Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 2. FORMATS FOR ELECTRONIC DOCUMENTS AND IMAGES

LESSON 4. PORTABLE DOCUMENT FORMAT (PDF)

NOTE

Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

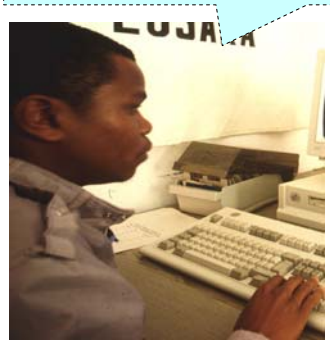
Objectives

At the end of this lesson, you will be able to:

- understand **when to use PDF** format;
- understand the main **features** of PDF; and
- understand the difference between PDF and **embedded TIFF**.

When to use PDF?

I am creating my document using MS Word, but I have some doubts about which format to use for its dissemination…



The best format for document delivery depends on what you're doing with the document.

To create and edit a document, **Microsoft Word** can be an appropriate format.

However, this is not necessarily the best format for dissemination of documents.

When to use PDF?

People have to be able to read and print the document. Therefore, the document must display correctly, regardless of the software and hardware being used.

Portable Document Format (**PDF**) is a procedural mark-up language that allows page-formatted documents to be **viewed and printed** in their **original page layout** on almost **any software platform**.

PDF is the most common format for exchanging documents where the **page format** of the original document must be **preserved**.

What is PDF?

The PDF mark-up language is based on an image model, whereby a document contains a set of pages, which are described by **three main object types:**

Path objects contain a **description** of a set of points and the way they are connected by lines or curves, equivalent to a vector graphic format (e.g. for displaying computer generated graphics).

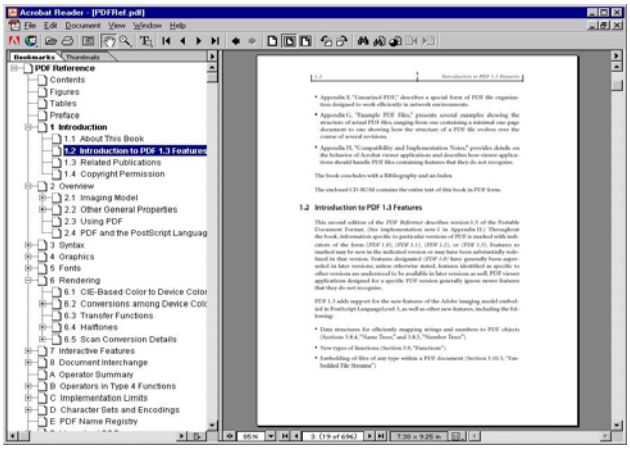
Image objects are a rectangular array of **image points**, equivalent to a raster graphics format (e.g. for displaying photographs).

TEXT

Text objects contain a set of glyphs (images representing text characters), which shapes are described by a separate font.

Now, let's look at PDF characteristics in detail...

Features of PDF



PDF was created by **Adobe Systems Inc.** in the early 1990s.

PDF is based on Adobe's PostScript page description language, but has more sophisticated features. The underlying structure of the document is exposed and accessible to software processors. This makes it easier to support features such as **document navigation** and **hyperlinking**.

The figure shows a PDF document with a navigation tree of bookmarks.

Features of PDF

This document will be translated in Arabic, Chinese, Japanese and Korean. But how will these characters be displayed?

Because the PDF file contains all the **font information** required to render the document, it is an ideal format for both scientific documents that contain **unusual symbols,** and for **multilingual documents**, particularly those using double byte character sets. It is also useful for mixing different languages in the same document.

Although **HTML or XML** documents using Unicode character encoding support the full character set required by scientific or multilingual documents, the **display** of those characters will **depend on the software system** used to render them, which cannot be guaranteed to do that correctly.

Features of PDF

A PDF file contains all the information necessary to render a document on screen or on the printed page, so that it looks **the same on any computing platform**, regardless of the software packages used to create and render it.

So a PDF file created on an **Apple Macintosh** will look the same to a user of a PDF viewer on a **UNIX workstation** or a reader of the document printed from a **Windows PC**.

This feature is called **portability**.

Features of PDF

The following are other important features of PDF:

COMPRESSION
PDF supports several compression algorithms. This reduces PDF files to minimal size, particularly for transmission over the Internet.

FONT MANAGEMENT
PDF font management allows font information from the original system that generated the document to be embedded in the file.
PDF also has direct support for 14 commonly used fonts. These can be used without embedding additional information.

SECURITY
There are two main security features in PDF. Documents can be encrypted, with different access permissions assigned to the creator and users of the document, so that they can update, view or print it. Documents can also be digitally signed so that their authenticity can be verified.

RANDOM ACCESS
The random access of objects in a PDF files means that viewing software can support features such as bookmarks, thumb-nails, tables of contents, annotation and hyperlinking (both within a document and to external resources).

INCREMENTAL UPDATE
PDF files can be incrementally updated, so that changes are added to the end of the document, rather than being applied to objects scattered throughout it. This means that large documents can be edited and saved very quickly, and also ensures that the original content of the document is preserved, no matter how many updates are made.

EXTENSIBILITY
The PDF specification has been designed to be extensible, so that new features can be added, whilst previous versions remain valid and older software still renders newer versions without error. In addition, applications can store their own information inside a PDF file, so that it can be processed in a specific way by that application, whilst still being printable and viewable by other applications (which ignore information specific to any other application).

Features of PDF

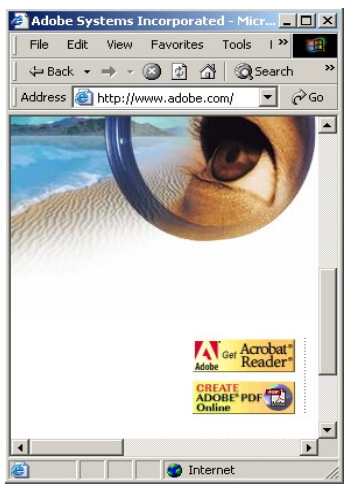
The document can be disseminated through the Internet. People can connect to the website and download it.

The compression and incremental loading features of PDF make it well suited for **transmission** of documents over the **Internet**.

Because Adobe have published the PDF specification and made the Acrobat Viewer freely available, you can be confident as an information creator and distributor that the **users** of your PDF documents **will be able to obtain a Viewer easily** and without having to pay for it.

It seems that this works, since it is estimated that over 300 million copies of the Viewer have been downloaded!

Features of PDF



In one sense PDF is an open format: the PDF specification was published by Adobe in 1993, at the same time as Adobe launched the Acrobat suite of products for creating and viewing PDF documents. There are now commercial products and open source software tools available from many sources other than Adobe.

However, the PDF specification is still **owned by Adobe Systems Inc**, who reserve the sole right to maintain and extend it, so it is not an open standard in the same way that HTML or XML are.

If you are interested in further features of PDF, look at the Adobe website:

www.adobe.com

PDF Software Applications

What do you need to **create a PDF document**?

PDF files can be created directly by software applications which can direct generate PDF, e.g. **Adobe PageMaker** or **Corel Ventura**.

In other applications you can save to another format (such as PostScript), then use a **PDF translator** to create PDF.
The most widely used PDF translator is Adobe Acrobat Distiller.

Any application that is able to print documents can also create PDF indirectly by installing a **PDF print driver**.
Adobe's own PDF print driver is called PDF Writer, but there are print drivers available from many other commercial and open sources.

PDF Software Applications

Adobe Systems, Inc. supply the Acrobat product suite, which includes Acrobat Distiller, Acrobat Viewer and Acrobat Capture (which can scan and convert paper documents to PDF).

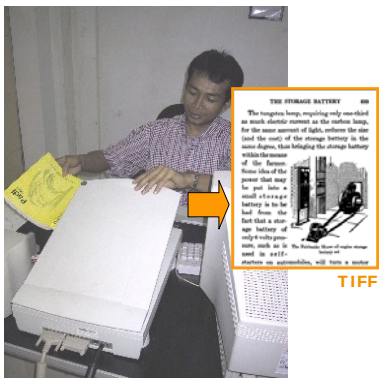
There are **many other products** available from other vendors **for creating, viewing, editing and printing PDF** documents.

Since PDF is also a mark-up language with a published specification, there are also many **smaller toolkits and utilities** available for processing PDF files or performing specific functions such as extracting text or breaking into individual pages. Many of these are available as open source software. See:

www.pdfzone.com

www.pdfstore.com

Embedded TIFF and true PDF



Sometimes, the entire pages of a PDF document are **image objects**.

This is the case when a document printed on paper is scanned and each page is saved as a **TIFF image** (a popular raster graphics format) and **then** converted to **PDF**.

This type of PDF is called **embedded TIFF**.

Embedded TIFF and true PDF

Embedded TIFF is **not a "true" PDF**: only when the page images are converted to a set of PDF text objects do we have a true PDF.

Embedded TIFF doesn't allow you to index, search, link, copy and edit the text. In fact many of the features and capabilities of the PDF format **will not work** when complete pages are represented as images.

However, there are other reasons to use TIFF images. One would be to preserve an **exact replica** of the printed page, including handwritten marks or annotations that could only be captured properly in a page image.
Another reason to use TIFF images would be to prevent users **copying or altering the content** of the document. However, the latest versions of PDF contain encryption and digital signature features that can achieve this, even with text objects.

Summary

• **PDF** (Portable Document Format) is a procedural mark-up language that allows page-formatted documents to be **viewed and printed** in their **original format** on almost **any software platform**.

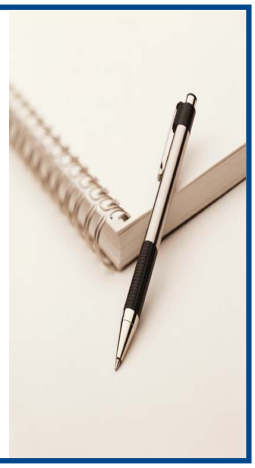
• PDF is an ideal format for scientific documents that contain **unusual symbols, and for multilingual documents**.

• The compression and incremental loading features of PDF make it well suited for **transmission** of documents over the **Internet**.

• Many software packages can be used to create PDF documents, and PDF viewers are available free of charge.

• A PDF document contains a set of pages which are described by **three main object types: path** objects, **image** objects and **text** objects.

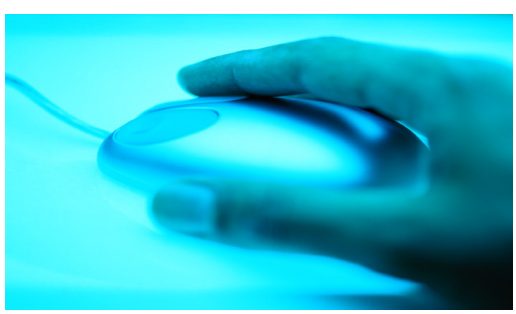
• **Embedded TIFFs** are PDF documents where the entire pages are TIFF images.



Exercises

The following three exercises will help you test your understanding of the concepts covered in this lesson, and provide you with feedback.

Good luck!



Exercise 1

In which of these situations would you select the PDF format?

- ○ For a document that still has to be modified by others.
- ○ For a document that has to be displayed and printed in its original format.
- ○ For a document that has to be displayed online.

Click on the answer of your choice

Exercise 2

What does it mean that PDF is a portable format?

- ○ A PDF file looks the same on any computing platform.
- ○ The size of a PDF file is appropriate for dissemination through the Internet.
- ○ PDF documents can be encrypted, with different access permissions.

Click on the answer of your choice

Exercise 3

In which of these cases would it be more appropriate to use embedded TIFF rather than true PDF?

- ○ To disseminate online a journal containing many photographs.
- ○ To disseminate online the original copy of a manuscript.
- ○ To disseminate online a document that should not be modified.

Click on the answer of your choice

If you want to know more...

Adobe Systems Inc (www.adobe.com) the home of PDF and the Acrobat product range. You can also download a copy of the PDF Reference.

PDF Reference published (2000) by Addison-Wesley (ISBN 0-201-61588-6)

Planet PDF (www.planetpdf.com) an independent global resource for Adobe Acrobat PDF products, tools and information.

PDF Store (www.pdfstore.com) an online store for PDF software products (linked to Planet PDF).

PDF Zone (www.pdfzone.com) online information source for PDF and related technologies.

Corel (www.corel.com) home of Ventura, a desktop publishing package with direct PDF creation capabilities.

Quark (www.quark.com) home of QuarkXpress a desktop publishing package that can create PDF by saving as Postcsript and using Acrobat Distiller