IDR SOLUTIONS
PDF & IMAGE SPECALISTS

1.4 EDITION

10 COMMON ISSUES

WITH PDF FILES

1. Poor Text Extraction

The Problem

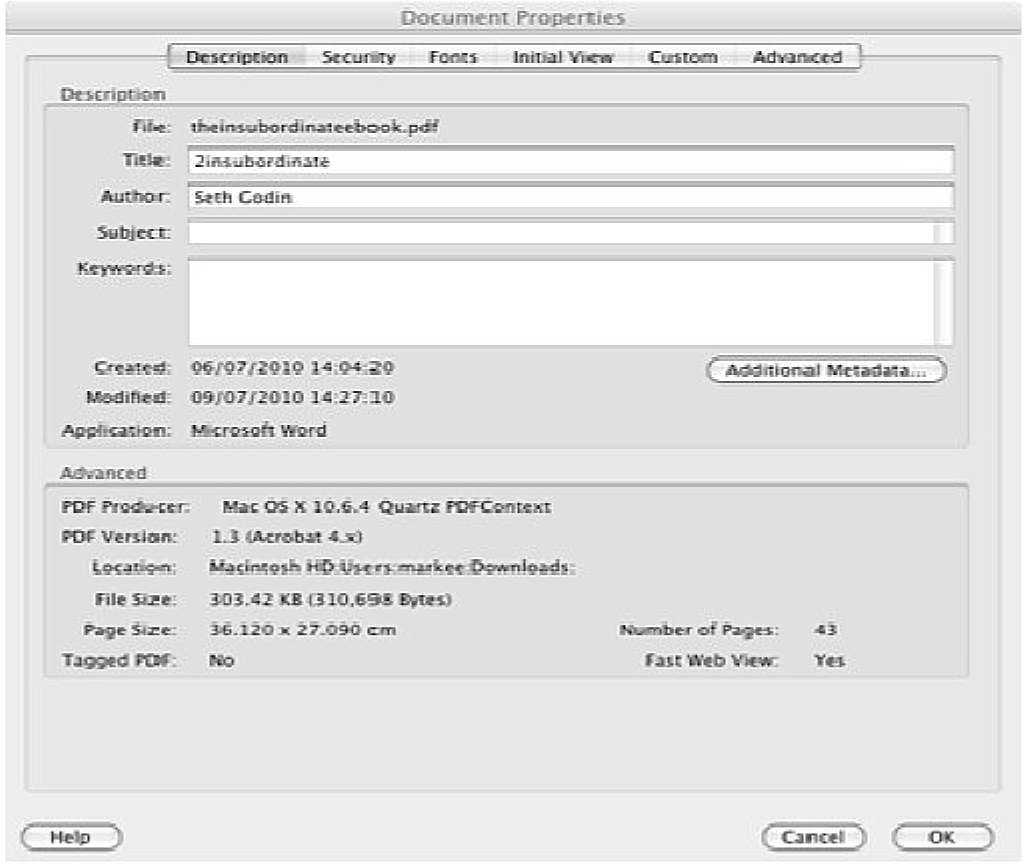
The text cannot be extracted very well. By default PDF files do not contain any text structure so all paragraphs and formatting of information is lost.

The Answer

Create the PDF file as 'Marked Content' which includes all this structure.

It turned out that people wanted to extract text from PDFs (and not just view them). So Adobe added a feature called 'marked content'. This allows the PDF file to contain additional tags as information, preserving the structure of the text.

However, this features needs to be used in the creation of the PDF, otherwise the additional information is not there!



Description

File: theinsubordinateebook.pdf
Title: 2insubordinate
Author: Seth Godin
Subject:
Keywords:
Created: 06/07/2010 14:04:20
Modified: 09/07/2010 14:27:10
Application: Microsoft Word

Advanced

PDF Producer: Mac OS X 10.6.4 Quartz PDFContext
PDF Version: 1.3 (Acrobat 4.x)
Location: Macintosh HD:Users:markee:Downloads:
File Size: 303.42 KB (310,698 Bytes)
Page Size: 36.120 x 27.090 cm   Number of Pages: 43
Tagged PDF: No   Fast Web View: Yes

1. Poor Text Extraction

The Answer (Continued)

There is a very easy way to tell if the PDF file has been created this way.
Open the file in Acrobat Reader and look at the properties menu. The Tagged PDF menu option (bottom left entry on the advanced section) tells you if the PDF contains these extra tags.

This file does not. So this PDF file will contain only limited structure tags.
If you can created Tagged PDF, it is worth setting this on by default, he files are not much larger and it makes text extraction much more viable if you need it in the future.

2. Colours are not cross platform

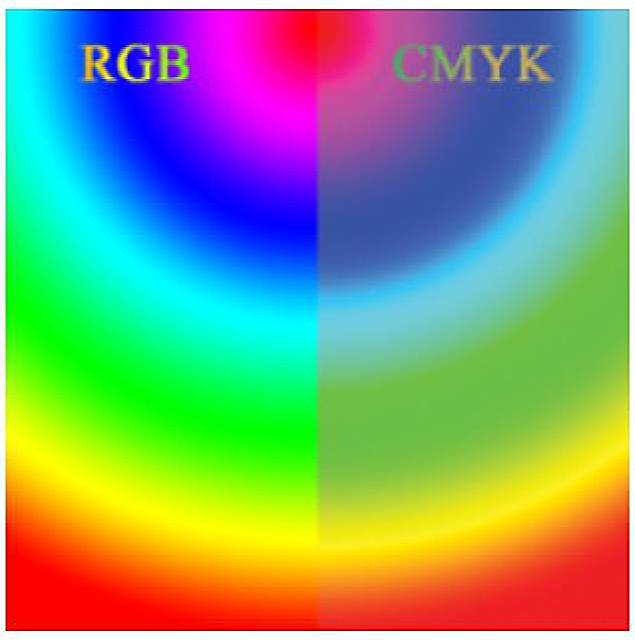
The Problem

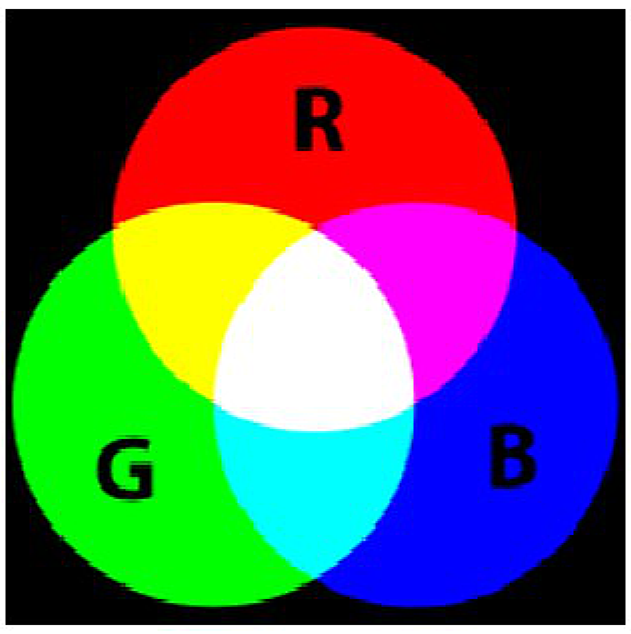
The colors in my PDF files appear different on other machines.

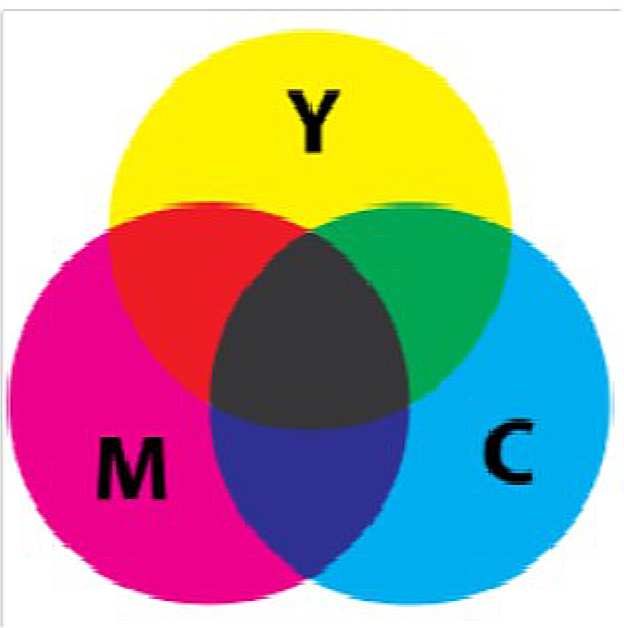
The PDF file format allows the use of many different types of color and not all of these are calibrated or guaranteed to look identical.

The Answer

Try sticking to specific color-spaces (try CMYK).







3. Font is not uniform cross platform

The Problem

The text is not appearing with the font that I use on other machines.

You can create PDF files with any font on your machine. If you do not include the font (its called embedding), the user will only be able to use it if they also have the font installed.

The Answer

Embed the font.

You need to be careful with fonts in PDF files. I was sent a PDF file which did not display properly in our software so I opened it in Acrobat on my Mac, and guess what... It did not display properly in that either! It did however work correctly in Acrobat on Windows, the problem turned out to be in the fonts.

3. Font is not uniform cross platform

The Answer (Continued)

Many PDF creation tools let you add fonts into a PDF file if they are on your system. But they do not include critical font data (like the widths of the characters so the fonts can be approximated or the actual font data). So when you try to open the file on a different machine (without this font), the font doesn't look correct.

In theory, the PDF file format provides a set of standard font families you can use, however one of the fonts that was not correctly displayed on my 'problem PDF' was Symbol, the Mac version Acrobat seemed confused because it was WIN encoded.

The best solutions is to embed the font which includes off the information needed to draw it an makes no assumption about what is on the viewing machines. If you subset the font, only the minimum data to draw the required glyphs is included, making it compact.

So be careful with your fonts, and if you are not embedding them, make sure you test the PDF files on any viewing platforms.

4. Poor PDF Image Printing

The Problem

The PDF images do not print very well.

PDF files can be optimized for screen display or printing. Screen displays can use lower resolutions images which means smaller PDF file sizes, this results in them not printing very well.

The Answer

Choose the appropriate image quality for your intended usage.

Most elements of a PDF file are device independent Vector Graphics, this means they will always appear 'smooth and clear' no matter the resolution. By contrast images are bit-mapped.

If the image is stretched it can appear pix-elated. If the image is too small, it will be stretched and the pixels will become blocky, this happens when the image is of a lower resolution.

4. Poor PDF Image Printing

The Answer (Continued)

There is a trade-off here, the better the quality, the bigger it (and the PDF) will be, that is why most PDF creation tools allow you to specify whether you are producing files for screen or printing.

With a PDF file you have a raw image and a 'scaling' command to fit it into a slot on the page (CTM matrix).

If the raw image is bigger than the 'slot' you can zoom into the page and print it at high resolution.

There are drawbacks to this, we once saw a large PDF file with raw images of 26,000 x 26,000px. Combining the raw image size and the CTM does actually allow you to work out the image quality to decide how well the PDF file will scale and fit.

4. Poor PDF Image Printing

The Answer (Continued)

Here's an example :

Imagine we have an image which is 585 x 585px. It is shown on the page with this CTM value :

(843 0 0)
(0 843 0)
(x y 1)

Using this information we can work out the dpi value with this formula :

dpi = (int)(imageWidth/GraphicsStateValue*100);

This gives us a value of 69. This is reasonable for screen display (where 75 is the target figure) but not great for printing or for fine zooming. For printing, 300+ dpi should be the target.

5. The PDF File is very large & slow

The Problem

It is possible to embed huge images (when much smaller images would look just as good) and embed the whole of a font when only a few characters are used.

The Answer

Choose the correct image size for extended purposes ad subset any fonts (this includes data on only the characters actually used).

6: Cant search or select PDF Text

The Problem

I can see text in my PDF file but I cannot search or select it.

The PDF file is very flexible and I is perfectly possible to include text in the form of shape o images. In this case it is not really text.

The Answer

Make sure the PDF files contain text and have an OCR tool, if you have to extract text from such files (Acrobat X has a pretty good one built in).

7. Transparency wont work in Viewer

The Problem

PDF Files have some very complex transparency settings which do not always display very well in most PDF viewers.

Adobe built some very sophisticated and high end features into the PDF file format which are only 100% supported in Acrobat.

You can often simplify them by re-saving the PDF file through using a lower PDF version (I.e 1.4).

The Answer

Simply your PDF files and avoid very complex features unless you absolutely need to use them.

8. PDF loads slow over the internet

The Problem

The PDF file was originally designed with vital information at the end of the file, so the whole file would have to be available before it can be displayed.

When downloading files from the Internet, this means the whole files has to be downloaded before it can be accessed and the content shown.

The Answer

Create 'Linearized PDF files', where some critical information is stored at the start of the file and it can be displayed very quickly.

When you access a PDF file across the Internet, it can take some time to open the file, this is down to the way a PDF file is designed. It consists of PDF objects and a table linking these objects to each page.

This make I very fast on a file system as the PDF viewer just reads the table (at the end of the PDF file) and loads just the required objects for any page using Random Access.

8. PDF loads slow over the internet

The Answer (Continued)

A file system allows you to access any bytes in a file without having to start at the beginning. However with a URL stream you cannot do this, you have to read them in order from the start. An internet connection does not allow for Random Access, and to be able to read from the end of the file you must download it, you cannot just skip to the end of the stream.

However you can create PDF files so they store the table and the objects for the first page at the start of the file, this means that the PDF can be displayed MUCH faster.

This is known as 'Linearized PDF'. It allows you to view the PDF before it is fully downloaded and access the pages as soon as they available.

If your PDFs are 'Linearized' then you can access them faster, otherwise you will have to download the whole file because the important info is at the end.

9. PDF File doen't save text properly

The Problem

The PDF file contains selectable text but it is garbage if I save it.

Some PDF creation tools do not properly embed the data needed to extract the PDF text as text.

The Answer

Use a decent PDF creation tool and check the text is correctly embedded.

10. My Editied PDF file doesn't work

The Problem

My PDF file (which I sent in an email or accidentally edited in a text editor) does not work.

The PDF file is a binary file format where changing a single byte can make the whole file unusable and corrupted.

The Answer

Suspect the PDF may be corrupted and get another copy.

What do I do Next?

There are several great sites worth checking out.

We do a blog on general Java/PDF issues at blog.idrsolutions.com.

You can also follow us on twitter at twitter.com/javapdf and other social media pages (see below)

Learn more about us at www.idrsolutions.com

Did you enjoy this Free Guide?

Download our other free Guides