IDR SOLUTIONS
PDF & IMAGE SPECALISTS

1.5 EDITION

A QUICK GUIDE

PDF FOR JAVA DEVELOPERS

Introduction

PDF Files are a major file format which many Java developers
will encounter, it is not a simple file format and it is not easy to handle in Java.

We have spent over 13 years manipulating PDF files with Java so we have put together this simple guide, hopefully it will fill you in quickly ad give you the essential information you need to know.

We hope you find this guide useful.

If you have any suggestions on how we can improve it or want to know more about PDF issues in general, please feel free to contact us via :

www.idrsolutions.com/contact

Introduction

PDF files are documents created according to a version of the Adobe PDF reference guide.

They are rather complex binary files which need a third party piece of software (Acrobat Reader is just one examples) but they do look good and they are the same across computers systems, therefore they are a very popular format for displaying documents and they can also have form elements, interaction and even embedded video and sound.

As a Java developer you will probably meet them in one of the following ways :

- I need to create/edit a PDF file.
- I need to view or print a PDF file.
- I need to extract data from a PDF file.
- Convert PDF files to something else (generally images).
- Search a PDF for text.
- I need to manipulate a PDF file (for example add or remove a password).

We will give you a introduction to all of these in this guide

What are PDF Files?

How much support does Java have for PDF?

Very little I am afraid, and the PDF file format is also pretty complex it is not something you want to be trying to hack directly,luckily there are lots of Java libraries available (both free and commercial).

As the developers of Jpedal, one of the most popular viewer/print/extraction libraries, we want to share our knowledge about PDF with you. This guide is not about Jpedal but a general PDF guide which will give you a working knowledge about PDF files.

N/B : There are several version of the PDF file format starting with1.1 and going up to 1.7 (1.7 related to Acrobat 9.1, however beyondthis the older addition of numbers to get the version is no longer valid.(Check out the PDF Reference Manual for more information). You need to make sure you use a library which supports the version being used in your files.

The PDF specification is now a ISO 32000 standard however there is development of a new standard known as PDF 2.0 which will be released at a later date

Working with PDF in Java

I Need to Create / Edit a PDF File

Most people find PDF files confusing because they expect them to be like HTML or Word files, they are not, a PDF file is a tree of objects which can be read very quickly so that any page can be displayed.

A reference table is loaded into memory and then this is used to load the objects from the file as required, any page can be displayed very quickly without having to load the rest.

The reference table needs to be correct for the PDF file to open so we strongly recommend that you do not try to create or edit a PDF directly.

Working with PDF in Java

You can Create a PDF in 2 Ways

1. Printing' a file to PDF (using Distiller, Ghostscript, and lots of other tools), this does all the hard work for you and if you want to convert XML to PDF, you can use the free tool FOP from Apache – otherwise a third party product will probably be required as Java Printing alone does not really support PDF.

2. Programmatically create a PDF from your code. You can use iText to design the PDF and then iText will create the PDF for you.
Editing the PDF should again be done with a tool because much of the data, it may be compressed and/or encrypted inside binary objects and you do not want to break the reference table since text in a PDF file loses much of its structure and flow so do not hope to carry out complex editing of text with reflow in a PDF file.

Therefore you are better off editing the file in its original format and then recreating the PDF.

Working with PDF in Java

I Need to View or Print a PDF File

Depending on the library you use you can get excellent display of PDF files in Java, most of these provide you with a standard Jpanel component or a Bean so that you can integrate it into your own Swing code.

However, some high end features (such as very complex transparency or embedded flash) will not work because Java cannot support them.
Java Printing will send the PDF to a printer, so if the printer directly supports PDF files it will be able to print them, however most do not so you will need to render them into a printable image on the computer.

The mechanics of this are virtually identical to viewing them; so many viewers also offer this feature.

Working with PDF in Java

I Need to Extract Data From a PDF File

As well as the content of your documents, PDF files can contain metadata as a set of values or an XML file, so you can store all sorts of things inside the file.

The process of creating a PDF loses much of the structural information in the text and converts the images into binary objects; you can extract text and images from a PDF but be careful in your expectations.

For example, you may wish to extract text from a document but the text may actually be an image instead of a sequences of characters.

Working with PDF in Java

Convert PDF to Something Else (e.g Images)

It is relatively straightforward to convert the PDF files to an image (again it is technically very similar to viewing or printing) and lots of libraries offer this ability. If you want to convert the PDF into other formats (Word, Excel and Powerpoint), it is easy to do through using a variety of different libraries some of which are available to try for free online.

Converting PDF to HTML used to be a difficult process as the PDF file format contains lots of very specific features which without special preparation do not translate easily however the process is much more easier now due to the industries move towards the HTML5 standard and the markets move towards smartphone, tablets and other devices. We offer different solutions to converting PDF to HTML5 including a Java Software Solution and a Online Converter

Search a PDF for Text

While it is hard to extract structured text content from a PDF file, the PDF contains the exact location of every text character on the page so it is possible to extract all the words and their location on the page.

Working with PDF in Java

Manipulate a PDF (Add/Remove Password)

There are several tasks you may need to do to existing PDF files such as adding or removing password protection.
Libraries like iText offer these functions in Java but be warned you may need some information.

10 Tips for PDF Files

1. Do Not Think of a PDF Files as a 'File'

When you start to learn HTML, you can open a file through hacking it in a text editor and see what happens…you can't do this with a PDF file.

This is it is essentially a binary data structure – lots of the information cannot be seen if you open the raw file and editing one byte could potentially break the whole file.

There are lots of really good tools out there on multiple platforms for examining the contents of a PDF file so you should not need to try and open the file directly.

2. PDF is All About Objects

What the PDF file essentially contains is a whole lot of PDF objects.
They all have a unique ID of the format number generation R (so you might see 3 0 R, 144 0 R), most of the time the generation is zero but not always.

There are lots types of objects – a 'Page' object describes a particular page, a 'Font' object contains all the information about a specific font, a 'Form' object contains information.

2. PDF is All About Objects - Continued

Objects can reference other objects, so Page object 5 0 R might reference Resources object 10 0 R which contains a list of Font objects used for the page, including Font objects 16 0R, 17 0R, 18 0R.

The objects can also be thought of as a Tree, this is what allows any page to be opened quickly, the PDF root object points to the list of pages which point to the resources they use and their contents.

3. Identical PDFS Can Be Deceiving

The PDF specification is very broad and flexible so there are lots of different ways to achieve the same result however the specification does not enforce any approach so all the PDF creation tools do things in different ways.

If you have a strange PDF, it is always worth seeing what the Producer or Creator settings are.

4. Images are Ripped up Inside a PDF

When a PDF is created, images are broken up into their pixel and colour data so that they can be compressed as efficiently as possible, with JPEG data it may well be stored in a JPEG compression format (DCTDecode or JPXDecode) but it may still need to have colour information applied.

5. The PDF Reference Guide is a Must Have

Adobe produces a detailed specification of the PDF Reference guide which is free to download. It is very big and there is an awful lot to it. Ideally, a beginner should start with the outline of the file format and just the areas they need to understand.

The PDF specification goes into considerable detail on the specification, but it may not be written from the precise viewpoint you need and also Adobe allows considerable interpretation in of what is acceptable, while there are lots of examples, it is possible for tools to do things in other ways.

6. What Makes a PDF File?

A PDF file should ideally have a .pdf file type, an xref pointer in the last 1024 bytes of its data and the file line of a PDF should be the version.
But there is quite a lot of variation in what is actually allowed in a PDF and how useful a PDF is. A PDF file can contain fonts and editable text or just be a raw around an image.

At the end of the day, if it opens in Acrobat it is accepted as a PDF and you need to handle it...

7. PDF is a Collection of Other Technologies

There are lots of other technologies used inside the PDF file format Including compression algorithms, encryption, font technologies, javascript and so on.
This makes it harder to understand because you need to have a grasp of these as well to understand what is going on.

8. Use the Tools

There are lots of tools (both free and commercial) on all platforms and in different languages (C, Java, Perl, Php, etc).

They make it much easier to work with PDF files and also experimenting with them (especially if you can access the source code) is a good way to understand how PDF works.

9. Become a PDF Expert is Not Easy!

I started working with PDF files over 10 years ago and I still learn new things every day.

PDF is a big, complex file format including a lot of technologies so it will need time to become proficient with it.

10. There are People to Ask!

I remember meeting Tom Phelps, the developer or Multivalent, at a conference in 2002.

We were so pleased to find someone else we could actually have a conversation with, we spent the whole night discussing PDF issues at the pub afterwards.

Everyone else in the bar complained it was the most boring night of their lives, but we both had a good time... Thanks to the Internet, you can discuss PDF issues without totally destroying your street credibility!

Many of the people or companies producing PDF tools run mailing lists or discussions forums (my first job every morning is to check the JPedal Support forums) and there are more general forums, I personally find stackoverfow a good place to ask questions.

Bonus Tip: I Need to Extract Data From a PDF File

As well as the content of your documents, PDF files can contain metadata as a set of values or an XML file, so you can store all sorts of things inside the file.

The process of creating a PDF loses much of the structural
information in the text and converts the images into binary objects; you can extract text and images from a PDF but be careful in your expectations. For example, you may wish to extract text from a document but the text may actually be an image instead of a sequences of characters.

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)

Also keep an eye on www.planetpdf.com which is the number one PDF site and is now back under dynamic new management.

There are lots of free PDF libraries and the commercial ones have trial versions, so get out there!!... and see what is available.

Let us know how you get on and if you have any suggestions for this guide.

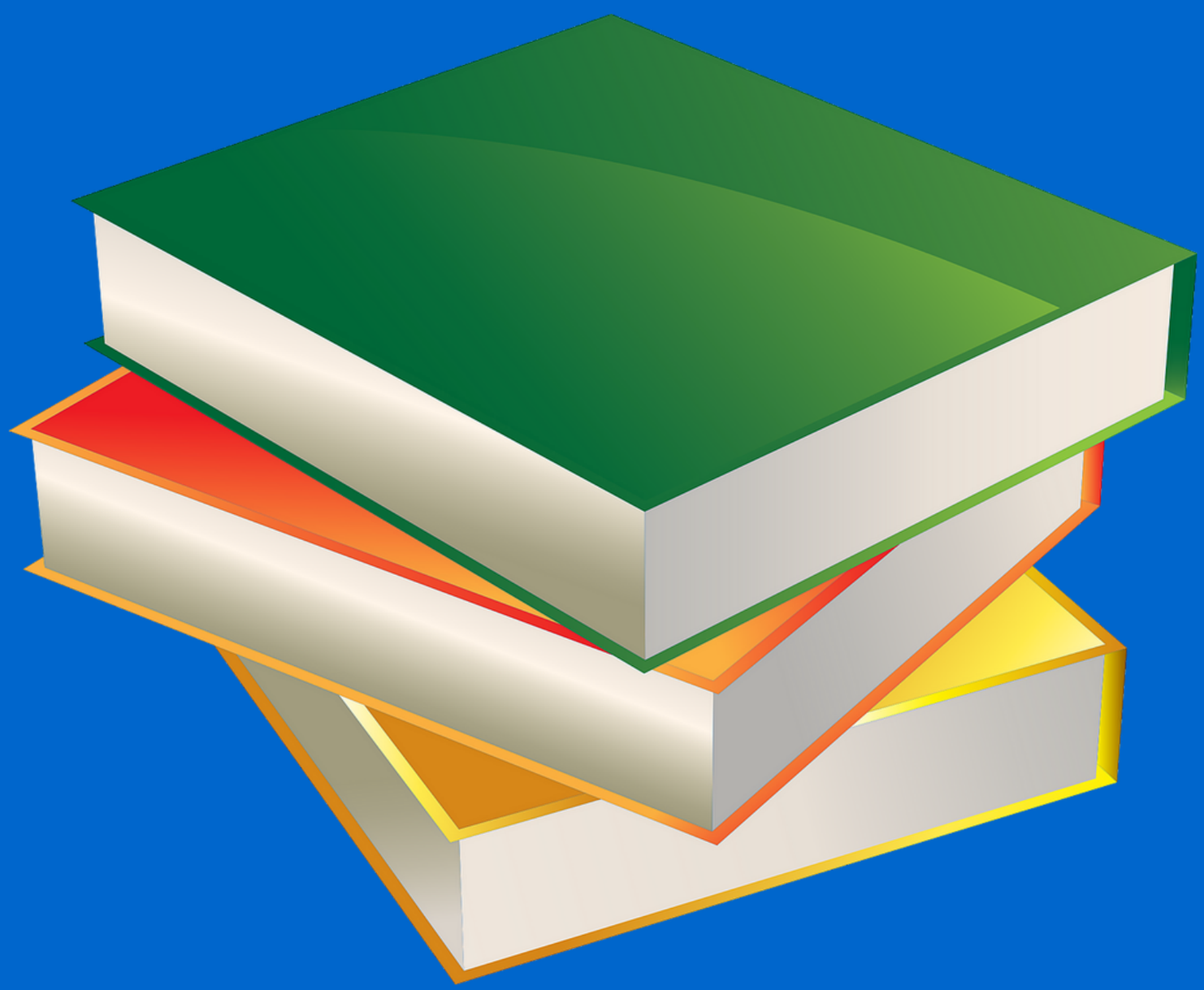
Acknowledgement

Thanks to the developers and staff of IDR Solutions and the JPedal PDF library in making this document:

Mark Stephens, Brendan Hillary,
Julie Stewart, Chris Wade, Simon Barnett,
Rachel Reed, Kieran France, Mariusz Saturnus,
Arran Titchmarsh, Sam Howard, Daniel Meredith,
Chika Okerche, Leon Atherton, Lyndon Armitage,
Sudharman Varatharajah, Alex Marshall
George Perry, Simon Lissack, Nathan Howard,
Sophia Matarazzo ,Sylwia Kedzia, Ernest Duodu,
Georgia Ingham, Bethan Palmer, Zain Arshad
and Rob Foley.

Did you enjoy this Free Guide?

Download our other free Guides



Download Free Ebooks