Pay No Attention to that Man Behind the Curtain
PS, PDFs, and Data Smuggling Oh My!

Matt Davis
(enferex)

Ruxmon July 2012
mattdavis9@gmail.com

July 27 2012

Act I: The Beginning

Act I: The Beginning ...
... or how I came to love the PDF.

I am human, I like things...

I like figuring out how things work...

I also like animals...

I also like animals...

I also like animals...

SHIEP

SHIEP Handles Information Extraction and Processing

- Project for work (2008), where I was the sole code monkie.
- Good excuse to learn about PDFs ... and get paid for it.
- Goal: Locate tables in PDFs, and then populate a database with that information.
- Personal Goal: Not using a PDF parsing library, but write my own.
- This was 2008, not many free and open source PDF libraries existed.
- Not relying on an external library enhances portability :-) YAY!
- I also had to learn about PDFs... in other words: I had to get the specification and read!

PDF Specification

- PDF Reference Sixth Edition: Adobe Portable Document Format.
- Published November 2006.
- Covers PDF version 1.7 still current today.
- Free to download... not behind a pay wall.
- Current ISO Standard: ISO 32000-1:2008.
- But this nugget is what caught my eye.....

PDF Specification: Interesting Piece...

2.2.7 Incremental Update

Applications may allow users to modify PDF documents. Users should not have to wait for the entire file—which can contain hundreds of pages or more—to be rewritten each time modifications to the document are saved. PDF allows modifications to be appended to a file, leaving the original data intact. The addendum appended when a file is incrementally updated contains only those objects that were actually added or modified, and includes an update to the cross-reference table. Incremental update allows an application to save modifications to a PDF document in an amount of time proportional to the size of the modification rather than the size of the file.

In addition, because the original contents of the document are still present in the file, it is possible to undo saved changes by deleting one or more addenda. The ability to recover the contents of an original document is critical when digital signatures have been applied and subsequently need to be verified.

This will become clear in a minute...

Act II: PS PDF WTF?

PS: Adobe PostScript

- Programming language originating in 1982 at Adobe.
- "Page description language."
- Can describe graphics and fonts/text on pages.
- Turing complete, so the language can be used outside the graphics, printing and page layout domain.
- Stack based.
- PostScript printer drivers interpret the PS code and convert it to graphics and text.
- Wanna play with PS as a language, try the free and open source Ghostscript (chances are you already have this).

PS: Adobe PostScript Example

Code snippit using Ghostscript interpreter. Based on example in the Adobe PS Language Manual (see Resources slide).

```
/Times-Roman findfont
100 scalefont
setfont
0 0 moveto
(Ruxmon!!!) show
```

What it does:

- Find the Times-Roman font and place it on the stack. [Times-Font]
- Place 100 on the stack. [100, Times-Font]
- Call the 'scalefont' operator and use the two values on the stack as arguments. This returns a font onto the stack as a result. [Font]
- Call the 'setfont' operator on the scaled font on the stack [Font]. This pops the Font off the stack, and returns nothing to the stack. []
- Push both '0' and '0' values onto the stack and call the 'moveto' operator on the stack [0 0], which pops both 0s from the stack. []
- Push some awesome text onto the stack and call 'show' to display the text as graphics.

PS: Adobe PostScript Security and ...

- Recall: most printers have a PS interpreter to print documents in oh so awesome ways.
- Recall: PS is Turing complete.
- Conclusion: What happens if you send a PS file that is nothing more than the equivalent to a busy-wait/infinite-loop/really-really-pointless operation?

... Security and Printer Hell

- PS can perform File IO (read/write files).
- PS can execute commands using your system shell:
  (%pipe% whoami > somefile)
- Ancient vulnerability:
  http://www.cert.org/advisories/CA-1995-10.html
- Printer Hell? Infinite loop in one line of PS:
  {} *loop*
- NOTE: This might force interpreting on your local machine first, therefore putting your PS interpreter into a never ending loop.

PDF Genesis... What is a PDF?

PS

PDF Genesis... What is a PDF?

PS +

PDF Genesis... What is a PDF?

PS +

PDF Genesis... What is a PDF?

PS +

=

PDF Genesis... What is a PDF?

PS + 

= PDF

PDF Genesis... What is a PDF?

PDF: Adobe Portable Document Format

- Originated by Adobe in 1993.
- A subset of PS for graphics and positioning.
- A page description language.
- Fonts can be embedded into the document, allowing anyone to view the document as it was intended.
- Embed other data formats as "objects" inside the document, such as: JPEGs, JavaScript, flash, etc.
- Adds the ability to have form data.
- Adds the ability to compress data.
- Adds the ability to encrypt data.
- You can think of it as a really really beefed-up PostScript wrapper.

PDF Basic File Structure

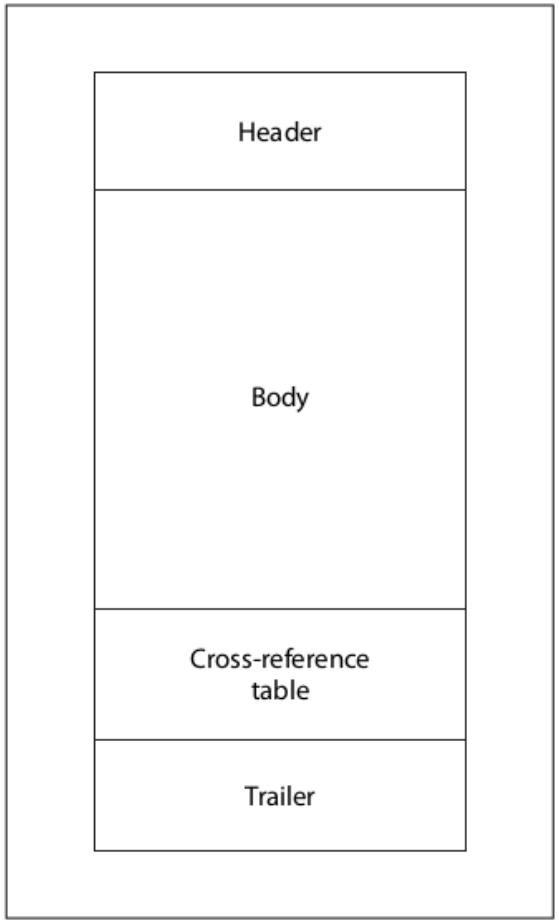

FIGURE 3.2 *Initial structure of a PDF file*

PDF Basic File Structure

- Header - PDF version.
- Body - Objects (pictures, text, cool things).
- Cross Reference Table - Table of contents, where objects are located inside the document. This is based on file offsets. *Important!*
- Trailer - Data about the Cross Reference Table (where its located).

PDFs and Data Hiding

Let's revisit the PDF specification:

2.2.7 Incremental Update

Applications may allow users to modify PDF documents. Users should not have to wait for the entire file—which can contain hundreds of pages or more—to be rewritten each time modifications to the document are saved. PDF allows modifications to be appended to a file, leaving the original data intact. The addendum appended when a file is incrementally updated contains only those objects that were actually added or modified, and includes an update to the cross-reference table. Incremental update allows an application to save modifications to a PDF document in an amount of time proportional to the size of the modification rather than the size of the file.

In addition, because the original contents of the document are still present in the file, it is possible to undo saved changes by deleting one or more addenda. The ability to recover the exact contents of an original document is critical when digital signatures have been applied and subsequently need to be verified.

PDFs and Data Hiding

- The PDF Cross Reference Table is based on the offsets of objects in the file.
- Modifying, adding, or removing objects would change the size of the object. Therefore, all object offsets after that modified/added/removed object would change. This means the Cross Reference Table would also have to be updated.
- Solution: Don't change/delete/insert data!
- PDF writer tools just copy and modify the object and append the changes or additions to the end of the PDF. And an updated cross reference table is finially added to the end of the resulting PDF.
- This "appending" of modified/added data means that older data is still in the document.
- If an object is removed, it's still there, just the table of contents is told that it should not be displayed.

PDFs and Data Hiding: Safety

How to get around this history smuggling?

- Saving the document as a new PDF *Save as New*. There is user-trust in that the PDF creation tool will only generate a PDF without previous history information.

Act III: PDFResurrect

Act III: PDFResurrect

PDFResurrect

- Written in 2008.
- Linux but should be easy to port to Windows.
- Low dependencies, internally built PDF library.
- What: Tool that takes a PDF file, finds the older table of contents, and places that at the end of the file.
- Result: PDF readers read the older table of contents, and render the older versions of the document.
- Should be in Debian, Fedora, and AUR repositories, others?
- Source: `http://757labs.org/wiki/Projects/pdfresurrect`

PDFs in the Wild **rawr**

- Scrape the web, Google, Yahoo, for PDFs
- Example in Google: "filetype:pdf domain:.gov"
- Scrape the results, download the PDFs.
- Run a 'pdfresurrect -q' on each PDF.

What About Other Document Formats?

- Open Document Format(ODF):
  - Can maintain previous history, XML schema, showing who and when someone changed something. Nice for maintaining differences between versions. History tracking can be disabled.
- Microsoft Word ('97-2007 Binary Format .doc)
  - Can retain previous versions, which supports annotations, merging, etc. User tracking for these changes can be disabled. But history can be maintained, which seems similar to PDF. The latter depends on the saving mode being used (i.e. fast or full-save).

The End

Questions?

I was born not knowing and have had only a little time to change that here and there.

–Richard Feynman

Resources: Part 1 of 2

- Adobe PostScript 3 Language Manual:
  `www.adobe.com/products/postscript/pdfs/PLRM.pdf`
- Adobe PostScript Cookbook:
  `http://www-cdf.fnal.gov/offline/PostScript/BLUEBOOK.PDF`
- Adobe Portable Document Format:
  `http://en.wikipedia.org/wiki/Portable_Document_Format`
- Adobe PDF Specifications:
  `http://www.adobe.com/devnet/pdf/pdf_reference_archive.html`
- PDFResurrect source and paper:
  `http://757labs.org/wiki/Projects/pdfresurrect`
- Hakin9 PDFResurrect Article: "Faith in the Format: Unintentional Data Hiding in PDFs", March 2010.
- Ghostscript: `http://pages.cs.wisc.edu/~ghost/`
- PostScript: `http://en.wikipedia.org/wiki/PostScript`

Resources: Part 2 of 2

- Tree Goats:
  `http://blog.moment.ee/2007/01/viljad-on-valminud-tree-goats.html`
- Sheep:
  `http://en.wikipedia.org/wiki/File:Barbados_Blackbelly.JPG`
- Cholula Sauce: `http://cholula.com`
- Wizard of Oz:
  `http://pharmamkting.blogspot.com.au/2010/10/pharma-social-media-behind-curtain.html`