How to use PDFTOTEXT to convert PDFs into text

Jeff Porter
IRE and NICAR
jeff@ire.org

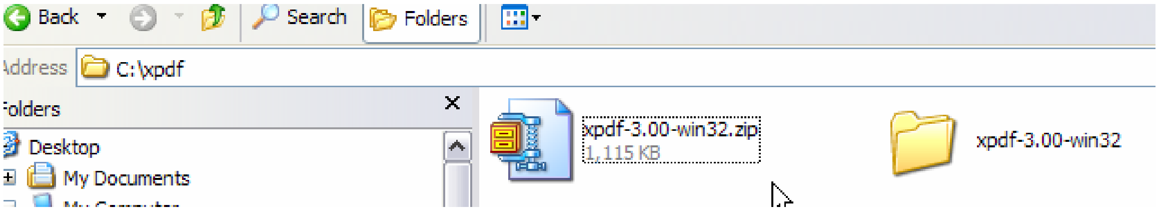
You've seen the Acrobat files containing columns and rows of numbers and other information, waiting to be analyzed. The problem: Acrobat files – ending with .PDF – can't be directly read into spreadsheet or database software. Now, there's a tool to make a PDF into formatted text, ready to import.

How to 'install' the software

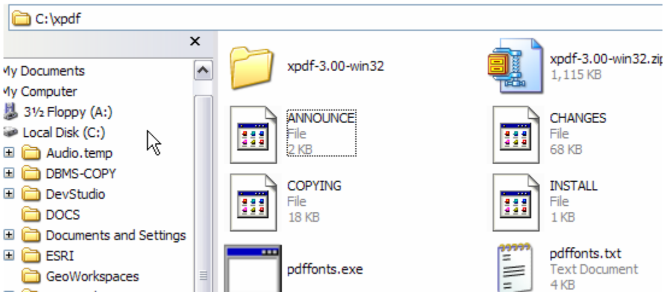
Make a folder on a hard drive of your machine. For our example, we'll call it c:\xpdf.

The Web page to download this file is http://www.foolabs.com/xpdf/download.html. If you're a typical Windows user, scroll down to "Precompiled binaries" and you'll see a paragraph starting with "x86, DOS/Win32." Choose the first download link that that paragraph.

Download that file into your c:\xpdf folder. Unzip the file as you would unzip any compressed file. It will place it in a subfolder under your c:\xpdf folder like this:



For simplicity's sake, move all the files inside the subfolder called xpdf-3.00-win32 into the main folder your created earlier: c:\xpdf. It should now look like this:

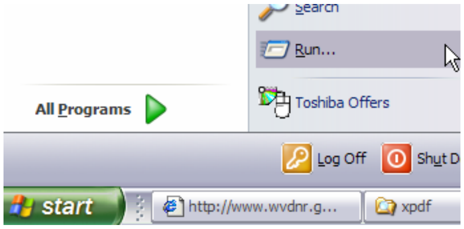


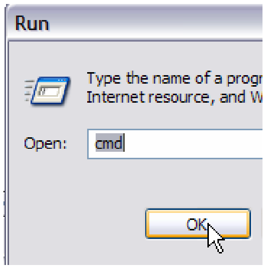
That's it – the software is "installed!"

How to use PDFTOTEXT

Find the PDF you want to convert. Ideally, the file is simple a table of columns and rows; not that PDFTOTEXT can't handle it other types, but that will save some cleanup later. For example, you can download this PDF file about homeland security grants: http://www.dhs.gov/dhspublic/interweb/assetlibrary/Grants-ODP-04.pdf. Save it in the same folder you created: c:\xpdf.

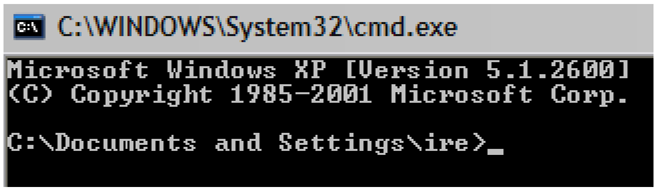
Now, you're ready to convert. Go to Start > Run… and when you get that Run window, type in cmd and hit OK.





Open: cmd

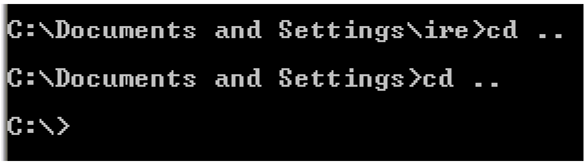
You'll get this little command window (yours might look a little different):



```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\ire>_
```

You'll need to navigate to the right folder. If you start on the same drive, but have some folders identified as the example above, type in cd .. and hit Enter. Do it until you get this:

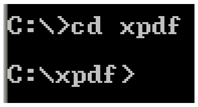


```
C:\Documents and Settings\ire>cd ..

C:\Documents and Settings>cd ..

C:\>
```

Now change your directory to c:\xpdf with another "cd" command. Your screen should look something like this:



```
C:\>cd xpdf

C:\xpdf>
```

Now you're ready to run this simple command. Type it in just like this and hit Enter:

```
C:\xpdf>pdftotext -layout Grants-ODP-04.pdf
```

Let’s parse that out: The first part (c:\xpdf>) simply identifies the folder you’re looking into. The next word (pdftotext) is the name of the program you’re telling the computer to use. The next (-layout) is telling that program to use an option to retain the appearance, spacing and all, of the PDF file you’re converting. The last part (Grants-ODP-04.pdf) is the name of the file we downloaded.

When it’s done, you’ll get back to the c:\xpdf> line. Type “exit” and hit Enter. Now, go to that folder and there you’ll find a text file named Grants-ODP-04.txt. Open it up with Notepad.

Compare the PDF with the text file, and you’ll notice that the layout is quite similar. For using it in Excel, you should run the import wizard three times, since each page is structured slightly differently.

In other examples, especially if the PDF has lots of blank spaces, you’ll get few numbers slightly off-kilter. A text editor like Notepad, or even a word processor like Word, can help you fix that type of problem. For more heavy-duty text editing, consider using one like UltraEdit (www.ultraedit.com) or VEdit (www.vedit.com).

A few more caveats

It won’t work at all if the file is a scanned image (for example, a 990 form from www.guidestar.org).

If the file is encrypted, PDFTOTEXT won’t convert it.

The more complicated the layout, the more text editing you’ll need to do later.

Always print out at least part of the PDF so you can easily compare the two.

More details

For the ultra-geek information, read the accompanying text file called PDFTOTEXT.TXT, located in your c:\xpdf folder. Note that you can tell PDFTOTEXT to start and stop at certain page numbers in the PDF.

Now, importing the text file

Close the Notepad file and start Excel. Navigate to the correct folder and go File > Open … and choose Text under the dialog box of Files of Type. Identify your file and tell Excel to open. That’ll start the Import Wizard.

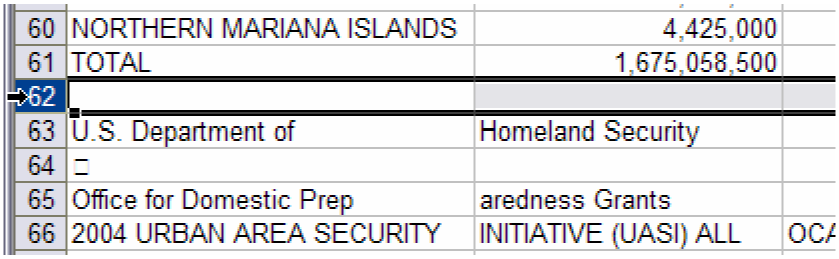
Excel should guess correctly that it’s fixed width. Note that below, it’ll ask you where to start your import. For reading in those PDFs, you’ll often choose another row to start with. In this case, let’s choose to start at row 4, where the headers of the data begin. Click Next.

That takes you to a new dialog to let you choose to make, delete or adjust columns. Before you click Next, scroll down entirely. You’ll catch a couple of problems. First, note that one of the “state” names, Northern Mariana Islands, is too long. Click on the break line and move it to position 24. Second, note below that, there are two other sets of numbers. For now, ignore them. We’ll have to deal with that material in two more imports. Now, click Next.

We’re almost there, and it’s tempting to just click Finish. Resist the temptation and review a couple of things. This is the section in which you can made decisions on whether a column is text, numbers or a date. “General” means you’ll allow Excel to guess. It’s correct most of the time; it’ll guess correctly that the dollar amounts listed are numbers. It’ll also correctly guess that the state names are text. If there is any question, you can change it by click in any column and choose a specific format.

Now click Finish, and you’re almost there.

Scroll down to row 62, and you’ll see that the second set of numbers is formatted differently. So let’s get rid of those by click on the number 62 itself and, still holding the left mouse button down, scroll down until you get the bottom of the file, highlighting each row.



60	NORTHERN MARIANA ISLANDS	4,425,000	
61	TOTAL	1,675,058,500	
62			
63	U.S. Department of	Homeland Security	
64	□		
65	Office for Domestic Prep	aredness Grants	
66	2004 URBAN AREA SECURITY	INITIATIVE (UASI) ALL	OC

Go Edit > Delete, and the problem rows will disappear. You’re done! Now, start another worksheet, use the Import Wizard again, choose row 71 and bring in the next set of numbers, then make one more pass to capture the third set. After you’re finished, you’re a true expert!