Portable Document Format

From Wikipedia, the free encyclopedia.

> *"PDF" redirects here. For other uses, see PDF (disambiguation).*

Portable Document Format (**PDF**) is a file format developed by Adobe Systems for representing documents in a manner that is independent of the original application software, hardware, and operating system used to create those documents. A PDF file can describe documents containing any combination of text, graphics, and images in a device independent and resolution independent format. These documents can be one page or thousands of pages, very simple or extremely complex with a rich use of fonts, graphics, colour, and images. PDF is an open standard, and anyone may write applications that can read or write PDFs royalty-free.

In addition to encapsulating text and graphics, PDF files are most appropriate for encoding the exact look of a document in a device-independent way. In contrast, markup languages such as HTML defer many display decisions to a rendering device such as a browser, and will not look the same on different computers.

Free readers for many platforms are available for download from the Adobe website [1], and there are several free open source readers, including Xpdf [2] for POSIX-like systems with the X Window System; KPDF [3], a viewer based on *Xpdf* for KDE; GPdf [4], a derivative of *Xpdf* for GNOME, Evince [5], a document viewer for GNOME (fork of GPdf) that can view PDF-files; GSPdf [6] and ViewPDF [7], for GNUstep; and front-ends for many platforms to Ghostscript.

Proper subsets of PDF, collectively called PDF/X, have been standardized by ISO.

Contents

[show]

-

[edit]

Technology

PDF is primarily the combination of three technologies:

- a cut-down form of PostScript for generating the layout and graphics,
- a font-embedding/replacement system to allow fonts to travel with the documents, and
- a structured storage system to bundle these elements into a single file, with data compression where appropriate.

[edit]

PostScript

Portable Document Format

From Wikipedia, the free encyclopedia.

> *"PDF" redirects here. For other uses, see PDF (disambiguation).*

Portable Document Format (**PDF**) is a file format developed by Adobe Systems for representing documents in a manner that is independent of the original application software, hardware, and operating system used to create those documents. A PDF file can describe documents containing any combination of text, graphics, and images in a device independent and resolution independent format. These documents can be one page or thousands of pages, very simple or extremely complex with a rich use of fonts, graphics, colour, and images. PDF is an open standard, and anyone may write applications that can read or write PDFs royalty-free.

In addition to encapsulating text and graphics, PDF files are most appropriate for encoding the exact look of a document in a device-independent way. In contrast, markup languages such as HTML defer many display decisions to a rendering device such as a browser, and will not look the same on different computers.

Free readers for many platforms are available for download from the Adobe website [1], and there are several free open source readers, including Xpdf [2] for POSIX-like systems with the X Window System; KPDF [3], a viewer based on *Xpdf* for KDE; GPdf [4], a derivative of *Xpdf* for GNOME, Evince [5], a document viewer for GNOME (fork of GPdf) that can view PDF-files; GSPdf [6] and ViewPDF [7], for GNUstep; and front-ends for many platforms to Ghostscript.

Proper subsets of PDF, collectively called PDF/X, have been standardized by ISO.

Contents

[show]

-

[edit]

Technology

PDF is primarily the combination of three technologies:

- a cut-down form of PostScript for generating the layout and graphics,
- a font-embedding/replacement system to allow fonts to travel with the documents, and
- a structured storage system to bundle these elements into a single file, with data compression where appropriate.

[edit]

PostScript

PostScript is a computer language — more precisely, a page description language — that is run in an interpreter to generate an image. This process requires a fair amount of resources.

PDF is a subset of those PostScript language elements that define the graphics, and only requires a very simple interpreter. For instance, flow control commands like `if` and `loop` are removed, while graphics commands such as `lineto` remain.

That means that the process of turning PDF back into a graphic is a matter of simply reading the description, rather than running a program in the PostScript interpreter. However, the entire PostScript world in terms of fonts, layout and measurement remains intact.

Often, the PostScript-like PDF code is generated from a source PostScript file. The graphics commands that are output by the PostScript code are collected and tokenized; any files, graphics or fonts the document references are also collected; and finally everything is compressed into a single file.

As a document format, PDF has several advantages over PostScript. One is that a document resides in a single file, whereas the same document in PostScript may span multiple files (graphics, etc.) and probably occupies more space. In addition, PDF contains already-interpreted results of the PostScript source code, so it is less computation-intensive and faster to open, and there is a more direct correspondence between changes to items in the PDF page description and changes to the resulting appearance of the page. Also, PDF (starting from version 1.4) supports true object transparency while PostScript does not. Finally, if displayed with Adobe Reader, a font-substitution strategy ensures the document will be readable even if the end-user does not have the "proper" fonts installed. PDF also allows font embedding to ensure that the "proper" fonts are displayed. While this is possible with PostScript, such files cannot normally be distributed freely because of font licensing agreements.

[edit]

History

When PDF first came out, in the early 1990s, it was slow to catch on. At the time, not only did the only PDF creation tools of the time (Acrobat) cost money, but so did the software to view and print PDF files. Early versions of the PDF format had no support for external hyperlinks, reducing its usefulness on the web. Additionally, there were competing formats such as Envoy, Common Ground Digital Paper, DjVu and even Adobe's own PostScript file format (.ps). Adobe started distributing the Acrobat Reader program at no cost, and continued to support PDF through its slow multi-year ramp-up. Competing formats eventually died out, and PDF became a well-accepted standard.

In 2005 Microsoft presented a competing format referenced by the code name "Metro". It is developed together with Global Graphics. Metro is based on XML, but requires a license. Metro is scheduled to be included in the next version of Microsoft Windows Vista.

[edit]

Macintosh

PDF was selected as the "native" metafile format for Mac OS X, replacing the PICT format of the earlier Mac OS. Mac OS X's imaging model, Quartz 2D, is based on both the Display

PostScript is a computer language — more precisely, a page description language — that is run in an interpreter to generate an image. This process requires a fair amount of resources.

PDF is a subset of those PostScript language elements that define the graphics, and only requires a very simple interpreter. For instance, flow control commands like `if` and `loop` are removed, while graphics commands such as `lineto` remain.

That means that the process of turning PDF back into a graphic is a matter of simply reading the description, rather than running a program in the PostScript interpreter. However, the entire PostScript world in terms of fonts, layout and measurement remains intact.

Often, the PostScript-like PDF code is generated from a source PostScript file. The graphics commands that are output by the PostScript code are collected and tokenized; any files, graphics or fonts the document references are also collected; and finally everything is compressed into a single file.

As a document format, PDF has several advantages over PostScript. One is that a document resides in a single file, whereas the same document in PostScript may span multiple files (graphics, etc.) and probably occupies more space. In addition, PDF contains already-interpreted results of the PostScript source code, so it is less computation-intensive and faster to open, and there is a more direct correspondence between changes to items in the PDF page description and changes to the resulting appearance of the page. Also, PDF (starting from version 1.4) supports true object transparency while PostScript does not. Finally, if displayed with Adobe Reader, a font-substitution strategy ensures the document will be readable even if the end-user does not have the "proper" fonts installed. PDF also allows font embedding to ensure that the "proper" fonts are displayed. While this is possible with PostScript, such files cannot normally be distributed freely because of font licensing agreements.

[edit]

History

When PDF first came out, in the early 1990s, it was slow to catch on. At the time, not only did the only PDF creation tools of the time (Acrobat) cost money, but so did the software to view and print PDF files. Early versions of the PDF format had no support for external hyperlinks, reducing its usefulness on the web. Additionally, there were competing formats such as Envoy, Common Ground Digital Paper, DjVu and even Adobe's own PostScript file format (.ps). Adobe started distributing the Acrobat Reader program at no cost, and continued to support PDF through its slow multi-year ramp-up. Competing formats eventually died out, and PDF became a well-accepted standard.

In 2005 Microsoft presented a competing format referenced by the code name "Metro". It is developed together with Global Graphics. Metro is based on XML, but requires a license. Metro is scheduled to be included in the next version of Microsoft Windows Vista.

[edit]

Macintosh

PDF was selected as the "native" metafile format for Mac OS X, replacing the PICT format of the earlier Mac OS. Mac OS X's imaging model, Quartz 2D, is based on both the Display

PostScript standard and PDF, and is sometimes referred to as Display PDF. Due to OS support, all OS X applications can create PDF documents automatically as long as they support the Print command.

[edit]

PDF and accessibility

PDF can be accessible to people with disabilities. Current PDF file formats can include tags (essentially XML), text equivalents, captions and audio descriptions, and other accessibility features. Some software, such as Adobe InDesign, can output tagged PDFs automatically. Leading screen readers, including Jaws, Window-Eyes, and Hal, can read tagged PDFs; current versions of the Acrobat and Acrobat Reader programs can also read PDFs out loud. Moreover, tagged PDFs can be reflowed and zoomed for low-vision readers.

However, many problems remain, not least of which is the difficulty in adding tags to existing or "legacy" PDFs; for example, if PDFs are generated from scanned documents, accessibility tags and reflowing are unavailable and must be created either by hand or using OCR techniques. Moreover, that process itself is inaccessible. Nonetheless, well-made PDFs can be a valid choice as long-term accessible documents. (Work is being done on a PDF variant based on PDF 1.4. The PDF/A or PDF-Archive is specifically scaled down for archival purposes.)

Microsoft Word documents can be converted into accessible PDFs, but only if the Word document is written with accessibility in mind - for example, using styles, correct paragraph mark-up and "alt" (alternative) text for images, and so on.

[edit]

PDF on the Web

Because HTML/XHTML rendering across web browsers has historically been inconsistent and sometimes unpredictable, PDF use online is becoming increasingly common. This is particularly true for order forms, catalogues, brochures, and other documents which are primarily formatted for printing. The ubiquity of the Adobe Reader web browser plugin, however, has inspired some (mostly corporate) web authors to publish a wider variety of information as PDF. This trend is compounded by the simple operation and wide corporate availability of WYSIWYG PDF authoring tools. While the end user experience of an XHTML document can vary significantly depending on browser, platform, and screen resolution, a PDF file can be reasonably expected to look exactly the same to every viewer.

Critics of this practice cite several reasons for avoiding it. Accessibility, particularly by the blind or sight-impaired is a common issue [8]. PDF files tend to be significantly larger than XHTML/SVG files presenting the same information, making it difficult or impossible for users with low-bandwidth connections to view them. Adobe Acrobat Reader, the de facto standard PDF viewer, has historically been slow to start and caused browser instability, particularly when run alongside other browser plugins (though the release of Adobe Reader 7 addressed many of these concerns).

PostScript standard and PDF, and is sometimes referred to as Display PDF. Due to OS support, all OS X applications can create PDF documents automatically as long as they support the Print command.

[edit]

PDF and accessibility

PDF can be accessible to people with disabilities. Current PDF file formats can include tags (essentially XML), text equivalents, captions and audio descriptions, and other accessibility features. Some software, such as Adobe InDesign, can output tagged PDFs automatically. Leading screen readers, including Jaws, Window-Eyes, and Hal, can read tagged PDFs; current versions of the Acrobat and Acrobat Reader programs can also read PDFs out loud. Moreover, tagged PDFs can be reflowed and zoomed for low-vision readers.

However, many problems remain, not least of which is the difficulty in adding tags to existing or "legacy" PDFs; for example, if PDFs are generated from scanned documents, accessibility tags and reflowing are unavailable and must be created either by hand or using OCR techniques. Moreover, that process itself is inaccessible. Nonetheless, well-made PDFs can be a valid choice as long-term accessible documents. (Work is being done on a PDF variant based on PDF 1.4. The PDF/A or PDF-Archive is specifically scaled down for archival purposes.)

Microsoft Word documents can be converted into accessible PDFs, but only if the Word document is written with accessibility in mind - for example, using styles, correct paragraph mark-up and "alt" (alternative) text for images, and so on.

[edit]

PDF on the Web

Because HTML/XHTML rendering across web browsers has historically been inconsistent and sometimes unpredictable, PDF use online is becoming increasingly common. This is particularly true for order forms, catalogues, brochures, and other documents which are primarily formatted for printing. The ubiquity of the Adobe Reader web browser plugin, however, has inspired some (mostly corporate) web authors to publish a wider variety of information as PDF. This trend is compounded by the simple operation and wide corporate availability of WYSIWYG PDF authoring tools. While the end user experience of an XHTML document can vary significantly depending on browser, platform, and screen resolution, a PDF file can be reasonably expected to look exactly the same to every viewer.

Critics of this practice cite several reasons for avoiding it. Accessibility, particularly by the blind or sight-impaired is a common issue [8]. PDF files tend to be significantly larger than XHTML/SVG files presenting the same information, making it difficult or impossible for users with low-bandwidth connections to view them. Adobe Acrobat Reader, the de facto standard PDF viewer, has historically been slow to start and caused browser instability, particularly when run alongside other browser plugins (though the release of Adobe Reader 7 addressed many of these concerns).

Currently, no web browser natively supports PDF, forcing viewers to run a seperate application to access these documents online. Since the PDF specification is not published by the W3C, this is unlikely to change.

[edit]

Searching for a text in a collection of files

Adobe Acrobat Reader 6.0 and above allow searching a collection of PDF files.

Using a search program to search for a text in a collection of files of different types, it may or may not be possible to also search PDF files, depending on the program. This is because the text is stored in coded form, and a program searching for some text must interpret the code and search the result, not just search the code.

Search programs that do not work include that of Windows XP and Agent Ransack. However, for searching the Web, some search engines, such as Google and Yahoo!, include PDF files in searches. The option to view the PDF in HTML format is also commonly offered (this conversion does not include images).

Mac OS X, having PDF as a core element of the operating system, fully supports searching PDF files with the Preview application, used to view PDF files. The Spotlight feature in Mac OS X v10.4 extends this ability across the whole operating system, allowing information in PDF files (as well as almost all others) to be found from a single search box.

On the Windows platform, text in PDF files can be searched using Google Desktop Search and also Windows Desktop Search when installed with an appropriate iFilter available from Adobe.

[edit]

Types of content

A PDF file for e.g. a map is often a combination of vector graphics layer, text, and raster graphics, e.g., the general reference map of the US [9] uses:

- vector graphics for coastlines, lakes, rivers, highways, markings of cities, and Interstate highway symbols — on zooming in, the curves remain sharp, they do not appear as consisting of enlarged pixels (i.e. rectangles of pixels)
- text stored as such — scalable, and also one can copy the text
- raster graphics for showing mountain relief — on zooming in, this consists of enlarged pixels (the blue of the sea and lakes is "filled" neatly to the vector graphics coast line, hence not in raster graphics).

An example of a PDF map without raster graphics is the CIA World Factbook's map of the Arctic. In the same publication's European map, the blue of the sea is not "filled" neatly to the vector graphics coast line, but just raster graphics, giving a cruder result (noticeable when highly zoomed in).

Currently, no web browser natively supports PDF, forcing viewers to run a seperate application to access these documents online. Since the PDF specification is not published by the W3C, this is unlikely to change.

[edit]

Searching for a text in a collection of files

Adobe Acrobat Reader 6.0 and above allow searching a collection of PDF files.

Using a search program to search for a text in a collection of files of different types, it may or may not be possible to also search PDF files, depending on the program. This is because the text is stored in coded form, and a program searching for some text must interpret the code and search the result, not just search the code.

Search programs that do not work include that of Windows XP and Agent Ransack. However, for searching the Web, some search engines, such as Google and Yahoo!, include PDF files in searches. The option to view the PDF in HTML format is also commonly offered (this conversion does not include images).

Mac OS X, having PDF as a core element of the operating system, fully supports searching PDF files with the Preview application, used to view PDF files. The Spotlight feature in Mac OS X v10.4 extends this ability across the whole operating system, allowing information in PDF files (as well as almost all others) to be found from a single search box.

On the Windows platform, text in PDF files can be searched using Google Desktop Search and also Windows Desktop Search when installed with an appropriate iFilter available from Adobe.

[edit]

Types of content

A PDF file for e.g. a map is often a combination of vector graphics layer, text, and raster graphics, e.g., the general reference map of the US [9] uses:

- vector graphics for coastlines, lakes, rivers, highways, markings of cities, and Interstate highway symbols — on zooming in, the curves remain sharp, they do not appear as consisting of enlarged pixels (i.e. rectangles of pixels)
- text stored as such — scalable, and also one can copy the text
- raster graphics for showing mountain relief — on zooming in, this consists of enlarged pixels (the blue of the sea and lakes is "filled" neatly to the vector graphics coast line, hence not in raster graphics).

An example of a PDF map without raster graphics is the CIA World Factbook's map of the Arctic. In the same publication's European map, the blue of the sea is not "filled" neatly to the vector graphics coast line, but just raster graphics, giving a cruder result (noticeable when highly zoomed in).

Tools exist, such as pdfimages (bundled with Xpdf) to extract the raster images from a PDF file. This can be extremely useful if the PDF is simply a collection of scanned pages.

[edit]

See also

- Display PostScript
- Scalable Vector Graphics
- XSL-FO

[edit]

Other Wikipedia articles about tools, utilities and products related to this article

- Ghostscript — Displays PDF files, converts to and from PS.
- iText
- OpenOffice.org — Can transform many types of documents into PDF documents.
- Panda library
- PdfTeX — Generates TeX output directly in PDF.
- PDFCreator — A GPL/AFPL PDF printer driver for Windows.

[edit]

References

This article was originally based on material from the Free On-line Dictionary of Computing, which is licensed under the GFDL.

[edit]

External links

[edit]

Adobe software

- Acrobat, for creating PDFs
- Adobe Reader, for viewing them
- Create Adobe PDF Online — Online service for creating PDF files from many different document types, including Microsoft Word
- Online conversion tools for Adobe PDF documents — The official Adobe online tool for converting from PDF to Text or HTML; also accepts emailed documents

[edit]

Format information

Tools exist, such as pdfimages (bundled with Xpdf) to extract the raster images from a PDF file. This can be extremely useful if the PDF is simply a collection of scanned pages.

[edit]

See also

- Display PostScript
- Scalable Vector Graphics
- XSL-FO

[edit]

Other Wikipedia articles about tools, utilities and products related to this article

- Ghostscript — Displays PDF files, converts to and from PS.
- iText
- OpenOffice.org — Can transform many types of documents into PDF documents.
- Panda library
- PdfTeX — Generates TeX output directly in PDF.
- PDFCreator — A GPL/AFPL PDF printer driver for Windows.

[edit]

References

This article was originally based on material from the Free On-line Dictionary of Computing, which is licensed under the GFDL.

[edit]

External links

[edit]

Adobe software

- Acrobat, for creating PDFs
- Adobe Reader, for viewing them
- Create Adobe PDF Online — Online service for creating PDF files from many different document types, including Microsoft Word
- Online conversion tools for Adobe PDF documents — The official Adobe online tool for converting from PDF to Text or HTML; also accepts emailed documents

[edit]

Format information

- PDF Specification, also available as a book describing PDF 1.4 (ISBN 0201758393)
- Adobe: PostScript vs. PDF
- History of PDF at prepressure.com
- The Camelot Paper — the paper in which John Warnock outlined the project that created PDF

[edit]

Related formats

- PDF/X Frequently asked questions
- PDF/X-3
- PDF-X — Includes PDF/X-1a and PDF/X-3
- AIIM — Information about PDF/A specification for archiving
- Under the Hood of PDF/X-1 by Scott Tully, Vertis, March 21, 2002.

Retrieved from "http://en.wikipedia.org/wiki/Portable_Document_Format"

- PDF Specification, also available as a book describing PDF 1.4 (ISBN 0201758393)
- Adobe: PostScript vs. PDF
- History of PDF at prepressure.com
- The Camelot Paper — the paper in which John Warnock outlined the project that created PDF

[edit]

Related formats

- PDF/X Frequently asked questions
- PDF/X-3
- PDF-X — Includes PDF/X-1a and PDF/X-3
- AIIM — Information about PDF/A specification for archiving
- Under the Hood of PDF/X-1 by Scott Tully, Vertis, March 21, 2002.

Retrieved from "http://en.wikipedia.org/wiki/Portable_Document_Format"