Structure of a PDF file? [closed]

Asked 16 years, 3 months ago Modified 4 years, 7 months ago

Viewed 126k times

85

Closed. This question is seeking recommendations for software libraries, tutorials, tools, books, or other off-site resources. It does not meet Stack Overflow guidelines. It is not currently accepting answers.

We don't allow questions seeking recommendations for software libraries, tutorials, tools, books, or other off-site resources. You can edit the question so it can be answered with facts and citations.

Closed 4 years ago.

Improve this question

For a small project I have to parse pdf files and take a specific part of them (a simple chain of characters). I'd like to use python to do this and I've found several libraries that are capable of doing what I want in some ways.

But now after a few researches, I'm wondering what is the real structure of a pdf file, does anyone know if there is a spec or some explanations anywhere online? I've found a link on adobe but it seems that it's a dead link :(

pdf

Share

Improve this question

Follow

asked Sep 17, 2008 at 23:11

Valentin Jacquemin

2,245 ● 1 ● 19 ● 33

12 Answers

Sorted by: Highest score (default)

51

Here is a link to Adobe's reference material

http://www.adobe.com/devnet/pdf/pdf_reference.html

You should know though that PDF is only about presentation, not structure. Parsing will not come easy.

Share Improve this answer

Follow

answered Sep 17, 2008 at 23:13

minty

22.4k ● 40 ● 91 ● 106

Ok... Greant the link is ok now... When I did my researches I wasn't able to download the last reference.
– Valentin Jacquemin Sep 17, 2008 at 23:19

70 Don't stare at it too long; you'll go insane. – user1228 Sep 17, 2008 at 23:41

7 I am new to work on pdf parsing, and I found some links that I want to share, link1, link2 and link3. – RBK Mar 28, 2015 at 17:44

Not useful as an answer – A X Feb 13, 2021 at 19:04

2 Archived version of the link web.archive.org/web/20210125064408/http://www.adobe.com/devnet/… – jkmartindale Nov 25, 2021 at 19:28

49

I found the GNU Introduction to PDF to be helpful in understanding the structure. It includes an easily readable example PDF file that they describe in complete detail.

Other helpful links:

- PDF Succinctly book is longer and has helpful pictures.
- Introduction to the Insides of PDF is a presentation that isn't as in-depth but gives a quick overview and has lots of pictures.

Share Improve this answer

Follow

edited Jan 5, 2016 at 10:28

vard

4,136 ● 2 ● 28 ● 49

answered Aug 12, 2014 at 15:31

Jeff Moser

20k ● 6 ● 66 ● 85

1 GNU links have gone stale – dwarring Oct 12, 2015 at 2:15

7 @dwarring I fixed them with webarchive links for posterity. – vard Jan 5, 2016 at 10:31

Adobe no longer hosts the intro preso where they used to. Looking for it, I found it here:

mafiadoc.com/slides_59c4bd4c1723dd6ff3e29059.html – LAFK says reinstate Monica Aug 2, 2020 at 13:00

your "other helpul links" are dead :(– T.M15 Feb 27, 2021 at 20:05

25

When I first started working with PDF, I found the PDF reference very hard to navigate. It might help you to know that the overview of the file structure is found in syntax, and what Adobe call the document structure is the object structure and not the file structure. That is also found in Syntax. The description of operators is hidden away in Appendix A - very useful for understanding what is happening in content streams. If you ever have the pain of working with colour spaces you will find that hidden in Graphics! Hopefully these pointers will help you find things more quickly than I did.

If you are using windows, pdftron CosEdit allows you to browse the object structure to understand it. There is a free demo available that allows you to examine the file but not save it.

Share Improve this answer

Follow

answered Sep 18, 2008 at 13:26

danio

8,655 ● 6 ● 49 ● 57

2 +1. Looks like CosEdit is a great introductory browser, not perfect but much better than trying to grep through the raw binary file. :/ – Jason S May 8, 2009 at 21:12

I downloaded CosEdit, but it rejected my PDF. The same PDF is accepted by other programs. CosEdit may be right, but it didn't help me determine what was wrong with my PDF. – LarsH Dec 20, 2013 at 19:28

10

Here's the raw reference of PDF 1.7, and here's an article describing the structure of a PDF file. If you use Vim, the pdftk plugin is a good way to explore the document in an ever-so-slightly less raw form, and the pdftk utility itself (and its GPL source) is a great way to tease documents apart.

Share Improve this answer

Follow

answered Sep 17, 2008 at 23:18

jmah

2,216 ● 14 ● 16

1 The raw reference seems pointless. It contains only a single page? – Carcamano Jan 21, 2016 at 14:10

@Carcamano The raw reference is a (large) package with a number of attachments. The first attachment describes the PDF format and is 1310 pages long. – banbh Nov 10, 2018 at 20:35

7

I'm trying to do pretty much the same thing. The PDF reference is a very difficult document to read. This tutorial is a better start I think.

Share Improve this answer

Follow

answered Jul 9, 2009 at 7:13

Noran

Probably the newer link: planetpdf.com/… – aderchox May 17, 2020 at 9:34

6

This may help shed a little light: (from page 11 of PDF32000.book)

> PDF syntax is best understood by considering it as four parts, as shown in Figure 1:
>
> • Objects. A PDF document is a data structure composed from a small set of basic types of data objects. Sub-clause 7.2, "Lexical Conventions," describes the character set used to write objects and other syntactic elements. Sub-clause 7.3, "Objects," describes the syntax and essential properties of the objects. Sub-clause 7.3.8, "Stream Objects," provides complete details of the most complex data type, the stream object.
>
> • File structure. The PDF file structure determines how objects are stored in a PDF file, how they are accessed, and how they are updated. This structure is independent of the semantics of the objects. Sub- clause 7.5, "File Structure," describes the file structure. Sub-clause 7.6, "Encryption," describes a file-level

> mechanism for protecting a document's contents from unauthorized access.
>
> • Document structure. The PDF document structure specifies how the basic object types are used to represent components of a PDF document: pages, fonts, annotations, and so forth. Sub-clause 7.7, "Document Structure," describes the overall document structure; later clauses address the detailed semantics of the components.
>
> • Content streams. A PDF content stream contains a sequence of instructions describing the appearance of a page or other graphical entity. These instructions, while also represented as objects, are conceptually distinct from the objects that represent the document structure and are described separately. Sub-clause 7.8, "Content Streams and Resources," discusses PDF content streams and their associated resources.

Looks like navigating a PDF file will require a little more than a passing effort.

answered Jul 30, 2011 at 3:54

Josh Albert

1,124 ● 13 ● 16

5

If You want to parse PDF using Python please have a look at PDFMINER. This is the best library to parse PDF files till date.

Share Improve this answer

Follow

answered Sep 17, 2013 at 11:54

codingscientist

1,136 ● 1 ● 11 ● 12

1 PDFMiner is great. Especially try `pdf2txt -t html -d -Y exact -o foo.html foo.pdf`. It's a pretty good tool for getting a look at the structure of a PDF page. I'm also working on some improvements to it, for our own project. – LarsH Dec 20, 2013 at 20:09

4

Didier have a tool to parse the PDF:

http://didierstevens.com/files/software/pdf-parser_V0_4_3.zip

or here:

http://blog.didierstevens.com/programs/pdf-tools/ which cataloged several related pdf-analysis tools.

Another tool is here:

http://mshahzadlatif.wordpress.com/2011/09/28/view-pdf-structure-using-adobe-acrobat-or-a-free-tool-called-pdfxplorer/

Share Improve this answer

Follow

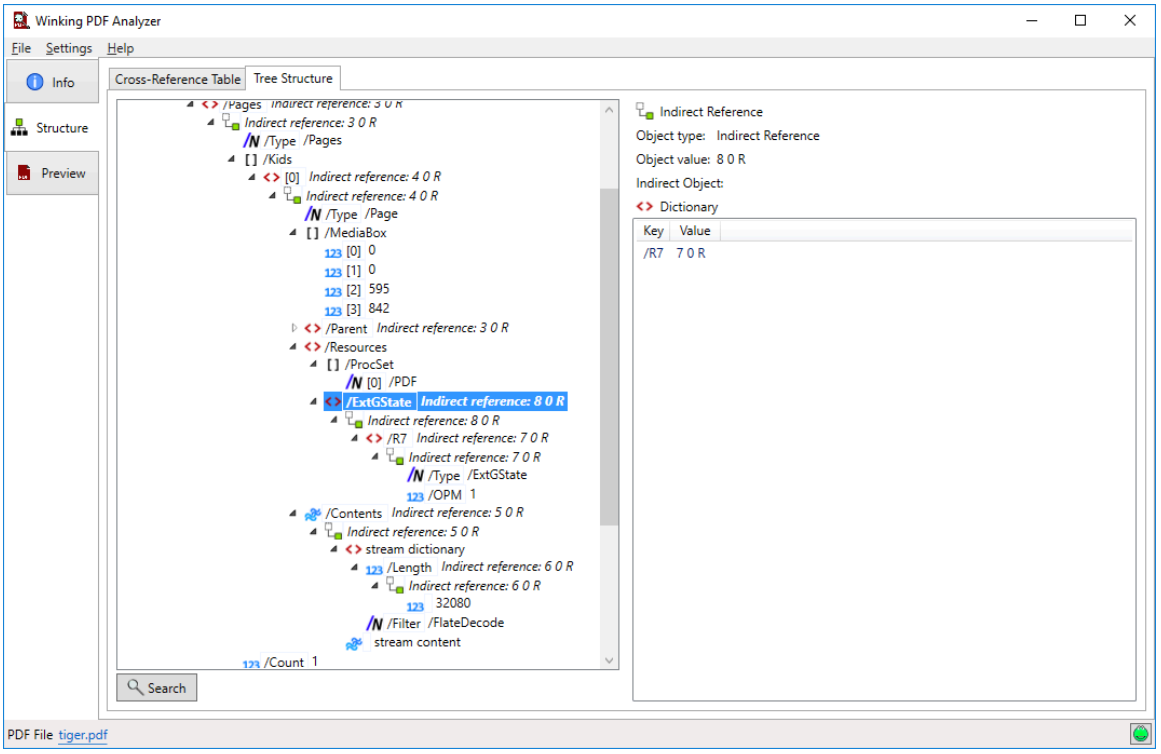
answered Mar 2, 2014 at 3:44

Peter Teoh
6,665 ● 4 ● 44 ● 59

3

You need the PDF Reference manual to start reading about the details and structure of PDF files. I suggest to start with version 1.7.

On windows I used a free tool PDF Analyzer to see the internal structure of PDF files. This will help in your understanding when reading the reference manual.



(I'm affiliated with PDF Analyzer, no intention to promote)

Share Improve this answer

Follow

edited Jul 10, 2019 at 8:21

answered Dec 17, 2018 at 8:06

juFo
18.6k ● 10 ● 109 ● 151

PDF has been an ISO standard for 10 years now. Thus, shouldn't one suggest starting with the ISO document instead of with the Adobe PDF Reference, in particular as Adobe published a copy of ISO 32000-1 (with exchanged page headers) for free? – mkl Dec 17, 2018 at 14:25

as a start the PDF Reference manual will give you a good understanding of the basics. Once you have mastered them you can read the ISO, this will give you an insight on why some changes have been made. The basic parsing will still be the same when reading the Reference Manual. As an advise it is good to read also multiple versions of manuals as they give sometimes subtle changes. – juFo Dec 17, 2018 at 15:45

Indeed it can make sense to read different versions of the documentation on some topic but in my eyes one should start with the current version and not archaic ones. – mkl Dec 17, 2018 at 21:30

2

Extracting text from PDF is a hard problem because PDF has such a layout-oriented structure. You can see the docs and source code of my barely-successful attempt on CPAN (my implementation is in Perl). The PDF data structure is very cool and well designed, but it's easier to write than read.

Share Improve this answer

Follow

answered Sep 19, 2008 at 2:51

Chris Dolan
8,965 ● 2 ● 37 ● 73

2

One way to get some clues is to create a PDF file consisting of a blank page. I have CutePDF Writer on my computer, and made a blank Wordpad document of one page. Printed to a .pdf file, and then opened the .pdf file using Notepad.

Next, use a copy of this file and eliminate lines or blocks of text that might be of interest, then reload in Acrobat Reader. You'd be surprised at how little information is needed to make a working one-page PDF document.

I'm trying to make up a spreadsheet to create a PDF form from code.

Share Improve this answer

Follow

answered Aug 24, 2010 at 16:52

Daniel Kim

21 • 1

0

To extract text from a PDF, try this on Linux, BSD, etc. machine or use Cygwin if on Windows:

```
pdfinfo -layout some_pdf_file.pdf
```

A plain text file named `some_pdf_file.txt` is created. The simpler the PDF file layout, the more straightforward the .txt file output will be.

Hexadecimal characters are frequently present in the .txt file output and will look strange in text editors. These hexadecimal characters usually represent curly single

and double quotes, bullet points, hyphens, etc. in the PDF.

To see the context where the hexadecimal characters appear, run this grep command, and keep the original PDF handy to see what character the codes represent in the PDF:

```
grep -a --color=always "\\\\[0-9][0-9][0-9]"
some_pdf_file.txt
```

This will provide a unique list of the different octal codes in the document:

```
grep -ao "\\\\[0-9][0-9][0-9]"
some_pdf_file.txt|sort|uniq
```

To convert these hexadecimal characters to ASCII equivalents, a combination of grep, sed, and bc can be used, I'll post the procedure to do that soon.

Share Improve this answer

Follow

answered Jul 26, 2019 at 12:28

keithchristian

5 ● 5

You mean `pdftotext -layout` (pdfinfo only gives you the header and page infos) – thilo Nov 6, 2023 at 20:39