2020.
Documentation	See Standards by ISO/TC 171/SC 2 for the current ISO standards under the auspices of the ISO subcommittee responsible for most of the PDF standards, including the base PDF standards (ISO 32000), PDF/A standards (ISO 19005), PDF/R (ISO 23504), PDF/E (ISO 24517), PDF/UA (ISO 14289). Also listed are some related standards, such as those for XMP (Extensible metadata platform, IS0 16648) and ECMAScript for PDF (ISO 21757). The standards for PDF/X are maintained as parts of ISO 15930 by WG 2 of a different ISO Technical Committee ISO/TC 130 Graphic technology. See Specifications below for links to the ISO catalog records for ISO 32000-1:2008 (PDF 1.7) and ISO 32000-2:2020 (PDF 2.0), the current standards for the base PDF format. In addition, the PDF Association announced in an April 2023 press release that it provides no cost downloads of the ISO 32000-2 (PDF 2.0) bundle at https://www.pdfa-inc.org/product/iso-32000-2-pdf-2-0-bundle-sponsored-access/. The PDF Association provides an archive of earlier versions of PDF specifications at https://pdfa.org/resource/pdf-specification-index/.
Adoption	Extremely widely adopted as a platform-independent format for disseminating page-oriented documents. For some general observations on trends, see the slides in PDF statistics – the universe of electronic documents, from a talk by Duff Johnson at PDF Europe Days in May 2018. One area in which PDF is particularly widely used is for prepress workflows. See How PDF changed prepress production dramatically in the last 25 years, a blog post from October 2018. Adobe Reader software for viewing PDF files is freely distributed and bundled with most personal computers.
Licensing and patents	Adobe has a number of patents covering technology that is disclosed in the Portable Document Format (PDF) Specification, version 1.3 and later. An annotated summary of information on the Adobe Web site in September 2010 (see http://partners.adobe.com/public/developer/support/topic_legal_notices.html via Internet Archive) follows. Note that, based on a 20-year period for U.S. Patents, all the patents listed on this Adobe page are expected to have expired in the U.S. by 2019-05-06. Comments welcome. To promote the use of PDF for information interchange the following patents are licensed by Adobe on a royalty-free, non-exclusive basis for the term of each patent for developing software that produces, consumes, and interprets PDF files : 5,634,064 (filed 1996-08-02, granted 1997-05-27, probably expired as of 2019-03-01); 5,737,599 (filed 1995-12-07, granted 1998-04-07, probably expired as of 2019-03-01); 5,781,785 (filed 1995-09-26, granted 1998-07-14, probably expired as of 2019-03-01); 5,819,301 (filed 1997-09-09, granted 1998-10-06, probably expired as of 2019-03-01); 6,028,583 (filed 1998-01-16, granted 2002-02-22, probably expired as of 2019-03-01); 6,289,364 (filed 1997-12-22, granted 2001-09-11, probably expired as of 2019-03-01); 6,421,460 (filed 1999-05-06, granted 2002-07-16, probably expiring 2019-05-06). Patent 5,860,074 (filed 1997-08-14, granted 1999-01-12, probably expired as of 2019-03-01) is similarly licensed on a royalty-free, non-exclusive basis for its term but only for the purpose of developing software that produces PDF files (thus specifically excluding software that consumes and/or interprets PDF files). A similar statement was submitted in 2004 to IETF as part of Adobe Systems Incorporated's Statement about IPR claimed in draft-zilles-pdf. In association with the adoption of PDF, version 1.7 as an ISO standard (ISO 32000-1:2008), Adobe issued a Public Patent License, granting "every individual and organization in the world the royalty-free right, under all Essential Claims that Adobe owns, to make, have made, use, sell, import and distribute Compliant Implementations."
Transparency	Depends upon compliant software tools to read. Building tools requires sophistication. In most PDF files, the content is compressed. Many PDFs include embedded content in binary form, for example for images, and for annotations in audio or video. See Notes below for more detail on compression of "stream objects" in PDF files.
Self-documentation	Starting with PDF 1.4, chronological versions of PDF can include metadata "streams" in the XMP (Extensible Metadata Platform) format at the level of the document or for individual objects. The XMP format, developed by Adobe in 2001, is a framework for including arbitrary blocks of metadata, using a representation in RDF. XMP was approved as ISO 16684-1:2012 in 2012.

	Use of document-level XMP is mandatory in some standard subsets for PDF, including the PDF/A_family. **Accessibility Features** Depending on implementation, PDF files can strongly support accessibility. One component is logical structure features that allow applications that produce PDF files to choose what structural information to include and how to represent it. According to W3C's PDF Techniques for WCAG 2.0, a PDF document's logical structure is stored separately from its visible content, with pointers from each to the other. This separation allows the ordering and nesting of logical elements to be entirely independent of the order and location of graphics objects on the document's pages. The logical structure of a document is described by a hierarchy of objects called the structure hierarchy or structure tree. At the root of the hierarchy is a dictionary object called the structure tree root, located by means of the StructTreeRoot entry in the document catalog. See Section 14.7.2, ("Structure Hierarchy") in PDF 1.7 (ISO 32000-1): Table 322 shows the entries in the structure tree root dictionary. The K entry specifies the immediate children of the structure tree root, which are structure elements." In addition, tagged PDF "(PDF 1.4) is a stylized use of PDF that builds on PDF's logical structure framework. It defines a set of standard structure types and attributes that allow page content (text, graphics, and images) to be extracted and reused for other purposes." These include" • Simple extraction of text and graphics for pasting into other applications. • Automatic reflow of text and associated graphics to fit a page of a different size than was assumed for the original layout. • Processing text for such purposes as searching, indexing, and spell-checking. • Conversion to other common file formats (such as HTML, XML, and RTF) with document structure and basic styling information preserved. • Making content accessible to people who rely on assistive technology."
External dependencies	Faithful rendering requires that fonts be embedded. PDF variants in the PDF/A_family, intended for archival purposes, PDF/X, for prepress exchange, and PDF/E-1, for engineering documentation, require that fonts be embedded.
Technical protection considerations	The PDF format offers several forms of technical protection, including encryption and password protection, that would prevent custodians of digital content ensuring accessibility in future technological environments.

Quality and functionality factors

Still Image	
Normal rendering	PDF is designed for page-oriented documents. Scaling, zooming, printing are expected functionalities for PDF viewers. The quality of raster images depend on the quality of the embedded image. Note that, in general, PDF is not a preferred archival or master format for images.
Clarity (high image resolution)	High-resolution images can be embedded using professional tools. See PDF/X, a standard version of PDF used by the printing industry.
Color maintenance	Parameters to support color management, including CIE-based and ICC-based color spaces, can be stored in the file using professional tools. See PDF/X, a standard version of PDF used by the printing industry.
Support for vector graphics, including graphic effects and typography	Extensive support for graphic elements. Versions after PDF 1.4 support a transparent imaging model in addition to the opaque model used for earlier versions. Hence images composed of layers can be stored without pre-composing into a single image.
Support for multispectral bands	PDF is designed to support printing and visual display on screens. PDF is not designed to support analysis of multispectral image data, which can include bands/channels outside the visible range. For scientific communication, raw multispectral image data may be converted to an image in a color space and format supported in the PDF specifications. See, for example, Introduction to Image Processing from the website of the Hubble space telescope. Comments welcome.
Functionality beyond normal rendering	PDF has extensive support for annotations of several types. PDF, Version 1.7, ExtensionLevel 3 (PDF_1_7_ext03), introduced with Acrobat 9.0, and incorporated into PDF 2.0 (PDF_2_0) supports capabilities for embedding geospatial data in association with points within 3D and geospatial images. PDF 2.0 also supports "rich media" annotations, and can contain measurement properties for use in 2D images, 3D artwork, and rich media annotations.
Text	
Normal rendering	Good support is possible, but not guaranteed. The PDF format allow creators to disallow printing and extraction of text for quotations. PDF can also be used to create documents from scanned page images; such files do not necessarily support indexing of the document text.

	Although for most PDFs that do incorporate character-based text, the text can be reliably extracted and indexed, problems can occur, because the PDF internal structure for text is based primarily on identification of glyphs within fonts and not on Unicode code points. If Unicode code points are not present, perhaps in order to make the file as small as possible, extracted text will be unintelligible. See Why is the extraction of text from a PDF document such a hassle?, a blog post by Dr. Hans Bärfuss of pdf-tools.com. See also a useful response to the problem Cannot copy non-latin characters from PDF document on Stack Exchange.
Integrity of document structure	The logical structure of a document is only represented in a PDF file if the creator or process during creation takes steps to incorporate structural tagging.
Integrity of layout and display	PDF is designed to represent the layout of page-oriented documents.
Support for mathematics, formulae, etc.	Can be represented visually by embedded graphics or using mathematical fonts and specialized tools for mathematical layout. See, for example, MathML Samples, a PDF 1.5 document from Antenna House, a vendor of specialized typesetting tools. Antenna House Formatter can use XSL-FO (XSL Formatting Objects) to format MathML for printing or display. The typesetting system LaTeX has a Math Mode that can be used to create PDFs. See Very Basic Mathematical Latex, a PDF created using LaTeX that shows sample code and rendered equations and formulas. PDF 2.0 introduced a standard way to incorporate the source MathML markup in a way that supports content re-use and accessibility via tools that convert MathML to braille. See, for example, Positive Impacts of EPUB 3: MathML and Braille Mathematics. A widely used software library for generating braille, Liblouis, has a version that can translate MathML into a braille representation; see Liblouisutdml User's and Programmer's Manual.
Functionality beyond normal rendering	Supports annotations and bookmarks. Supports embedding of media objects (in binary format) and links to external media objects, such as images, audio, or video. Audio and video are considered "annotations." PDF Support for 3D artwork in the U3D format was added in Adobe's PDF 1.7, ExtensionLevel 3 and has been incorporated into PDF 2.0. PDF 2.0 also added support for 3D artwork in the PRC format.

File type signifiers and format identifiers

Tag	Value	Note
Filename extension	pdf	
Internet Media Type	application/pdf	Registered with IANA (see Application Media-Types), described originally in IETF (Internet Engineering Task Force) RFC 3778, which was obsoleted by RFC 8118. Reported for PDF files by JHOVE PDF-hul module for file identification.
Internet Media Type	application/x-pdf application/acrobat application/vnd.pdf text/pdf text/x-pdf	Selected media types listed at The File Extension Source.
Magic numbers	Hex: 25 50 44 46 ASCII: %PDF	From Gary Kessler's File Signatures Table.
Indicator for profile, level, version, etc.	See note.	PDF files should have a chronological version identified in the header with the 5 characters %PDF– followed by a version number. For example, PDF 1.7 would be identified as %PDF-1.7. However, this version identification can be over-ridden by a version value stored in the document's *Catalog*. See Notes below, for more detail.
Pronom PUID	See note	Pronom PUIDs at subtype level only.
Wikidata Title ID	Q42332	See https://www.wikidata.org/wiki/Q42332.

Notes

General	In 2023, the PDF Association published a series of PDF Cheat Sheets which "are designed to help developers work more efficiently while ensuring that their knowledge of PDF is technically correct. They are highly condensed summaries, with logical groupings of information to help jog one's memory about nuances or details that are often forgotten without regular and repeated use. These cheat sheets use simple sentences, illustrations and color coding wherever possible to optimize support for the global community of PDF developers." These no cost resources cover four main topics areas: Basics, Graphic operators and operands, Common Objects and Color.

Maximum size for PDFs: This topic has been discussed in a number of online forums. On one Adobe forum, in response to a 2012 question Is there a PDF size limit?, a very high theoretical page-count limit is described: "There's no explicit page number limit but there is a limit on indirect objects of 8,388,607 in a 32-bit PDF rendering application--Acrobat and Adobe Reader are both 32-bit code--and because each page consumes at least one indirect object, every PDF file created by or opened by Acrobat must have less pages than that. If you were to create a native x64 PDF application you could add more pages, but the resulting files wouldn't open at all in 32-bit apps." This forum entry goes on to say, "Architecturally there is only one limit in the PDF standard: the overall file size must be below ~10GB as the cross-reference tables which define the PDF structure use 10 bits."

The preceding paragraph offers a generous view of the potential size for a PDF. Many commentators argue that the limit for practicality is lower than those stated above. What matters is whether you can open a given PDF in any reasonable application, including Acrobat and Adobe Reader, mentioned above. Online forums also include reports like these examples: "It seems that the iPad has a limit of 30MB for displaying PDF files," and "users of GoodReader have reported flawless performance with files over 1 gig in size." The practical limits imposed by applications might also include limits set by indexers if the PDF includes searchable text.

Self-identification of chronological versions of PDF: Identification of chronological versions of PDF can be given in two places in a PDF file. All PDF files should have a version identified in the header with the 5 characters *%PDF–* followed by a version number of the form 1.N, where N is a digit between 0 and 7 or a version number of 2.0. For example, PDF 1.7 would be identified as *%PDF–1.7*. However, beginning with PDF 1.4, a conforming PDF writer may use the Version entry in the document Catalog to override the version specified in the header. The location of the Catalog within the file is indicated in the Root entry of the file trailer/footer. This override feature was introduced to facilitate the incremental updating of a PDF by simply adding to the end of the file. As a result, it is necessary to locate the Catalog within the file to get the correct version number. Unless the PDF is "linearized," in which case the Catalog is up front, this will require reading the trailer and then using the reference there to locate the Catalog, which will typically be compressed. This has practical implications because format identification tools, including DROID, typically look for particular characters at the beginning of a file (i.e., in the header), to permit identification with minimal effort. DROID can look for characters at the end of the file, but is not able to follow an indirect reference or decompress file contents. When the version number is not the same in the header and the Catalog, there is potential for format identification errors.

Compression of "stream objects" in PDF files: Stream objects in a PDF file are often compressed. A number of compression schemes are supported for PDF files, indicated by Filter values defined in the specification. Filter names correspond to the decoding/decompression that must be applied to recover the original data. Filters can be combined into pipelines. The Filters listed below are permitted in generic PDF files. However, some filters are not permitted in the "subset standards for PDF," such as PDF/A, PDF/X, and PDF/E.

- ASCIIHexDecode: data encoded in an ASCII hexadecimal representation. This filter was developed for Postscript, at a time when it was useful to transform binary data into a representation that could be safely used in mail messages. Binary data can be converted to ASCII by representing each byte of binary data as two ASCII hexadecimal digits (0–9 and either A–For a–f). Newline characters are added in the encoded output at least once every 80 characters, thereby limiting the lengths of lines. The resulting data was twice as large as the original binary form. See PostScript Language Reference, 3rd edition link via Internet Archive.
- ASCII85Decode: data encoded in an ASCII base-85 representation. This filter was also used in PostScript and was preferred to ASCIIHexDecode, because it yielded smaller files. Generally, for every 4 bytes of binary data, this compression scheme produces 5 ASCII printable characters. ASCII85 is the encoding scheme used by the uuencode and uudecode programs available on UNIX systems. See PostScript Language Reference, 3rd edition (link via Internet Archive).
- LZWDecode: data encoded using LZW (Lempel-Ziv-Welch) compression. LZW is a lossless, variable-length, adaptive compression method that can be applied to any form of data. LZW is supported in the TIFF 6.0 image format. Until 2004, it was protected by a patent.
- FlateDecode: data encoded using the ZLIB/DEFLATE compression method defined in IETF RFC 1950 and IETF RFC 1951. Similar to LZW, it had the advantage of being patent-free. Permitted for PDF 1.2 and above. The most commonly used method to compress the text content of PDFs.
- RunLengthDecode: data encoded using a byte-oriented run-length encoding algorithm. See Wikipedia entry for Run-length Encoding.
- CCITTFaxDecode: data encoded using the CCITT (now ITU-T) Group 3 or Group 4 facsimile standard. Used for monochrome image data at 1 bit per pixel. See ITU-T Group 4 FAX Compression.

	• JBIG2Decode: data encoded using the JBIG2 standard, as defined by ISO/IEC 14492:2001 and its amendments 1 and 2, excluding amendment 3. Used for monochrome (1 bit per pixel) image data. Permitted for PDF 1.4 and above. See Wikipedia entry for JBIG2. • DCTDecode: data encoded using a DCT (discrete cosine transform) technique based on the JPEG standard. Baseline JPEG encoding is permitted for all chronological versions of PDF. For PDF 1.3 and above, Progressive JPEG encoding is permitted. See JPEG_DCT_BL and JPEG_DCT_PRG. • JPXDecode: data encoded using the JPEG 2000 compression method and the JPX data structure as defined in ISO/IEC 15444-2. Permitted for PDF 1.5 and above. Data used in PDF image XObjects shall be limited to the JPX baseline set of features. Enumerated colour space 19 (CIEJab) is not supported. However, enumerated colour space 12 (CMYK), which is part of JPX but not JPX baseline, is permitted in a PDF. The Crypt filter (introduced in PDF 1.5) can be used to specify the encryption algorithm that has been applied to a datastream. Many of the encryption algorithms supported in earlier chronological versions of PDF are now deprecated. **Tagged PDF**: The concept of a tagged PDF was introduced in PDF 1.4. In addition to the content tree that is part of any PDF, a tagged PDF also has a structure tree. Tags provide a logical structure that governs how the document content is presented through assistive technology. Each tag identifies the associated content element, for example paragraph <P>, heading level three <H3>, list item <LI>, image <Figure>, or table data cell <TD>. The order of the tags defines the reading order. PDF Reference (third edition) (link via Internet Archive), which specifies PDF 1.4, indicates that *Tagged PDF* is a stylized use of PDF that uses a set of standard structure types and attributes that allow page content (text, graphics, and images) to be extracted and reused for other purposes. Tagged PDFs follow a set of rules for representing text in the page content so that characters, words, and text order can be determined reliably for use by tools that perform operations such as: • Simple extraction of text and graphics for pasting into other applications • Automatic reflow of text and associated graphics to fit a page of a different size than was assumed for the original layout • Processing text for such purposes as searching, indexing, and spell-checking • Conversion to other common file formats (such as HTML, XML, and RTF) with document structure and basic styling information preserved • Making content accessible to the visually impaired Many PDF files are created by "printing to PDF" or other methods that to not create tag structures. In general, the logical structure of a document is only represented in a PDF file if the creator or a process during creation takes steps to incorporate structural tagging. See What is Tagged PDF? from 2004, in which Duff Johnson said, "Tags may be generated automatically for any PDF file using Acrobat 6.0 Professional, but unless the document is very simple indeed, automated tagging alone is unlikely to produce satisfactory results, and is certainly not a quick-fix for compliance with Section 508." Some applications for which the native format is inherently structured appropriately have specialized exports that can create tagged PDFs that represent that structure. For example, according to the description of exporting to accessible PDF (link via Internet Archive) from Framemaker (desktop publishing software), "Tagged Adobe PDF provides the following capabilities: • Ensures that information is in the correct reading order on a page. • Includes paragraph attributes used to correctly reflow the document contents into different-sized devices, such as eBook reading devices. • Ensures the reliable translation of text into Unicode. This approach recognizes ligatures and hyphens, so that a Windows screen reader can correctly read all characters and words. • Recognizes alternative text descriptions for graphics in anchored frames. • Enables the document to be exported more reliably to Rich Text Format (RTF) and XML from Acrobat 7.0 for reuse in other documents. The guidance on accessibility using Adobe InDesign encourages users to "apply accessibility features within your InDesign document, rather than having to make major changes in Adobe Acrobat software. PDF tags, alt tags, and the content order you assign stay with the document as you revise it." The Tagged PDF Best Practice Guide: Syntax, issued in 2019 by the PDF/UA Technical Working Group (PDF/UA TWG) provides detailed guidance aimed at accessibility.
History	Adapted from PDF Reference, Third Edition: The origins of PDF and the Adobe Acrobat product family date to early 1990. At that time, the PostScript page description language was rapidly becoming the worldwide standard for the production of the printed page. PDF builds on the PostScript page description language by layering a document structure and interactive navigation features on PostScript's underlying imaging model,

	providing a convenient, efficient mechanism enabling documents to be reliably viewed and printed anywhere. See Wikipedia entry on History of the Portable Document Format (PDF) and descriptions for chronological versions for later history. In October 2009, ISO authorized a new project to develop the PDF 2.0 standard, which was published as ISO 32000-2:2017. See PDF_2_0. The ISO 32000-2 (PDF 2.0) bundle is available for no cost download from the PDF Association. The agenda for the ISO TC 171 SC 2 "PDF Week" – Spring, 2019 indicated that a "dated revision" of ISO 32000-2 (PDF 2.0) was in process and that several new editions for the subset standards are waiting for that revision. In December 2020, the dated revision was published. See announcement from December 16, 2020, titled The new PDF 2.0 and subset standards.

Format specifications

- ISO 32000-2 (PDF 2.0)
  - See PDF 2.0 (https://www.loc.gov/preservation/digital/formats/fdd/fdd000474.shtml).
  - ISO 32000-2:2020. Document management — Portable document format — Part 2: PDF 2.0 (https://www.iso.org/standard/75839.html). Catalog record, preview, and opportunity to purchase.
  - ISO 32000-2 (PDF 2.0) bundle - no cost download from PDF Assocation (https://www.pdfa-inc.org/product/iso-32000-2-pdf-2-0-bundle-sponsored-access/). Includes ISO 32000-2:2020, ISO 32000-2:2020/Amd 1, ISO/TS 32001:2022 and ISO/TS 32002:2022.
- ISO 32000-1 (PDF 1.7)
  - See PDF 1.7 (https://www.loc.gov/preservation/digital/formats/fdd/fdd000277.shtml).
  - ISO 32000-1:2008. Document management — Portable document format — Part 1: PDF 1.7 (https://www.iso.org/standard/51502.html). Catalog record, preview, and opportunity to purchase
  - PDF Version 1.7 specification from Adobe (Equivalent to ISO 32000-1) (https://opensource.adobe.com/dc-acrobat-sdk-docs/pdfstandards/PDF32000_2008.pdf). This page provides open access to an ISO approved copy of the July 2008 ISO 32000-1 Standards document, together with Adobe-specific extensions from 2008 and 2009.

Useful references

URLs

- Archive of PDF specifications from the PDF Association (https://pdfa.org/resource/pdf-specification-index/).
- Wikipedia entry on Portable Document Format (https://en.wikipedia.org/wiki/Portable_Document_Format).
- Wikipedia entry on History of the Portable Document Format (PDF) (https://en.wikipedia.org/wiki/History_of_the_Portable_Document_Format_(PDF)).
- PDF Association (https://www.pdfa.org/).
- PDF format becomes ISO standard, July 2008 (https://web.archive.org/web/20190415060741/https://www.iso.org/news/2008/07/Ref1141.html). An ISO announcement.
- PostScript Language Reference, 3rd edition. Adobe Systems Inc. 1999. Available via Internet Archive (https://web.archive.org/web/20210321144708/https://www.adobe.com/content/dam/acom/en/devnet/actionscript/articles/PLRM.pdf).
- Postscript Technical Note #5603 | Filters and Reusable Streams (Oct 1997). Link via Internet Archive (https://web.archive.org/web/20210322064243/https://www.adobe.com/content/dam/acom/en/devnet/postscript/pdfs/TN5603.Filters.pdf).
- Former home of the Acrobat & Reader Engineering team (http://web.archive.org/web/20141101010717/http://acroeng.adobe.com/wp/?). Link now via Internet Archive. Detailed information and guidance on important aspects of PDF, such as linearization (Fast Web View) and standards-compliant approaches to viewing PDF/A and PDF/X files.
- Adobe PDF References -- from former acroeng.adobe.com site (http://web.archive.org/web/20140110001341/http://acroeng.adobe.com/wp/?page_id=321). Contains links to every version of the PDF Reference published by Adobe as well as associated errata and addenda to the document. No longer online. Link via Internet Archive.
- Documentation for PDF from GNU PDF. No longer available online. Out of date as of 2018, but possibly still helpful. (http://web.archive.org/web/20141012024418/http://www.gnupdf.org:80/Category:PDF).
- JHOVE PDF-hul Module (http://jhove.sourceforge.net/pdf-hul.html).
- Original (2004) registration for application/pdf media type (https://www.rfc-editor.org/rfc/rfc3778).
- Updated (2017) registration for application/pdf media type (https://www.rfc-editor.org/rfc/rfc8118).
- Resources related to the 2020 dated revision of ISO 32000-2, the specification for PDF 2.0.
  - The new PDF 2.0 and subset standards (December 16, 2020) | announcement from PDFA. (https://www.pdfa.org/the-new-pdf-2-0-and-subset-standards/).
  - ISO 32000 Normative References | compilation by PDFA (https://www.pdfa.org/iso-32000-normative-references/). This resource includes normative references and ibliography items from ISO 32000-2, and additional materials.
  - Updated draft PDF 2.0 standard publicly available (March 2020) (https://www.pdfa.org/updated-draft-pdf-2-0-standard-publicly-available/). Describes changes to expect in the 2020 dated revision of the PDF 2.0 specification.
  - PDF/A-4, PDF/X-6 and the other new PDF standards (October 2020) | video of presentation by Dietrich von Seggern (https://www.youtube.com/watch?v=Rc6mPNWy5I4).
  - PDF's ISO-standardized subsets: a tour (May 2018) | video of presentation by Dietrich von Seggern (https://www.youtube.com/watch?v=9BX0Sek696A).

    - PDF's ISO-standardized subsets: a tour (May 2018) | slides from presentation by Dietrich von Seggern (https://www.pdfa.org/wp-content/uploads/2018/06/1145_von_Seggern.pdf).
- Preservation with PDF/A (Second Edition) 2017 | by Betsy Fanning (https://www.dpconline.org/docs/technology-watch-reports/1707-twr17-01-revised/file). DPC Technology Watch Report 17-01 July 2017
- PDF statistics – the universe of electronic documents | slides from PDF Europe Days (May 2018) (https://www.pdfa.org/wp-content/uploads/2018/06/1330_Johnson.pdf).
- How PDF changed prepress production dramatically in the last 25 years | blog post (October 2018) (https://www.callassoftware.com/en/blog/how-pdf-changed-prepress-production-dramatically-in-the-last-25-years).
- Why is the extraction of text from a PDF document such a hassle? (https://blog.pdf-tools.com/2014/01/why-is-extraction-of-text-from-pdf.html). January 2014 blog post from pdf-tools.com
- Very Basic Mathematical LaTeX | a PDF created in LaTeX (http://mathlab.cit.cornell.edu/support/latex/sample/sample.pdf). Includes examples of formulas and equations with source LaTeX code and rendered output.
- Advanced features for publishing mathematics, in PDF and on the Web | from TUGboat (for TeX and LaTeX) (https://tug.org/TUGboat/tb29-3/tb93moore.pdf).
- PDF/A, UA, X, VT and E (explanation from Luratech), now via Internet Archive. (https://web.archive.org/web/20161224065555/https://www.luratech.com/en/subscribe-to-newsletter/detail-vr/article/pdfa-ua-x-vt-and-e/).
- Wikidata entry for Q42332 (https://www.wikidata.org/wiki/Q42332). Information in Wikidata about PDF Family. Wikidata Title ID: Q42332.

Books, articles, etc.

- Morrissey, Sheila M., "The Network is the Format: PDF and the Long-term Use of Digital Content", Archiving 2012, (2012): pp. 200-203. ISBN: 978-0-89208-300-8 (print). Available online at Portico with permission of IS&T: The Society for Imaging Science and Technology.

Last Updated: 05/09/2024

Digital Preservation Home | Digital Formats Home