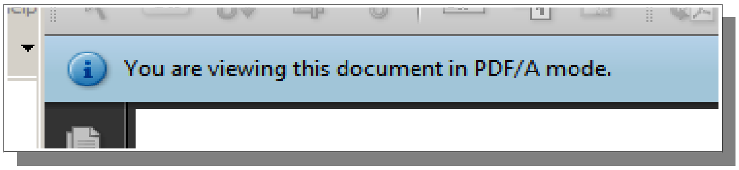
# The Content that Endures:
# What to know about PDF/A

By *Duff Johnson*
President, Document Solutions, Inc.



You are viewing this document in PDF/A mode.

What's the difference between a document and the software used to view the document?

In the paper or microfiche worlds, no software is needed, so the question is meaningless. The only potential barrier to legibility is the physical condition of the document.

Electronic documents are different. The physical condition of the document is assumed, otherwise software can't even begin to do its thing. Understanding precisely how to display and print that document, however, can be somewhat more complicated.

What happens in the year 2023, when someone has to open an Outlook PST file from 2003 to settle a lawsuit? For those to whom such questions matter, no-one wants to consign their precious electronic documents to proprietary software, or the fortunes of one company.

First released by Adobe Systems in 1993, PDF became *de facto* "electronic paper" during the late 90s. At 15, PDF still isn't old enough to vote. In 10, 15, or 30 years, who knows what software will be used to view today's documents? This is why PDF/A, an open international standard for archiving PDFs, is so important.

Ralph Cafiero, CVISION Technologies COO, thinks this is something that most corporate leaders don't usually think about – and yet when they do, they sit up straight. "They realize that when the software of the day fails to open the files they saved so carefully 15 years ago, they could be in a lot more trouble than they'd ever imagined," Ralph says. In Europe, they are already figuring this out, as we'll see.

PDF is an inherently flexible format, and can contain far more sophisticated content compared with images, fiche or tagged-text. Even so, the rules for creating PDF used to be (and to an extent, remain) loose, allowing lazy, sloppy or simply careless developers to make poorly-constructed and even broken PDFs that Adobe nonetheless feels compelled to attempt to open with their Reader. Some users complain that Reader is "bloated", and compared to many alternatives, it is. That said, unlike most alternatives, Adobe Reader will open almost any bundle of bytes that claims to be a PDF file.

Quite apart from such "difficult" PDFs, the power and flexibility of the format presents another challenge for archival purposes; the possibility of content contained within a PDF that viewing software can't render.

When created by quality software, PDF is already far more reliable for archival purposes than Word, Excel or Outlook. PDF/A, the Archive standard for PDF, is a core subset of the larger PDF Specification. PDF/A is designed specifically to address the concerns of organizations with electronic document retention policies. Launched by AIIM and NPES in 2001, the PDF/A standard was first published as ISO 19005-1 in October, 2005.

"By stipulating not only a file format but also viewer requirements (above and beyond those in PDF Reference) – PDF/A gives you both pieces of the puzzle." says Leonard Rosenthol, Adobe's PDF Standards Evangelist.

## *Why PDF at all? What's the matter with TIFF?*

TIFF images, while traditional for electronic archival, are a proprietary (albeit published) format, and there are many incompatible variations on the theme. According to Dr. Hans Bärfuss, CEO of PDF-Tools AG, tools for migrating large image archives have to deal with a wide variety of image format dialects and proprietary tags. "Developing such tools makes it clear how much better a well-designed standard such as PDF/A really is," he says.

Unlike TIFFs, PDF can contain text as well as color raster and vector graphics and metadata. Electronic-source and OCRed PDFs are searchable, a key feature for almost every organization, and they are usually smaller than TIFFs as well.

The PDF format facilitates the organization of pages into documents with associated metadata, and provides navigational features such as bookmarks and links. TIFFs can become PDF files, and even PDF/A files, quickly and easily. Olaf Drümmer, CEO of callas software, a leader in PDF/A technology, points out that "PDF and thus PDF/A, elegantly bridges the gap between analog (scanned) documents and electronic documents."

PDF/A includes two distinct conformance levels, PDF/A-1b and PDF/A-1a. PDF/A-1b focuses on ensuring the document displays and prints correctly. PDF/A-1a additionally requires that text include sufficient information to make Unicode mapping possible, and additionally, stipulates that document content be tagged to reflect the semantic relationships between objects such as text and images (concepts such as "heading", "paragraph", "table", "list" and so on). These requirements present serious technical and operational challenges. Says CVISION's Cafiero, "For most corporate purposes – retention of a printable page - PDF/A-1b is considered sufficient." Government agencies, however, often work under accessibility regulations such as Section 508, which stipulates the semantic tagging of content, implying PDF/A-1a compliance.

Dr. Bärfuss wants implementers to look beyond the format itself, and notice that unlike image-files, PDF includes a metadata standard; XMP. "I can't emphasize enough the importance of a metadata standard such as XMP for the archiving community," he says.

## *Is PDF/A really an archive standard in the same way as TIFF? Can a 100% reliable PDF/A viewer be created without reference to Adobe's software?*

In discussing archival formats, the Library of Congress covers PDF/A, but notes that in rendering normal text "Good support is possible, but not guaranteed." What does that mean, given that PDF/A is supposed to be a standard for archiving documents?

PDF/A depends on the PDF Reference (which became ISO 32000 as of July 3, 2008) to provide the specific information required to construct PDF files, and the PDF Reference is more of a dictionary than it is a cookbook for building PDF files. According to CVISION's Cafiero, "Certain features such as linearization are somewhat vaguely specified in the PDF Reference, making it difficult for non-Adobe providers to implement this feature."

Some interactive features are permitted, but PDF/A requires that an unambiguous visual representation for hyperlinks, comments and form fields be present. The Reference doesn't tell developers how to draw every possible object, so different applications may create different representations for the same functional object (a form-field, for example). This quality The authoring application has the final say, and that's as true to the essential intent of PDF as any other property of the format.

As Callas's Drümmer points out, PDF/A puts you lightyears ahead of any other single format in terms of inherent reliability and software independence, and offers far more functionality than simple image files.

Dwight Kelly, President of Apago, Inc. , says that today "...there are several implementations of PDF apart from Adobe that do a very good job of rendering PDF files." Once the file is rendered and (if required) corrected," Kelly says, "...we're confident that PDF/A is a reliable archiving format."

## *Implementing PDF/A workflows in business environments*

Callas's Drümmer says "the challenge of implementing PDF/A itself explains why PDF/A is so important." Workplaces collect files of all types and sources, PDF included. Flawless, exception-free automated conversion from non-PDF source documents to PDF/A is the holy grail in this business, and it's next-to-impossible to achieve completely. Even with PDF files, some pages can't be automatically fixed, or (more commonly) the fix involves the summary removal of scripts, movies or other content excluded by PDF/A, with no regard for how that change might affect the document's contents when rendered.

Adobe's Rosenthol focuses on the avoidable problems in PDF creation. "The greatest challenge for PDF/A solutions comes from tools that produce poorly made PDFs," he says. Of course, Adobe wants the world to use quality Adobe software to create PDFs, but others concur that shoddy PDF creation software is responsible for most of the hard PDF/A validation and correction problems. Organizations concerned with document archival should be careful in their choice of such software, and should think carefully before adopting (or accepting files from) an application that makes PDFs that aren't PDF/A-friendly.

Lousy software is only part of the problem. Apago's Kelly points out that there's very limited support for "Save as PDF/A" in major applications, and that most PDF/A software doesn't provide industrial-strength workflows handling millions of documents.

There's a cost to any retention policy, and PDF/A at least offers a realistic model to dramatically enhancing long-term retention. "Let's just stop to point out that it's not going to be any easier to open that Word Perfect or Outlook file in 15 years, to say nothing of 50 years," says Drümmer. Point well taken, most would agree.

## *The technical challenges in PDF/A solutions development*

A variety of companies have created software to certify that files are PDF/A compliant and to fix those that do not comply. There are three basic scenarios for PDF/A solutions:

PDF/A Competence Center

- Desktop applications designed to facilitate PDF/A creation directly by the author.

- Server (or server-like) applications designed to validate, correct and/or flag large volumes of PDF files in an automated or semi-automated process.

- Desktop (or server) applications addressing PDF/A-1a requirements for tagged PDF.

Today, there are over 80 members of the PDF/A Competence Center, ranging from interested parties to developers addressing one or more of the above scenarios. Clear standards for semantic structure (accessibility) in PDF are still a work in progress, so support for PDF/A-1a is proving slower to build. While Adobe has invested to a degree in applications to support PDF/A-1a requirements, it's been a few years since they substantially updated Acrobat with improved tagging and structure tools.

Given the obvious business interest in PDF/A-1b, most vendors concentrate on the WYSIWYG level of the standard, PDF/A-1b. Even there, they are finding challenges.

As a practical matter, users tend to consider a PDF as "valid" if Adobe Reader can display the file. Adobe makes Reader go through back-flips to render every possible assistance to the most lamely cobbled-together PDF file, so expectations have therefore been set very high.

Since there are a wide variety of ways in which it's possible to mis-create a PDF file, "writing a PDF/A validator/fixer is a very complex task," says Apago's Kelly. Depending on the source, up to 15% of PDF files can't be fixed (ie, brought up to PDF/A-1b standards) without changing the document or removing content that doesn't comply with the standard.

CVISION's Cafiero emphasizes that conversion problems affect any archiving format. "In some cases, the only solution is to render the file to an image-based PDF, and that process is no less reliable than the process of converting to TIFF or fiche," he says. Of course, unlike TIFF or fiche, PDF image files may be OCRed, the resulting text is stored in the same file to provide searchability.

Apart from difficult PDF files, developing reliable conversions from source formats to PDF/A poses significant challenges. Take email, an obvious candidate for archival. Modern email can include HTML, JavaScript and CSS, as well as who-knows-what attachments. The problem of archiving email, therefore, is the same "holy grail" problem of reliable automated conversion of arbitrary files.

Of course, these details are of little interest to customers. Every vendor agrees with Callas's Drümmer, a board member of the PDF/A Competence Center, who says that what's high on their agenda is simply "just make it work, don't make me think".

## *The top workflow practices inhibiting PDF/A compliance*

According to Drümmer the biggest single problem is a lack of document policies. If an organization specifies that only 10 or 15 formats or applications will be used in-house, for example, this single fact can immensely simplify the PDF/A challenge. But as he says, "if one accepts all kinds of documents, one accepts all kinds of trouble."

Coming a close second is the failure to embed fonts. PDF/A requires that fonts used in the document be embedded – and too often, they aren't. Fixing this problem isn't trivial – the choices involve identifying and correctly embedding the font (if available), or else replacing the font with a look-alike. Either approach can cause problems, but no worse than the problems of converting to TIFF or fiche.

Looking the problem right in the eye remains a key challenge for many organizations. “PDF/A is directed at avoiding the loss of information,” says PDFlib President and PDF/A Technical Working Group chair Thomas Merz. “In a sense, conversion to PDF/A provides a benchmark for how easy it will be to render the file in the future. If it's hard to render it perfectly today, it may be impossible in 20 years... so it makes sense to do it today.” Merz says.

## *Why are the Europeans leading the US on PDF/A?*

TIFF remains the dominant digital archival document format in the US, but PDF/A has already proven persuasive in Europe. The dense interrelated profusion of super-national (EU level), national, regional and local governments contributes to a drive towards the most broadly capable document archiving technology, and in Europe, that technology is PDF/A.

In contrast to the European view, in the US, PDF/A is often seen as a cost, not a profit generator. This view may be changing, however. Scalable search technology means corporate archives are increasingly seen as treasure-troves of information. If the risks of data-loss in conversion are more-or-less equivalent between PDF, TIFF and electronic-source fiche, the superiority of PDF over TIFFs – searchability, navigability, high fidelity and an effective unification of paper and electronic sources in a single format – become obvious.

Drümmer says that for his EU customers “...nobody will blame you for choosing PDF/A... but you might be blamed for any other choice.” In Europe, it seems, going with PDF/A today is like buying IBM 30 years ago.

## *Where is PDF/A going from here?*

While government archivists were among the first to ask the sorts of questions that eventually led to PDF/A, awareness of the need to ensure that today's documents are future-proofed has grown to include heavily-regulated industries, major corporations, law-firms, and others with an interest in assured longevity.

In the United States, corporate interest in PDF is led from the pharmaceutical, banking and financial sectors, according to the leading 3rd party US-based PDF/A solutions providers, Apago and CVISION Technologies. Adobe's Rosenthol cites the example of Airbus, who gave a presentation about their usage of PDF/A at the 1st International PDF/A conference in Amsterdam last April.

There's little question that governments and businesses are ready for an archival format that improves on TIFF. Is PDF/A that format? It's hard to think of a better option, even in theory. PDF facilitates the unification of heterogeneous paper and electronic document sources in a single, standardized format, along with the relevant semantics. Serious long-term storage depends on such attributes, and PDF can deliver.

The next few years will see more vendors provide “Save as PDF/A” functionality from within their applications – a crucial step to address the problems raised by the wide variety of formats. We can expect more and better batch-archiving tools as well.

Managers should expect to see PDF/A on the agenda in IT and document-management and retention policy meetings in the near future.

### *A partial list of organizations in the PDF/A space*

- ISO (Publishers of PDF/A - ISO 19005)
- Adobe Systems
- AIIM's US Committee page
- The PDF/A Competence Center
- Apago, Inc.
- callas software, GmbH
- CVISION Technologies
- Datalogics
- PDFlib, GmbH
- PDF Tools AG
- OpenOffice.org