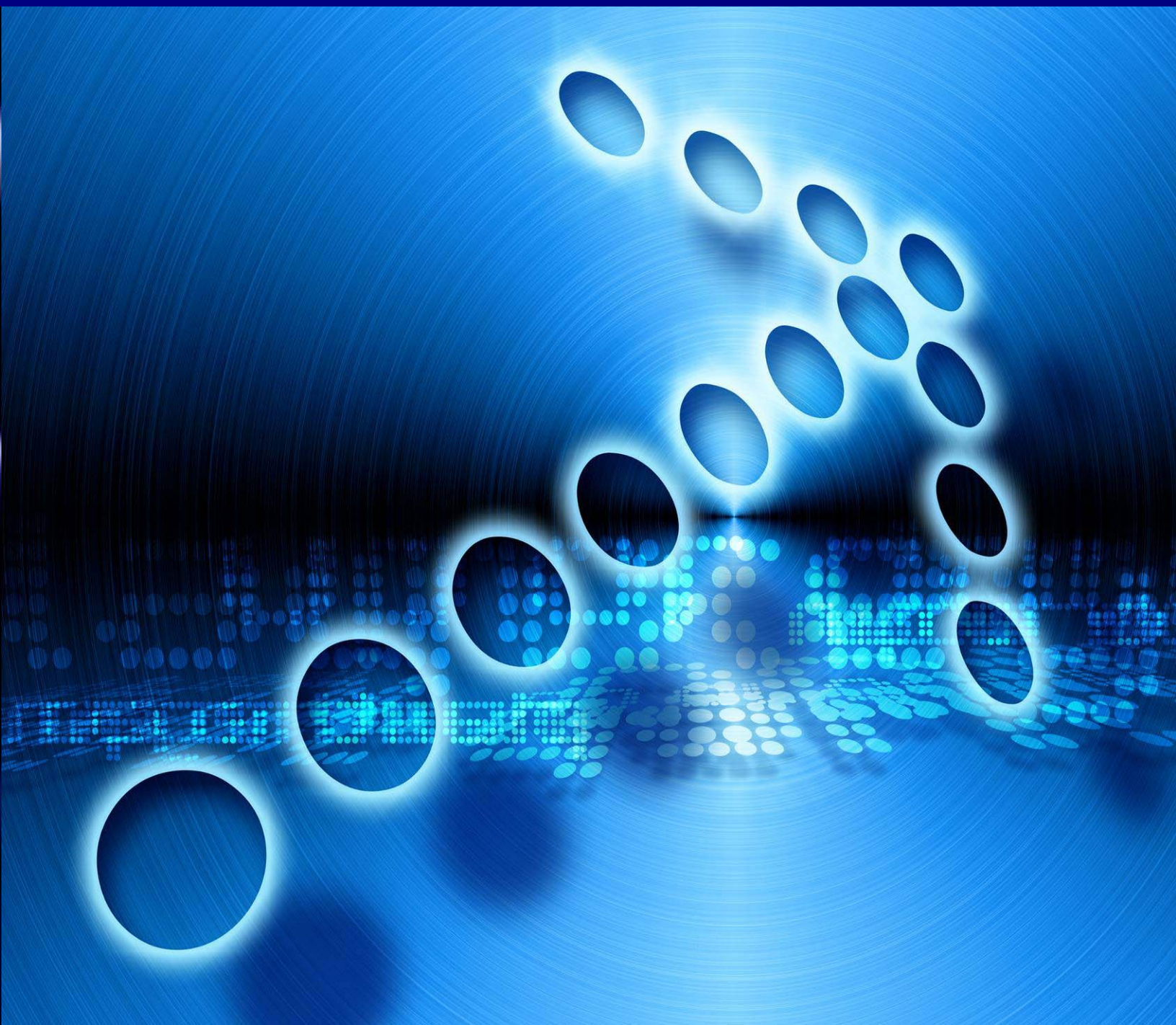
TIFF versus PDF For Document Storage

Aquaforest Limited

www.aquaforest.com



TIFF versus PDF for Document Storage

The choice of storage format for electronic documents can have significant and far-reaching consequences. This short white paper provides an overview of the TIFF and PDF document formats and discusses their relative merits as a format for electronic document storage.

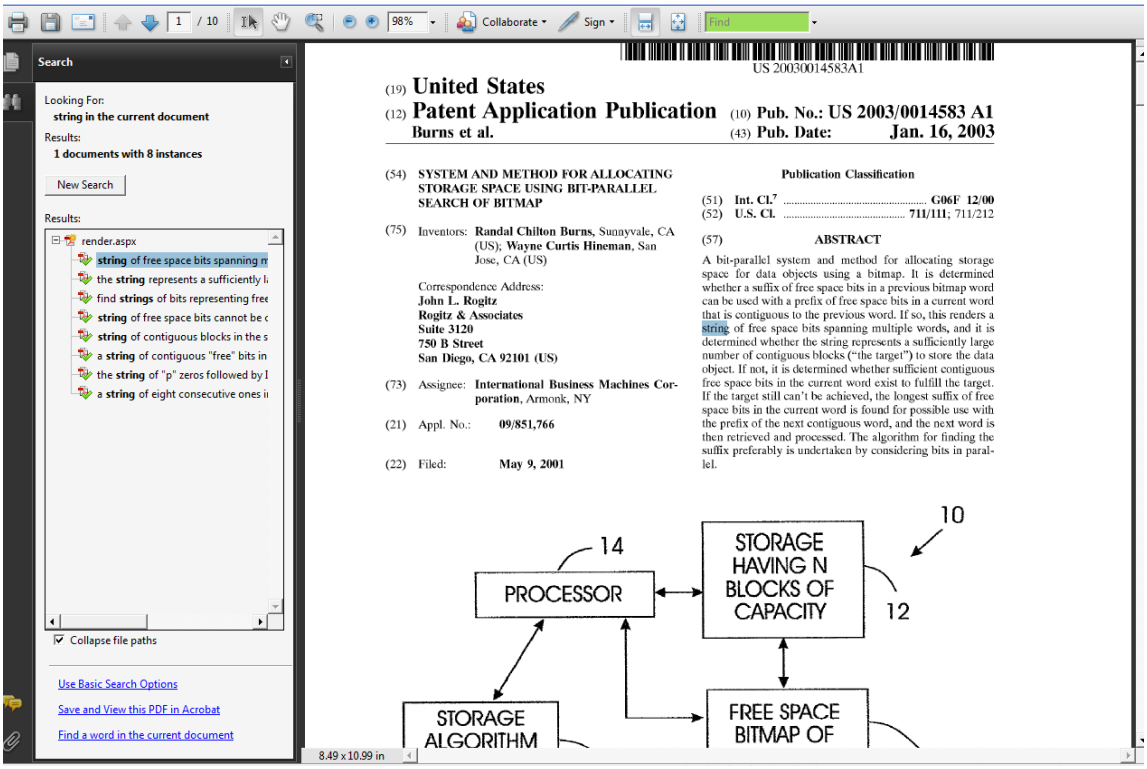
PDF Format - Overview

Background

The PDF format was created by Adobe in 1993 for the purpose of portable document exchange between systems and applications. The PDF specification has been updated and published and Adobe over the years but PDF is now an open standard **[2]** that was officially published on July 1, 2008 by the ISO as ISO 32000-1:2008.

PDF File Format

A PDF file encapsulates a complete description of a fixed-layout document that includes the text, fonts, raster images, and vector graphics which the document comprises. It includes support for JPEG, JPEG 2000, JBIG2, Group 3 and Group 4 images.



An example of searching a "searchable" PDF

Types of PDF

PDF Type	Description
Normal	This is the most common type of PDF and is most typically created from a document such as Microsoft Word. It contains the full text of the page with appropriate coding to define fonts, sizes, etc. and will provide a faithful print of the original.
Image Only	This is a PDF that has been created from one or more images – most commonly as a result of scanning a document either directly to PDF or by converting a scanned TIFF image to PDF. These files do not contain any searchable text and most often comprise a set of Group4 or JBIG2 images in a PDF "wrapper".
Searchable	A "Searchable" PDF is an "Image-Only" PDF that additionally contains a hidden layer of text generated by an OCR engine. This enables the file to be searched in the same fashion as a "Normal" PDF. Text can be copied and pasted.

TIFF Format - Overview

Background

The TIFF format was created by Aldus Corporation in the mid-1980s in order to create a standard file format for storage of scanned images. The TIFF specification **[1]** is now controlled by Adobe although no major update to the specification has taken place since 1992.

TIFF Format

It is important to understand that the TIFF format itself is a file format, not an image format. A TIFF file can be thought of as a container for one or more images each of which may be of a different type.

The most common types of image included in TIFF files are shown in the table below.

Compression Type	Typical Usage
Group4	Most commonly used for black and white ("bitonal") scanned images.
Group3	Used for Faxes
JPEG	Used for Grayscale and Color scanned documents. There are two definitions of JPEG in TIFF, type 6 and type 7 and there have been interpretation issues with type 6 in particular. Consequently these images are not always well supported.
LZW	Used for Grayscale and Color scanned documents. Due to historical patent issues this is not always well supported, although as the patents expired in 2004 more recent software should provide good support.
Uncompressed	Often output from graphics applications.
Others	Other less commonly used schemes include ZIP, Packbits, RLE.

TIFF Metadata Tags

Metadata may be stored in TIFF files by the use of tag fields. There is a set of baseline TIFF tags which should be supported by all TIFF software. The baseline tags include Scanner, Make, DateTime, Software etc. There is also a set of defined extension tags. Furthermore developers can implement “Private” tags.

TIFF or PDF: Key Decision Criteria

This section considers the key factors to be considered when making a decision on which format to use.

File Size

PDF can enable production of smaller files for the same document but there are a number of issues to be considered. It is helpful to consider in turn each of the three "types" of PDF previously described.

PDF Type	Comments on PDF File Size versus TIFF
Normal	"Normal" PDF files will typically be smaller that an TIFF image version of the same document as the text and internal PDF description of the pages will almost always be smaller than the image representation held in a TIFF file, especially for color documents. Note, however, that when fonts are embedded in a PDF file (as required with PDF/A) the PDF file size can increase and for documents will small numbers of pages, the equivalent TIFF could be smaller as a result.
Image Only	For Image-Only PDFs, the difference in size will primarily be down to compression techniques. A Group 4 TIFF will be slightly smaller than the PDF equivalent if the images in the PDF file if the PDF file also uses Group 4 compression for the images. However, if the PDF file uses more modern compression schemes such as JBIG2 for bitonal images, JPEG 2000 for color or more advanced MRC compression method for color, then the PDF file can be significantly smaller.
Searchable	A Searchable PDF can be expected to be 5 – 10% larger than the equivalent Image-Only file.

Longevity

Whilst the historical widespread adoption of TIFF will ensure the availability of viewers for some time to come, for long term standards-based archiving, the presence of the PDF/A standard makes PDF the clear choice. PDF/A (ISO 19005-1:2005) is a tightly defined subset of PDF 1.4 that is suited for long term archiving.

Searchability

TIFF was designed purely to store images, not text, so standard TIFF can only become searchable by having an OCR process create a separate text file which is separately indexed. Although Microsoft introduced “text searchable TIFFs” that can be created using Document Imaging component of Office Professional, it is not a widely supported feature.

PDF was designed to enable full searchability of documents as it can contains the text within the document itself. So unless the PDF is “image-only” it can always be searched - both within a PDF viewer application and via external search tools configured to index PDFs.

Security

TIFF does not have a security model so users can only be allowed or disallowed access to the document whereas PDF has a sophisticated security system which can be used to set document access passwords or restrict usage.

Portability

Both TIFF and PDF are portable across operating environments. The ubiquity of the Adobe viewer across all platforms including UNIX and Linux provides a consistent interface.

Metadata

TIFF files can make use of TIFF Tags to store simple metadata as does PDF via the PDF Information Dictionary. PDF allows for much more sophisticated XML-based metadata to be embedded in PDF files via XMP **[3].**

Document Structure

Standard TIFF does not include any method for defining document structure beyond sequencing pages. PDF documents can include Bookmarks, Hyperlinks, Tags and Annotations.

Accessibility

Appropriate use of PDF document structure enables effective use of assistive technologies such as screen readers. TIFF does not support the necessary document structures to enable this.

Suitability as an Input Format

Applications such as Microsoft Word and Graphics packages can typically accept TIFF images as input, but not PDF files.

Quality of Presentation

In general “Normal” PDF files will offer the best quality for printing and viewing, as the bitmaps stored in TIFFs or image PDFs by their nature have a limited resolution. However, for most practical purposes the image quality of TIFFs at a suitable resolution will be sufficient.

Availability of Viewers

Viewers for both PDF and TIFF are widely available free of charge, although the precise functionality of TIFF viewers in particular can vary significantly.

Suitability for Web Delivery

Like nearly all multi-page document file formats neither TIFF nor PDF files are natively supported by any common web browser. Consequently, to display the files as part of both require either a client-side plugin or server-side processing.

The free Adobe Reader provides a browser plug-in for PDF Files. A number of plugins for TIFF viewing are also available. Server-side products such as TIFF Server **[4]** enable browser delivery of both TIFF and PDF files without plugins.

PDF files can be “web optimized” (linearized) which optimizes them for web delivery to a plugin that enables pages to be displayed as soon as they are downloaded rather than waiting for the entire file to be received before displaying the first page.

Ease of Conversion

Conversion from TIFF to Image PDF or Searchable PDF can be made with a number of tools such as Autobahn DX **[5].**

Conversion from PDF to TIFF can also be made by such tools, although clearly there may be a loss of information such as document structure and metadata.

Legal Admissibility

Legal admissibility and the weight of evidence attached to electronic documents vary by country. For example, the position in the UK is dealt with by the BSI publication “Code of practice for legal admissibility and evidential weight of information stored electronically” **[6].** Both TIFF and PDF files can be perfectly acceptable as long as appropriate processes are followed.

Summary

TIFF has proven to be a robust file format for document storage over the last 20 years and still has a role to play in document storage. Increasingly, however, business needs dictate that the greater functional capabilities of the PDF format are required. The ISO standardization of the PDF format has put to rest any qualms about the long term openness of the format.

References

[1] Adobe TIFF Home Page

http://partners.adobe.com/public/developer/tiff/index.html

[2] PDF Specification

http://www.adobe.com/devnet/pdf/pdf_reference.html

[3] XMP

http://www.adobe.com/products/xmp/

[4] TIFF Server

http://www.aquaforest.com/en/tiff_plugin_tiff_server.asp

[5] Autobahn DX

http://www.aquaforest.com/en/autobahn.asp

[6] BSI Legal Admissibility

http://www.bsigroup.com/en/Shop/Publication-Detail/?pid=000000000030104568