

# Re: OT: Copying text from a PDF

---

*Source:* <http://sci.tech-archive.net/Archive/sci.electronics.design/2005-06/msg00358.html>

---

- *From:* Rich Grise <[richgrise@xxxxxxxxxxx](mailto:richgrise@xxxxxxxxxxx)>
  - *Date:* Fri, 03 Jun 2005 03:55:53 GMT
- 

On Thu, 02 Jun 2005 23:16:06 +0100, Terry Pinnell wrote:

> Rich Grise <[richgrise@xxxxxxxxxxx](mailto:richgrise@xxxxxxxxxxx)> wrote:

>

>>On Wed, 01 Jun 2005 07:08:01 +0000, mroberds wrote:

>>

>>> Terry Pinnell <[terrypinDELETE@xxxxxxxxxxxxxxxxxxxxxxxx](mailto:terrypinDELETE@xxxxxxxxxxxxxxxxxxxxxxxx)> wrote:

>>>>Quite often I have trouble extracting text from a PDF. I use the Text

>>>>tool, copy, but on then pasting into my text editor I get garbage.

>>>>

>>>> What's your text editor? Assuming you're under Windows, perhaps the

>>>> problem is trying to paste Unicode into an editor that can't handle it.

>>>> You might try pasting the text into Word or Wordpad to see what happens.

>>>>

>>>> You might also look at xpdf, <http://www.foolabs.com/xpdf/> . I don't

>>>> think you can run the PDF viewer under Windows, but the command-line

>>>> utilities, including a PDF-to-text converter, will work.

>>>>

>>>>Barely a day goes by that Slackware doesn't pleasantly surprise me!

>>>>It seems I got xpdf along with it, and lo and behold:

>>>>-----

>>>>30A, 50V, 0.040 Ohm, N-Channel Power

>>>>MOSFET

>>>>This is an N-Channel enhancement mode silicon gate power field effect

>>>>transistor designed for applications such as switching regulators,

>>>>switching converters, motor drivers, relay drivers and drivers for high

>>>>power bipolar switching transistors requiring high speed and low gate

>>>>drive power. This type can be operated directly from integrated circuits.

>>>>Formerly developmental type TA9771.

>>>>Ordering Information

>>>> PART NUMBER PACKAGE BRAND

>>>> BUZ11 TO-220AB BUZ11

>>>>NOTE: When ordering, use the entire part number.

>>>>

>>>>Features

>>>>· 30A, 50V

>>>>· rDS(ON) = 0.040

>>>>· SOA is Power Dissipation Limited

>>>>· Nanosecond Switching Speeds

>>> Linear Transfer Characteristics  
>>> High Input Impedance  
>>> Majority Carrier Device  
>>> Related Literature  
>> - TB334 "Guidelines for Soldering Surface Mount  
>> Components to PC Boards"  
>>Symbol  
>> D  
>> G  
>> S  
>>-----  
>>  
>>Cheers!  
>>Rich  
>  
> Thanks for the text paste.  
>  
> Must say I'm a bit lost on that site  
> <http://www.foolabs.com/xpdf/home.html>  
> Can you help me locate specifically the PDF to text converter please?  
> I'm wallowing in files with off-putting and Windows-alien names like  
> 't1lib-1.3.tar.gz'.

Click "Download" to get to <http://www.foolabs.com/xpdf/download.html> ,  
then scroll down to "Precompiled binaries" and it's probably either  
xpdf-3.00p13-win32.zip (1142558 bytes) for Win32 or  
xpdf-3.00p13-dos6.zip (1775202 bytes) for DOS.

Here's a chunk of the README for the win32 version, copied without  
permission:

-----<begin excerpt>-----

Xpdf

====

version 3.00  
2004-jan-22

The Xpdf software and documentation are  
copyright 1996-2004 Glyph & Cog, LLC.

Email: [derekn@xxxxxxxxxxx](mailto:derekn@xxxxxxxxxxx)  
WWW: <http://www.foolabs.com/xpdf/>

The PDF data structures, operators, and specification are  
copyright 1985-2003 Adobe Systems Inc.

What is Xpdf?  
-----

Xpdf is an open source viewer for Portable Document Format (PDF)

## Re: OT: Copying text from a PDF

files. (These are also sometimes also called 'Acrobat' files, from the name of Adobe's PDF software.) The Xpdf project also includes a PDF text extractor, PDF-to-PostScript converter, and various other utilities.

Xpdf runs under the X Window System on UNIX, VMS, and OS/2. The non-X components (pdftops, pdftotext, etc.) also run on Win32 systems and should run on pretty much any system with a decent C++ compiler.

Xpdf is designed to be small and efficient. It can use Type 1 or TrueType fonts.

### Distribution

-----

Xpdf is licensed under the GNU General Public License (GPL), version 2. In my opinion, the GPL is a convoluted, confusing, ambiguous mess. But it's also pervasive, and I'm sick of arguing. And even if it is confusing, the basic idea is good.

In order to cut down on the confusion a little bit, here are some informal clarifications:

- I don't mind if you redistribute Xpdf in source and/or binary form, as long as you include all of the documentation: README, man pages (or help files), and COPYING. (Note that the README file contains a pointer to a web page with the source code.)
- Selling a CD-ROM that contains Xpdf is fine with me, as long as it includes the documentation. I wouldn't mind receiving a sample copy, but it's not necessary.
- If you make useful changes to Xpdf, please make the source code available -- post it on a web site, email it to me, whatever.

If you're interested in commercial licensing, please see the Glyph & Cog web site:

<http://www.glyphandcog.com/>

### Compatibility

-----

Xpdf is developed and tested on a Linux 2.4 x86 system.

In addition, it has been compiled by others on Solaris, AIX, HP-UX, Digital Unix, Irix, and numerous other Unix implementations, as well as VMS and OS/2. It should work on pretty much any system which runs X11 and has Unix-like libraries. You'll need ANSI C++ and C compilers

to compile it.

The non-X components of Xpdf (pdftops, pdftotext, pdfinfo, pdffonts, pdftoppm, and pdfimages) can also be compiled on Win32 systems. See the Xpdf web page for details.

If you compile Xpdf for a system not listed on the web page, please let me know. If you're willing to make your binary available by ftp or on the web, I'll be happy to add a link from the Xpdf web page. I have decided not to host any binaries I didn't compile myself (for disk space and support reasons).

If you can't get Xpdf to compile on your system, send me email and I'll try to help.

Xpdf has been ported to the Acorn, Amiga, BeOS, and EPOC. See the Xpdf web page for links.

#### Getting Xpdf

---

The latest version is available from:

<http://www.foolabs.com/xpdf/>

or:

<ftp://ftp.foolabs.com/pub/xpdf/>

Source code and several precompiled executables are available.

Announcements of new versions are posted to several newsgroups (comp.text.pdf, comp.os.linux.announce, and others) and emailed to a list of people. If you'd like to receive email notification of new versions, just let me know.

#### Running Xpdf

---

To run xpdf, simply type:

```
xpdf file.pdf
```

To generate a PostScript file, hit the "print" button in xpdf, or run pdftops:

```
pdftops file.pdf
```

To generate a plain text file, run pdftotext:

pdftotext file.pdf

There are four additional utilities (which are fully described in their man pages):

pdffinfo — dumps a PDF file's Info dictionary (plus some other useful information)

pdfmfonts — lists the fonts used in a PDF file along with various information for each font

pdftoppm — converts a PDF file to a series of PPM/PGM/PBM-format bitmaps

pdfimages — extracts the images from a PDF file

Command line options and many other details are described in the man pages (xpdf.1, etc.) and the VMS help files (xpdf.hlp, etc.).

-----<end excerpt>-----

Good Luck!

Rich

---

• *Follow-Ups:*

◆ **[Re: OT: Copying text from a PDF](#)**

◇ From: Terry Pinnell

• *References:*

◆ **[OT: Copying text from a PDF](#)**

◇ From: Terry Pinnell

◆ **[Re: OT: Copying text from a PDF](#)**

◇ From: mroberds

◆ **[Re: OT: Copying text from a PDF](#)**

◇ From: Rich Grise

◆ **[Re: OT: Copying text from a PDF](#)**

◇ From: Terry Pinnell

• Prev by Date: **[keypad scanner](#)**

• Next by Date: **[Re: OT: How can I glue a broken nylon gear?](#)**

• Previous by thread: **[Re: OT: Copying text from a PDF](#)**

• Next by thread: **[Re: OT: Copying text from a PDF](#)**

• Index(es):

◆ **[Date](#)**

◆ **[Thread](#)**