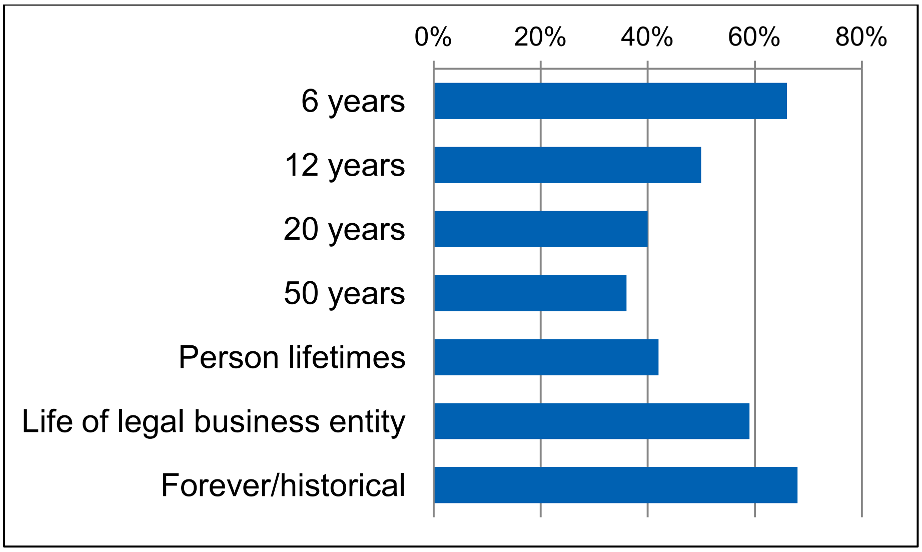
The Legal Case for PDF/A

It's not just law firms that need to keep documents for a long period of time, but it sometimes feels like this industry takes the concept to a whole new level! In addition to the thousands of paper documents nearly every firm accumulates, most companies now have the need to operate some form of digital document storage and make use of so-called "Document Management Systems". This means that many law firms now have the need to maintain (at least) two types of document archive, i.e. paper & digital.

The aim of this article is to draw attention to the fact that there's an underlying problem with digital document archiving, which may only present itself as a major problem in the future when there's a need to go back and view this digital data. The real, risk discussed below, is not whether someone can actually search, find and retrieve the digital information, but whether or not it will be possible for the retrieved file to be viewed exactly as it was when it was created. Will the pages & images be displayed correctly, and will cross-references still be accurate?

How long do you need to store your files for?

A recent survey by AIIM (Association for Information and Image Management) asked several companies how may years they needed to keep electronic records for, and the results are show in the table below.



It's clear from the above that many respondents had multiple retention strategies for their documents but the overall message is clear; most companies need to keep documents for many years. What's less clear is how these companies plan to ensure they can access the documents in the future.

Accessing and displaying an archived document correctly is critical for the legal sector, and it's become such an area of concern in the USA and Europe that new legislation relating to digital archiving is already being introduced in some areas to ensure that it's going to be possible to access legacy documents for many years to come.

Below, we will look at how PDF/A documents meet the requirements for long-term archiving and discuss how law firms can go about the business of converting their existing paper and digital archives to PDF/A format.

Durable Documents with the PDF/A Standard

There are many documents in the legal industry that need to be retained for a long period of time. In the days when everything existed on paper – in the pre-digital era – the main problem was remembering which index file, folder, or shoe box you'd used to store your letters or contracts. In today's world of digital documents, the task of archiving is fundamentally different.

Thanks to search functions or database solutions, even the most forgetful of us can easily find a particular document or photo on our computers. However, there are certain risks and uncertainties that might influence the shelf life of digital documents. These risks do not only arise from the physical durability of the storage medium used, although it is clear that magnetic tape, CDROMs, and DVDs will not necessarily last any longer than paper and ink.

And although photographic prints dating from the last century still exist today, it's still debatable whether or not we will be able to view the millions of digital snapshots being taken and stored on mobile phone memory cards all over the world many years from now.

In addition to the restrictions imposed by the limited lifetime of storage media, the document format and software used also present a considerable challenge for the durability of electronic documents.

Yesterday's, Today's and Tomorrow's Software

It's a common problem: Opening old documents in brand-new programs doesn't always work. The rate of success for the opposite direction (new documents in old programs) is even less encouraging. Software developers do try to achieve backward compatibility that enables files that are, say, five years old to be opened using a current program release. However, this can change the layout and page rendering, meaning that not everything is displayed exactly as it ought to be. More recent software tends to generate documents with additional features that older versions may not be able to display. In some cases, it is not even possible to open current files in previous versions of a program. For example, whereas a Microsoft Word 95 file can normally be opened in Word 2003, it is not possible to open a Word 2003 document in Word 95.

Because software production cycles are often very short – one major release per year is not unusual – the challenge that arises from new program developments is greater than that caused by the aging of storage media. The successful long-term archiving of digital files is at least as threatened by the constant rollout of new program versions as by damaged data or storage media.

TIFF as an archive format

For a long time, many public authorities and companies that need to store large quantities of correspondence, records, invoices, contracts, and similar information in digital archives have been using the pixel image format TIFF (Tagged Image File Format). This format digitises documents containing text and images pixel by pixel.

TIFF is an established image file format that has both advantages and disadvantages. Because pixel-based formats like TIFF store the *appearance* of documents, problems with missing graphics and fonts do not occur, since the format stores all of the document elements as an image. Since TIFF is widespread in use and is subject to few file handling complications when upgrading to a new program version, many users believe that the future of the format is guaranteed. However, while TIFF may indeed be a de facto standard, it is not an official norm for safe archiving. Other disadvantages include the relatively large file size and the fact scanned text documents saved as TIFF cannot be searched because they are only images. Any OCR text (Optical Character Recognition) created from the TIFF image must be stored in a separate file.

PDF data containers

The development of PDF (Portable Document Format), which Adobe Systems has been promoting since 1993, has significantly simplified data management and exchange for a great number of users from completely different fields. PDF allows obstacles that can arise during the transmission or storage of files to be neatly avoided.

- PDF files can be opened on all established operating systems. Free PDF readers are available for all of the important platforms including Windows, Apple, Linux, and mobile devices.
- With PDF, the document layout is true to the original. Since PDF can incorporate different types of content, such as text (and the relevant fonts), images, and graphics, nasty surprises relating to missing illustrations or incorrect fonts – like those that occur when Word documents are opened on another computer – are not usually a problem.
- PDF is an open format. This means that companies other than Adobe Systems (who invented PDF) can develop software for creating or displaying PDF. The “release” of PDF by Adobe has brought independence for both users and developers and, as a result, there is a high probability that there will still be programs for generating and displaying PDFs in decades to come.

So, can law firms who want to keep documents for long periods of time trust in PDF to make sure that their documents will work just as well in fifty years time as they do today? It might well be the case that ordinary PDF files created today will still work without any significant problems in the future. However, only the new PDF/A standard can guarantee that users will be able to view exactly the same content as when their documents were created. This format brings the kind of legal certainty that can be decisive in many business and administrative contexts.

Why PDF/A and not PDF?

Why has a special PDF standard now been defined for archiving documents? Are traditional PDF documents not “good enough” for long-term archiving?

PDF has some excellent characteristics that lend themselves to the creation of archive documents. Like a container, a PDF can incorporate completely different elements such as text, images, and fonts. In addition, it reproduces layouts that are true to the original and it is cross-platform capable.

However, certain requirements must be met in order to enable the exact reproduction of content. For example, it is essential that fonts must be embedded; a link to the font in question is not sufficient. If a font is ***not*** embedded in a PDF document it means that if, in 10 years time, a user who tries to open a document does not have a required font on his or her computer, special characters or symbols will not be displayed correctly. Imagine the problems this could create – a critical piece of information may be lost from a case file simply because the font used to display the missing characters is no longer available.

In simple terms, a PDF/A document is just a traditional PDF document that fulfils precisely defined specifications. In order to prevent users from repeatedly having to test and discuss the best appearance of a well-functioning archive PDF, industry experts decided in 2002 to work together to develop the PDF/A standard.

The introduction of the PDF/A standard

The PDF/A standard for long-term archiving was adopted by ISO (International Organization for Standardisation) in autumn 2005, published with the number ISO 19005-1:2005 and based on PDF specification 1.4. A more recent upgrade to the specification (PDF/A-2) now allows PDF/A documents to be based on the additional functionality found in PDF specification 1.7.

The PDF/A standard aims to enable the creation of PDF documents whose visual appearance will remain the same over the course of time. These files should be software-independent and unrestricted by the systems used to create, store, and reproduce them.

Many new PDF/A tools and solutions for creating and verifying files have entered the market since the introduction of the standard − from small tools for individual users who want to create PDF/A documents every now and again, to extensive server solutions from companies like LuraTech that can create a hundred thousand archive documents in just a few hours time, simultaneously adding OCR text and compressing them to a fraction of their original size without loss of quality.

These tools not only safeguard future access to the documents, but they also make it possible to search their contents and reduce the associated storage & power costs.

How to create archive PDFs

There are many different conditions that might be encountered when creating PDF/A files. The process differs depending on whether existing PDF documents are already available or whether they need to be generated from working files such as Word or PowerPoint files.

- **PDF/A files from digital files:** Law firms typically have many documents already in digital form, created using e.g. word-processors and spreadsheets. These “born-digital” files can all be converted to PDF/A using server-based solutions from LuraTech and others.
- **Converting scanned paper documents to PDF/A:** Often, documents that exist only on paper, such as contracts, case files etc., need to be digitised using a scanner. Historically, the results of the scanning process have usually been stored as Bitmap TIFFs, but immediate conversion to compressed PDF or PDF/A is increasingly being implemented.
- **Creating PDF/A from PDF:** Many users already have PDF documents that are not PDF/A-compliant. It is often not possible to recreate such documents from the source program because, for example, they were not created locally but were sent to the user in question by e-mail. Fortunately, there are programs from LuraTech and others that can take these existing PDFs and convert them to text-searchable PDF/A files.

- **Is this really a PDF/A file?** When working with PDF/A on a daily basis, file verification is also important. Is it sensible to believe the sender of a PDF document when he or she says that it’s a PDF/A file? Before received files are saved in an archive they must be checked to make sure that they are PDF/A-compliant, for which software tools already exist from companies who are members of the PDF/A Competence Centre (www.pdfa.org).

How can law firms benefit from PDF/A?

Many different types of content can be saved as PDF/A files:

Archiving e-mails as PDF/A: Today, more and more correspondence, some of it of a contractual nature, is being sent by e-mail. Archiving emails to PDF/A is easy to implement using the various PDF/A server-based tools available, including any attachments to the email (which can also be automatically converted to PDF/A).

Plans, maps, and design drawings: Digital maps, architectural drawings, and construction plans all form part of case archives and are usually compressed very efficiently to small PDF/A files that retain all of the formatting of these documents.

Signed digital contracts: An increasing amount of business correspondence is sent electronically. PDF/A documents can be digitally signed to enable legally effective contracts to be concluded using only digital means.

Correct colours in image documents: PDF/A also enables the accurate display of colours, an important advantage when working with digital image data in e.g. insurance claims and medical records.

Accessible PDF files: In the USA, accessibility in the digital world has been an issue for a long time – especially for the Internet. Enabling the accessibility of information to visually impaired members of society is now also on the agenda in Europe. Since PDF/A specifically supports structured content in PDF documents, it is ideal for processing accessible PDF documents that can be read out by screen readers.

Which file formats are suitable for archiving?

In most law firms, many users simply archive the original documents (for example, Word, Excel, or PowerPoint files). This can lead to nasty surprises with regard to reliable display and future usability of the files. This method of archiving is, therefore, not recommended.

TIFF-G4 has been the de facto standard for numerous companies and administrative departments for many years. TIFF-G4 files are monochrome, black and white TIFFs (bitmaps) that can be archived in a way that saves space thanks to compression features based on fax technology.

The JPEG image format is commonly used for colour documents, which has the benefit of producing relatively small file sizes.

The PDF/A standard and XPS (XML Paper Specification), which was developed at the end of 2006 by Microsoft, are relatively new formats. Both of these formats offer facsimile quality as well as supporting structured content, thereby allowing the reliable and complete indexing of text. PDF/A is already standardized, but XPS still has to stand the test of time.

The table below gives an overview of the long-term archiving features offered by both these formats.

	PDF/A	XPS	TIFF-G4	JPEG	DOC (Word)
ISO standard for archiving	Yes	No	De facto standard, but not official	No	No
Font security	Yes. Strictly defined specifications in the PDF/A standard	Yes. Fonts are embedded	No fonts exist, since the files are pixel images.	No fonts exist, since the files are pixel images.	No. The display of fonts can vary on different computers.
Searchable text	Yes. Created by OCR	Yes	No standard procedure for storing OCR text	No	Yes
Consistent colours	Yes	Possible	No	Possible	No
Images and graphics are fixed document parts	Yes	Yes	Yes	Yes	No
Structured data	Yes. With tagged PDF	Yes. With XML	No	No	Possible
Cross-platform capable	Yes	No	Yes	Yes	Only with restrictions (font problems)

Converting existing JPEG and TIFF-G4 archives

There are basically two options for converting large document archives that currently use TIFF-G4 or JPEG to PDF/A along with their existing inventory: Permanently or temporarily. If the number of documents handled is not too high, and regular access to the data is required, converting the image files to PDF/A is worthwhile. This involves using mass conversion solutions that package pixel information into PDF and can enable text searching using OCR (see below). However, if users only need to call up data from an archive every now and then, 'on the fly' solutions can be used to generate a PDF/A file from a particular original image file.

PDF/A Creation: Analogue, Digital, and Mass Processing

PDF/A is always the destination, but the point of departure can differ greatly from user to user. This chapter concentrates on three main tasks:

1. Converting paper documents to PDF/A,
2. Exporting Microsoft Office files and other documents in a way that allows them to be archived
3. Mass Processing PDF archive files.

The special process flow for converting existing PDF files to PDF/A is explained in detail later on.

PDF/A from Scanned Documents

'Analogue to digital' conversion is normally required when users have received the documents that need to be archived as printed pages rather than creating them themselves. For example, customer engagement documents are often sent as printouts by mail. In some cases, documents that need to be retained are only available as printouts because the digital originals have been deleted from users' computers. In addition, many old documents may have been created by typewriter or by hand in the days before computerisation.

In such cases, the only way to digitise document pages is by using a document scanner. As well as the type of scanner (flat-bed scanner or a device with bypass feed), the scope of features provided by the software used to scan the documents also has a bearing on whether or not the digitisation process can create a faultless PDF/A, and dictates which additional features can be used to enhance the usability of the document such as OCR & compression. This latter option is particularly useful if colour images are saved by the scanner, which are usually very large files that compress very efficiently to PDF/A.

In general, all modern scanners support the use of PDF as the initial format (in addition to image formats such as JPEG or TIFF) but not all scanners are currently able to generate PDF/A and few can compress the final file. This is not a problem, however, since software tools can e.g. monitor folders where the scanner is saving TIFF images and immediately OCR & compress them before saving as PDF or PDF/A.

Compression to PDF

The generation of PDF files from digitised paper documents has a disadvantage – the image data for such files normally requires more memory capacity than digital pages of text. In other words, a PDF generated directly from a Microsoft Word document will be considerably smaller than a PDF file that is generated by scanning the printout of the Word file.

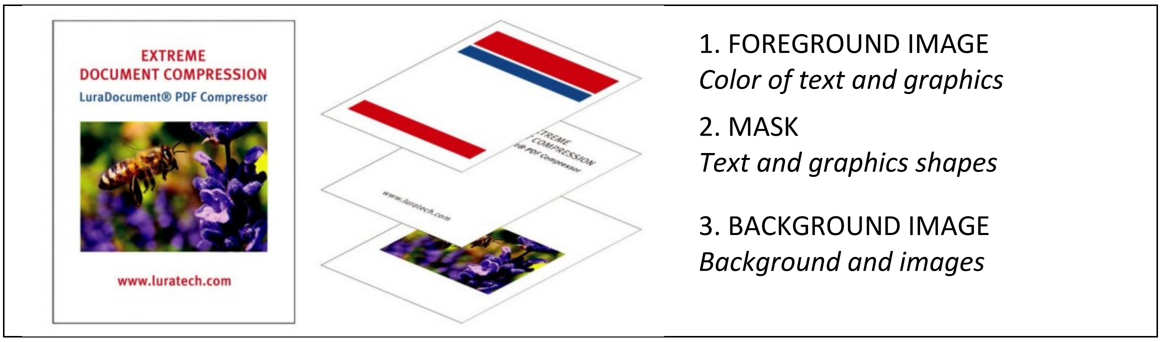
This comparatively high file size is particularly cumbersome if a large number of documents with many pages need to be archived, particularly if the documents are scanned in colour, which can result in extremely large TIFF files.

Various image compression types have been developed over the past years to enable users to save memory space when storing image data, the best-known of which is the JPEG file format. This file format is extremely common as a result of the increase in popularity of digital cameras for personal use, nearly all of which save their images as JPEG files on memory cards. However, JPEG is known as a “lossy” compression type, i.e. if it used for scanning colour documents, information on the document will be lost when it is converted from TIFF to JPEG. Fortunately, there are other types of compression available that do not degrade the image in the same way as JPEG, thereby retaining a higher level of quality in the compressed image.

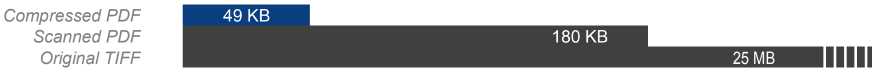
In addition to the ***type*** of compression, the ***degree*** of compression is also important for a scanned text because higher compression levels can render the image/text progressively less clearly. Most PDF Compressor tools allow the user to select the level of compression required and some also allow the selection of different compression types, thereby giving the best possible results for the types of document being compressed.

Advanced Compression Techniques

Where it is critical to maintain the quality of the compressed file, for example to be able to reliably search for dates and names in a legal case file and to be able to view details on photographs, it’s important to choose the right compression technology. This usually involves a technique called MRC Compression (Mixed Raster Content), which analyses a document prior to compression in order to determine the most appropriate compression technique to be used. This can be seen in the diagram below:



In the above example, the segmentation algorithms first separate colour information from the text as well as any photographs. Then, different types of compression are applied to each component of the original image, thereby giving maximum compression for the overall document with minimal loss of quality. Typical compression results achieved by the use of MRC compression are shown below.



Benchmark: LuraDocument PDF/A to other conventional formats
(Original: 300 DPI, full colour RGB, letter size)

Compression of archive material not only brings benefits in terms of reduced power & data storage costs, but also by facilitating remote archiving. There are many options for storing digital information offsite, using hosted facilities "in the cloud", but all come with one inherent problem: when the time comes to retrieve the data, the speed and quality of the user's Internet connection will directly affect how quickly a large file is displayed to the end-user. In other words, compressing your archive material to PDF/A means that it's going to be significantly quicker to retrieve the material in the future (up to 90% in the case of scanned colour documents).

PDF/A in the USA Federal Courts

The potential risk of not being able to reliably access archived digital information long into the future was identified by the Administrative Office of the US Courts in 2010, and has subsequently been taken up by multiple District Courts including New York and Wyoming. After 15 years of the federal court adopting PDF as the standard for their central case management system, and with approximately 500 million PDFs in the archive, the decision was made to ultimately require district courts to file in the PDF/A format.

The move to PDF/A was made in direct response to the ISO certification of PDF/A in 2005, and the realisation that PDF/A is the only way to guarantee future readability for legacy case related documents.

PDF/A Outside of the USA

Throughout Europe and Asia, PDF/A has been adopted as the standard for long-term government archives, and is mandated in many other industry verticals. Some examples are:

- In 2008, the Swiss Federal Council began requiring PDF/A format for all communications between the government and citizens.

- In Austria, all land register deeds must comply with PDF/A, in order to prove the authenticity of its documents through a qualified digital signature.
- In Germany, the use of PDF/A has been recommended for e-government applications.
- The European Commission has also included PDF/A in the recommended data formats for scanned documents and long-term archiving in order to standardise the exchange of documents between the European Commission and the governments of its member states.
- The European Court of Human Rights has recently begun to archive documents in PDF/A format and is using compression technology to minimise storage space and cost as well as facilitate document retrieval times
- The Brazilian & Dutch Governments have now mandated the use of PDF/A for archiving, along with the Australian Public Records Office

In parallel with the above, various options for archiving in PDF/A have begun to appear here in the UK. From document imaging systems to specialised eBible and data migration tools, the technical barriers to securing long-term accessibility to your archived documents have largely been removed and all that's required now is the commitment to make the change. It's uncertain when or if there will ever be a UK mandate for PDF/A in the legal vertical but while we wait to find out, there's little doubt that the time is right to begin protecting your future.