LINUX.COM EDITORIAL STAFF APRIL 27, 2006

Manipulating PDFs with the PDF Toolkit

Creating and reading PDF files in Linux is easy, but manipulating existing PDF files is a little trickier. Countless applications enable you to fiddle with PDFs, but it's hard to find a single application that does everything. The PDF Toolkit (pdftk) claims to be that all-in-one solution. It's the closest thing to Adobe Acrobat that I've found for Linux.

Developer Sid Steward describes pdftk as the PDF equivalent of an "electronic staple remover, hole punch, binder, secret decoder ring, and X-ray glasses." That's a lot of functionality for a 4MB application, but the software delivers. Pdftk can join and split PDFs; pull single pages from a file; encrypt and decrypt PDF files; add, update, and export a PDF's metadata; export bookmarks to a text file; add or remove attachments to a PDF; fix a damaged PDF; and fill out PDF forms. In short, there's very little pdftk can't do when it comes to working with PDFs.

You can download pdftk 1.12 as source or as a Debian or RPM package, FreeBSD port, or Gentoo Ebuild. Binaries are available for Windows and Mac OS X too. If you decide to compile pdftk, as I did, check the build notes before you begin, in order to find out about any dependencies for your Linux distro or your platform. The compilation process only took a few minutes on my computer, and there were no hitches.

Pdftk is a command-line tool, and the syntax can be complicated, especially for complex actions such as removing specific pages from a PDF file. You can expect to do a lot of typing, but that shouldn't put you off using the tool.

I put pdftk through its paces with a number of PDFs that ranged in size from 30KB to 2MB. I focused on the functions that I use most with other PDF software: joining and splitting PDFs, removing pages from a PDF, and

attaching files to a PDF. Except for one or two very minor issues, I wasn't disappointed with the results. Pdftk also produced output far more quickly than most other PDF tools that I've worked with.

Joining files

Pdftk's ability to join two or more PDF files is on par with such specialized applications as pdfmeld and joinPDF (discussed in this article). The command syntax is simple:

```
pdftk file1.pdf file2.pdf cat output newFile.pdf
```

`cat` is short for concatenate -- that is, *link together*, for those of us who speak plain English -- and `output` tells pdftk to write the combined PDFs to a new file.

Pdftk doesn't retain bookmarks, but it does keep hyperlinks to both destinations within the PDF and to external files or Web sites. Where some other applications point to the wrong destinations for hyperlinks, the links in PDFs combined using pdftk managed to hit each link target perfectly.

Splitting files

Splitting PDF files with pdftk was an interesting experience. The `burst` option breaks a PDF into multiple files -- one file for each page:

```
pdftk user_guide.pdf burst
```

I don't see the use of doing that, and with larger documents you wind up with a lot of files with names corresponding to their page numbers, like pg_0001 and pg_0013 -- not very intuitive.

On the other hand, I found pdftk's ability to remove specific pages from a PDF file to be useful. For example, to remove pages 10 to 25 from a PDF file, you'd type the following command:

```
pdftk myDocument.pdf cat 1-9 26-end output removedPages.pdf
```

I have used this syntax extensively to trim pages from work samples that I have posted on my company's Web site, and to extract articles from back issues of a magazine to which I contribute. The resulting files are small, and the PDFs retain excellent resolution.

Adding attachments

When I moved to Linux from Windows in 1999, I missed Adobe Acrobat's ability to attach files to a PDF. I regularly used this feature to include addenda, surveys, or additional information with a published PDF. Until I found pdftk, I was forced to move my PDF documents to a Windows box whenever I needed to attach a file.

Why attach a file to a PDF instead of sending an archive? The major appeal is convenience. If you move a PDF from one computer to another, and don't move the archive along with it, you won't have access to the attachments. And instead of pulling a file from an archive to view it, you just double-click on the attachment's icon to open the file from your PDF viewer.

Pdftk can attach binary and text files to a PDF with ease. You can even specify what page of the PDF you want the attachment to appear on. For example:

```
pdftk html_tidy.pdf attach_files command_ref.html to_page 24 output
html_tidy_book.pdf
```

I have attached OpenOffice.org Writer documents, tar.gz archives, and text and HTML files to various PDF documents, and aside from a noticeable increase in the size of the PDF file, there were no nasty side effects.

Attached files are denoted by a thumbtack icon in the PDF, but only in Adobe's Acrobat Reader. Attachments don't appear in Xpdf, Evince, KPDF, or gv.

Filling out forms

Most PDF files are static -- you read them, print them out, or copy text from them. But PDFs can also be interactive. It's possible to create PDF forms with fields that accept information. Companies and government departments post PDF forms on their Web sites to collect survey information and customer feedback, and even to submit tax returns.

Using pdftk's `fill_form` option, you can fill out forms using information in a separate file. However, the `fill_form` option isn't for the faint of heart. To perform this task, you need to create a Form Data Format (FDF) file containing the data that you want to merge into the form. You can do this using pdftk's `generate_fdf` directive.

The FDF file contains the names of each field in the PDF and the values you want to enter into those fields. The FDF file also contains a link to the name of the PDF form. An FDF file looks something like this:

```
%FDF-1.2
1 0 obj
<< /FDF
  << /Fields
   [ << /T (Name_field) /V (Fred Langan) >>
     << /T (Address_field) /V (1313 Mockingbird Lane) >>
     << /T (Age_field) /V (53) >>]
     /F (info_form.pdf)
  >>
>>
endobj
trailer
<< /Root 1 0 R >>
%%EOF
```

To fill out the form using an FDF file, use a command like this:

```
pdftk survey_form.pdf fill_form survey_answers.fdf output filled_survey.pdf
```

Unless you're comfortable creating FDF files, the `fill_form`option isn't really suited for completing the odd form here and there. However, if you're feeling adventurous, the book *PDF Hacks* explains how to use pdftk and a Web server running PHP to do this with Web-based forms.

A couple of infrequently used options

Pdftk has a number of options that you might use infrequently, but that are very useful when you need them -- such as `update_info` and `user_pw`.

When you create a PDF, it might contain no or incomplete metadata -- that is, information describing the PDF. Metadata can come in handy when you or your users need to organize or index a set of PDF files. Using pdftk and a text file, you can change or add metadata to the PDF:

```
pdftk DocBook_Overview.pdf update_info data.txt output DocBookOverview.pdf
```

In this usage, the contents of the file data.txt consist of an InfoKey and InfoValue pair, like this:

```
InfoKey: Keywords
InfoValue: DocBook,writing,documentation,background
```

You can change only the following metadata items with pdftk: title, author, subject, producer, and keywords.

If you're working with PDFs that contain sensitive information, you may want to require a password to read the PDF. If you want to make sure that only certain people can view a PDF, you can apply a password to it with the `user_pw`option:

```
pdftk sales_report.pdf output SalesReport.pdf user_pw PROMPT
```

You will be prompted for a password of up to 32 characters. When someone tries to open the PDF, they will be asked to enter a password.

If you use pdftk regularly, or if you're comfortable writing scripts to encapsulate the commands that you use, then you should have no problems working from the command line. Otherwise, check out Dirk Paehl's graphical front end for pdftk, GUI for PDFTK. It isn't the prettiest or most intuitive GUI around, but it does give you quick access to all of pdftk's functions.

Conclusion

Pdftk is one of the most useful tools for manipulating PDF files. It does as good a job as the single-function PDF tools available for Linux, and often the results are better.

Pdftk's flexibility is unmatched on Linux. While it's not the easiest software, with a bit of practice you'll get the hang of it. The pdftk Web site contains a number of useful tips and tricks.

Chances are you'll use only a handful of pdftk's features regularly. But when you need to call on some of pdftk's other functions, for things like repairing a PDF file or filling out PDF forms, you'll be glad you have this application on your hard drive.

Scott Nesbitt is a technical writer and journalist who spends way too much time fooling around with PDFs (and other types of documents).