PDF Scoping Report: Appendices

These appendices are working notes generated during the writing of the PDF Scoping Document, mainly related to the technical detail of the PDF format and document creation. They are presented as is without full formatting or editing. While we hope that they may be of use, other well-presented introductions to the subject are available on the web, and may be of interest. These are listed in the references section of the main report.

www.aidworld.org 109 Barnwell Road, Cambridge, CB5 8RQ UK Registered No: 4980791 +44(0)1223 331716 info@aidworld.org

Appendix 1

Technical Structure of a PDF File

This section provides information about the PDF file format; the aspects of the format which affect the size of files generated and options for presenting data contained in PDF documents.

The Wikipedia entry for PDF states

> "Portable Document Format (PDF) is a file format developed by Adobe Systems for representing documents in a manner that is independent of the original application software, hardware, and operating system used to create those documents. A PDF file can describe documents containing any combination of text, graphics, and images in a device independent and resolution independent format. These documents can be one page or thousands of pages, very simple or extremely complex with a rich use of fonts, graphics, color, and images. PDF is an open standard, and anyone may write applications that can read or write PDFs royalty-free." .

PDF also defines a structured storage system to compress text, images and other elements and bundle them together into a single file. This is based on a subset of PostScript. (PostScript is a page description language and programming language used primarily in the electronic and desktop publishing areas).

Creating PDF Documents: Original Document Data Available.

The simplest case involves generating documents directly from their original electronic versions. There are a number of ways to do this:

1. Applications supporting direct PDF document generation;
2. "Virtual printers" which output a PDF document instead of a printed page;
3. Applications for converting finished documents from other formats into PDF.

These methods are outlined below.

Direct PDF Document Generation

Some applications can create PDF documents directly. When the document is completed, an option will be available to save a copy as a PDF (sometimes called "Export as PDF"). Examples of applications possessing this ability include:

- OpenOffice;
- Adobe Acrobat;
- Scribus;
- The next version of Microsoft Office (due some time in 2006).

Virtual Printers

PostScript was originally designed with printing in mind, and PDF is based on a subset of PostScript. It is therefore relatively easy to insert a software layer to derive a PDF document from the information sent to the printing subsystem. Examples of such "virtual printer" software include:

- PDFCreator (Microsoft Windows);
- cups-pdf (Linux/Unix).

www.aidworld.org 109 Barnwell Road, Cambridge, CB5 8RQ UK Registered No: 4980791 +44(0)1223 331716 info@aidworld.org

Conversion to PDF

An application is run with a file as input and produces a PDF document as output. A common example is conversion of PostScript files (produced by various applications) into PDF documents - one important use of which is the conversion of scientific papers written in the LaTeX language to PDF.

Adobe run an online service for converting files to PDF. Viable input formats include DTP, CAD, PostScript and a number of image and office formats.

It is also worth mentioning that there exist automatic tools for converting PDF documents into various output formats such as text, HTML and PostScript. There is also software that extracts images from PDF documents.

Access Restrictions and Cryptography

Access Restrictions

Intrinsic restrictions may be set on what a user can do with a PDF document (defined in table 3.20 of the PDF Reference 1.6). Limitations may be placed on:

- Printing;
- Modifying the contents of a document;
- Copying or extracting text or graphics.

These restrictions are indicated by fields in the PDF document and enforced by the application reading the document. There is nothing preventing an application from ignoring them, allowing automatic conversion tools to access text and graphics. Access restrictions should however be copied into any new PDF document they produce. Obviously if the document is converted to HTML or text, any such restrictions will be lost.

Cryptography

This covers two aspects - digital signatures and encrypting the document.

Digital signatures allow the viewer of a document to have confidence that the document was authored (or at least approved) by a trusted person. Any change to the document, such as automated processing to reduce its size, would invalidate such a signature.

Encrypted documents are not readable without a decryption key. Not all of the document is encrypted - only its content (images and text). Fonts and document structure information are left unencrypted. The decryption key is most commonly supplied by processing a password. Automatic conversion would therefore require a password (or the key in some other form), and the document would need to be re-encrypted afterwards.

It is worth noting that it is not possible to compress encrypted data, so the compression built into PDF software would have no effect on the bulk of it. From a file size perspective it is therefore best to avoid encryption.

Creating PDF Documents: Scanned Documents

Some PDF documents are created by scanning in a paper copy of the document. Text can be extracted from these documents using Optical Character Recognition (OCR) technology. It is possible to store the OCR-ed text as metadata along with the scanned image in order to facilitate indexing and text searching, although inaccuracies associated with OCR technology may restrict the usefulness of such searches. A further step of inserting indexing links in the document might also help.

Scanning a document effectively creates an image of the document, leading to a large file size. This might be improved by scanning different parts of the document at different resolutions and/or colour depths. Low resolution monochrome would be sufficient for text. Images could then be scanned at a higher resolution (and in colour if appropriate).

Font Handling

From the PDF Specification, version 1.6:

> "A font is represented in PDF as a dictionary specifying the type of font, its PostScript name, its encoding, and information that can be used to provide a substitute when the font program is not available. Optionally, the font program can be embedded as a stream object in the PDF file." .

Incorporating fonts in a document will significantly increase their file size - a full TrueType font can consume 300 KB. Including the entire font may however not be necessary - it is possible to incorporate only the characters that are actually used. Typically this could add 5 - 10 KB to the file. This is the default option in at least some of the available PDF document generation tools.

There exists a set of 14 standard PostScript fonts renderable by default by all PDF document readers. If these fonts are used then there is no need to include extra font data in the PDF file. It is also possible to specify an arbitrary font which when unavailable on a user's system will be substituted with a standard font instead. This will lead to variations in how the document is displayed, but these should not be significant provided the original font is fairly similar to the substituted font.

Using more than one font will obviously lead to larger file sizes.

If a document is intended mainly for printing, then it is possible to "render" the fonts. This means that each letter of each word will be stored as a description of a series of curves. This guarantees the document will print as intended, with no variation due to differing font implementations. However this will lead to a significantly larger document.

Image Handling

When creating a PDF document from an electronic version, it is normally possible to exercise some control over the degree to which images are compressed. Quite a high level of compression may often be used without major quality degradation. It is important to note that as PDF documents are normally compressed, it is best not to insert already compressed images as the recompression may lead to visible "artifacts" in the image.

Repeated images (e.g. a logo on each page) lead to larger file sizes.

There are two categories of images - raster graphics and vector graphics. In general, vector graphics will lead to smaller file sizes (and better images). If raster graphics are used, it is good to start with a bitmap at the desired resolution and let the PDF document creation software handle the compression. This "desired resolution" will depend on whether the document is primarily intended for viewing on screen or for printing.

Compression

PDF documents are usually compressed when generated. They are compressed using a lossless compression method - i.e. the compression can be completely reversed to recover the original document. However the images contained in the document may be compressed with a lossy compression method. This yields a superior compression ratio, but at the cost of some loss of image quality. Compressing a PDF document will therefore not lead to a significantly smaller file size, but using a higher compression ratio for the images within the document may do.

PDF Versions

There are a number of versions of the PDF standard. Access control, types of compression, multimedia and cryptography support are some of the features that have been added or changed over different versions. This has implications for software wishing to read PDF files. There appears no compelling reason for a typical user to need compatibility with PDF versions beyond 1.4.

PDF vs HTML

Although variants of HTML such as XHTML exist we will refer only to HTML.

A document published as HTML generally exhibits smaller file sizes than the equivalent document published as PDF. HTML is easier to search and generally easier to read on a computer screen. Any simple text-and-graphics document that is typeset in a single column can easily be provided as an ordinary HTML and CSS web page.

HTML is typically superior to PDF from the point of view of disabled access. Tags in HTML provide context, and documents are usually easier to navigate, as long documents normally have links to aid in finding relevant sections. Web browsers allow the user to adjust font sizes, making it easier for people with poor eyesight to read. Screenreaders do however exist for PDF documents, and PDF documents can also have tags and links. A more detailed discussion can be found at http://www.alistapart.com/articles/pdf_accessibility/.

Unfortunately there are certain PDF documents that cannot be well represented as HTML. Reasons for this include:

- Footnotes and endnotes have no HTML equivalent.
- Complex mathematical expressions are difficult to display in HTML documents.
- Inclusion of the original image may be mandatory for PDF files consisting of a scan of a hardcopy document (e.g. legal documents).
- PDF documents can impose access and usage restrictions as mentioned above. Disabling the copying of text means that the document can only be redistributed in its entirety. Similar neutering features are not available in HTML.

Another limitation of HTML compared to PDF is that PDFs are generally better for printing, and some are optimised for that purpose (though for reading on screen, an HTML equivalent could still be provided).

www.aidworld.org 109 Barnwell Road, Cambridge, CB5 8RQ UK Registered No: 4980791 +44(0)1223 331716 info@aidworld.org

Appendix 2

Existing PDF Creation and Conversion Tools

There are many resources available for creating, inspecting and modifying PDF documents. They range in quality and price from enterprise level solutions from Adobe to small open source projects aimed at specific tasks. It is worth noting that some of the best software available such as GNU GhostScript is open source and that there are many commercial PDF products that produce poor results. At the time of writing, Adobe Acrobat 6 and 7 (both Standard and Professional versions) are the industry standard tools for working with individual PDF documents.

There are two scenarios for optimising a PDF document: at the time of creation and after creation when the source documents may be unavailable.

PDF Document Creation Tools

PDF documents can be created directly by an editor, by printing to a virtual printer, or by converting a document from another format. Here is a review of the most common tools.

Adobe Acrobat

Adobe Acrobat is the industry standard tool for creating PDF documents. The standard way to create a PDF document is to print the document from whichever application created the original. Acrobat interprets the printing instructions and uses them to create a PDF file. The user can then manipulate the PDF document before saving it.

The document could be optimised for printing or for viewing on screen. If the document is intended for printing and the Acrobat print profile is used, then the file size will be larger. The images will have a higher resolution and the fonts may be rendered.

Metadata (such as tags) can be added to a PDF document to improve the quality of searches within and across documents, with an associated file size penalty. There are options for high levels of compression that may reduce the quality of the images. It is also possible to add links and alternative text to images to improve accessibility.

While being the standard for PDF document generation, Adobe Acrobat does not appear to be designed with batches or automation in mind. Were it necessary to convert a set of documents, each would have to be processed manually with Acrobat.

It is possible to extract text and individual images manually from PDF documents using Acrobat - however again there exist no batch capabilities for doing so.

Adobe Web PDF Creation Service

This can be found at http://createpdf.adobe.com/. Documents can be uploaded and converted to PDF. Legal input formats include common office formats, DTP formats, image formats and PostScript. The service is very efficient from a file size point of view, and allows for fine control of the options applied to the PDF document.

PDFCreator

This is a commonly used open source solution for Windows which installs as a printer. When a user wishes to create a PDF document, he or she prints from their application to the PDFCreator printer driver. They are then presented with a dialog box to set the PDF options. PDFCreator can unfortunately be quite inefficient with file size.

www.aidworld.org 109 Barnwell Road, Cambridge, CB5 8RQ UK Registered No: 4980791 +44(0)1223 331716 info@aidworld.org

OpenOffice

OpenOffice is open source software which performs the normal range of office software tasks. It is able to produce PDF documents directly (from an option under the File menu).

PDF Conversion Tools

Open Source Command Line Tools

There are a number of command line tools for manipulating PDF documents.

- pdftops - converts a PDF document into a PostScript document;
- pdftohtml - converts a PDF document into HTML;
- pdftotxt - converts a PDF document into a text document;
- pdfimages - extracts the images from a PDF document;
- pdftk - can perform quite a number of operations to reformat PDF documents.

Adobe Online PDF Conversion Tools

Adobe runs an online web service that can convert PDF documents from PDF into either text or HTML.

www.aidworld.org 109 Barnwell Road, Cambridge, CB5 8RQ UK Registered No: 4980791 +44(0)1223 331716 info@aidworld.org