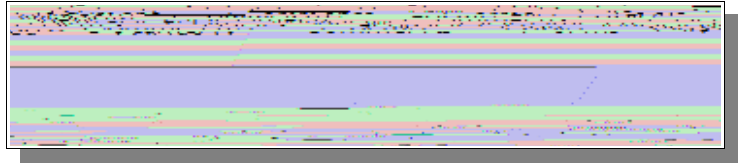


# The Content that Endures: What to know about PDF/A

By Duff Johnson  
President, [Document Solutions, Inc.](#)

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 2011, when someone has to open an (outlook) PS" file from 2000 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 1990s. By 2001, PDF still isn't old enough to vote. In 2002, 2003, or 2004 years, who knows what software will be used to view today's documents? This is why PDF2, an open international standard for archiving PDFs, is so important.

3alph 4afiero, 45 ISI (6 technologies 4 ( , thin)s 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 10 years ago, they could be in a lot more trouble than they'd ever imagined, 3alph 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, microfiche or tape. Even so, the rules for creating PDF used to be standard 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. PDF2, the archive standard for PDF, is a core subset of the larger PDF Specification. PDF2 is designed specifically to address the concerns of organizations with electronic document retention policies. Launched by [ISO](#) and [CEN](#) in 2000, the PDF2 standard was first published as [ISO 32000-1](#) in October 2008.

/By stipulating not only a file format but also viewer requirements :above and beyond those in PDF Reference; 7 PDF, gives you both pieces of the puzzle.0 says >Leonard Bosenthal, , Adobe's PDF Standards #vanalist.

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

"I\*\*ma!es, while traditional for electronic archival, are a proprietary :albeit published; format, and there are many incompatible variations on the theme. , ccording! to Dr. @ans Barfuss, 4# ( of PDF+ "ools , B, tools for mi!ratin! lar!e ima!e archives have to deal with a wide variety of ima!e format dialects and proprietary ta!s. /Developin! such tools ma)es it clear how much better a well+desi!ned standard such as PDF, really is,0 he says.

\$nli)e "I\*\*s, PDF can contain te9t as well as color raster and vector !raphics and metadata. #lectronic+ source and ( 43ed PDFs are searchable, a )ey feature for almost every or!ani8ation, and they are usually smaller than "I\*\*s as well.

"he PDF format facilitates the or!ani8ation of pa!es into documents with associated metadata, and provides navi!ational features such as boo)mar)s and lin)s. "I\*\*s can become PDF files, and even PDF, files, uic)ly and easily. (laf DrCmmer, 4# ( of callas software, a leader in PDF, technology, points out that /PDF and thus PDF, , ele!antly brid!es the !ap between analo! :scanned; documents and electronic documents.0

PDF, includes two distinct conformance levels, PDF, +-b and PDF, +-a. PDF, +-b focuses on ensurin! the document displays and prints correctly. PDF, +-a additionally re uires that te9t include sufficient information to ma)e \$nicode mappin! possible, and additionally, stipulates that document content be ta!!ed to reflect the semantic relationships between objects such as te9t and ima!es :concepts such as /headin!0, /para!raph0, /table0, /list0 and so on;. "hese re uirements present serious technical and operational challen!es. Says 45 ISI ( 6's 4afiero, /\*or most corporate purposes 7 retention of a printable pa!e + PDF, +-b is considered sufficient.0 Bovernment a!encies, however, often wor) under accessibility re!ulations such as Section 1&E, which stipulates the semantic ta!!in! of content, implyin! PDF, +-a compliance.

Dr. Barfuss wants implementers to loo) beyond the format itself, and notice that unli)e ima!e+files, PDF includes a metadata standard- F ? P. /I can't emphasi8e enou!h the importance of a metadata standard such as F ? P for the archivin! community,0 he says.

## Is PDF/A really an archive standard in the same way as TIFF? Can a 100% reliable PDF/A viewer be created with !t re"erence t Ad be's s "tware?

In discussin! archival formats, the >ibrary of 4on!ress [covers PDF, ,](#) but notes that in renderin! normal te9t /Bood support is possible, but not !uaranteed.0 What does that mean, !iven that PDF, is supposed to be a standard for archivin! documents?

PDF, depends on the PDF Reference :which became [IS \( '%&&&](#) as of Guly ', %&&E; to provide the specific information re uired to construct PDF files, and the PDF Reference is more of a dictionary than it is a coo)boo) for buildin! PDF files. , ccording! to 45 ISI ( 6's 4afiero, /4ertain features such as lineari8ation are somewhat va!uely specified in the PDF Reference, ma)in! it difficult for non+ , Adobe providers to implement this feature.0

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/A as any other property of the format.

As Dallas's DrCmmer 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 Helly, President of Palo, Inc., says that today "...there are several implementations of PDF/A apart from Adobe that do a very good job of rendering PDF files. Once the file is rendered and if required; corrected," Helly says, "...we're confident that PDF/A is a reliable archiving format."

## Implementing PDF/A workflows in business environments

Dallas's DrCmmer 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. \*Lawless, exception-free automated conversion from non-PDF/A source documents to PDF/A, is the holy grail in this business, and it's next-to-impossible to achieve completely. Even with PDF/A 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 Cosentino focuses on the avoidable problems in PDF/A 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/A 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.

Good software is only part of the problem. Palo's Helly 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.

"Here'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 (Lotus) file in 10 years, to say nothing of 100 years," says DrCmmer. 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. Here are three basic scenarios for PDF/A solutions:

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



Server :or server+li)e; applications desi!ned to validate, correct and2or fla! lar!e volumes of PD\* files in an automated or semi+automated process.

Des)top :or server; applications addressin! PD\*2 , +-a re uirements for ta!!ed PD\*.

"oday, there are over E& members of the [PD\\*2 . 4ompetence 4enter](#), ran!in! from interested parties to developers addressin! one or more of the above scenarios. 4lear standards for semantic structure :accessibility; in PD\* are still a wor) in pro!ress, so support for PD\*2 , +-a is provin! slower to build. While ,dobe has invested to a de!ree in applications to support PD\*2 , +-a re uirements, it's been a few years since they substantially updated , crobat with improved ta!!in! and structure tools.

Biven the obvious business interest in PD\*2 , +-b, most vendors concentrate on the WKSIWKB level of the standard, PD\*2 , +-b. #ven there, they are findin! challen!es.

, s a practical matter, users tend to consider a PD\* as /valid0 if ,dobe 3eader can display the file. ,dobe ma)es 3eader !o throu!h bac)+flips to render every possible assistance to the most lamely cobbled+to!ether PD\* file, so e9pectations have therefore been set very hi!h.

Since there are a wide variety of ways in which it's possible to mis+create a PD\* file, /writin! a PD\*2 , validator2fi9er is a very comple9 tas),0 says , pa!o's Helly. Dependin! on the source, up to -1 L of PD\* files can't be fi9ed :ie, brou!ht up to PD\*2 , +-b standards; without chan!in! the document or removin! content that doesn't comply with the standard.

4 5 ISI ( 6's 4afiero emphasi8es that conversion problems affect any archivin! format. /In some cases, the only solution is to render the file to an ima!e+based PD\*, and that process is no less reliable than the process of convertin! to "I\*\* or fiche,0 he says. ( f course, unli)e "I\*\* or fiche, PD\* ima!e files may be ( 4 3ed, the resultin! te9t is stored in the same file to provide searchability.

, part from difficult PD\* files, developin! reliable conversions from source formats to PD\*2 , poses si!nificant challen!es. "a)e email, an obvious candidate for archival. ? odern email can include @ " ? >, GavaScript and 4SS, as well as who+)nows+what attachments. "he problem of archivin! email, therefore, is the same /holy !rail0 problem of reliable automated conversion of arbitrary files.

( f course, these details are of little interest to customers. #very vendor a!rees with 4allas's DrCmmer, a board member of the [PD\\*2 . 4ompetence 4enter](#), who says that what's hi!h on their a!enda is simply I!ust ma)e it wor), don't ma)e me thin)I.

## The t #w r%"l w #ractices inhibitin\$ PDF/A c m#liance

, ccordin! to DrCmmer the bi!!est sin!le problem is a lac) of document policies. If an or!ani8ation specifies that only -& or -1 formats or applications will be used in+house, for e9ample, this sin!le fact can immensely simplify the PD\*2 , challen!e. But as he says, /if one accepts all )inds of documents, one accepts all )inds of trouble.0

4omin! a close second is the failure to embed fonts. PD\*2 , re uires that fonts used in the document be embedded 7 and too often, they aren't. \*i9in! this problem isn't trivial 7 the choices involve identifyin! and correctly embeddin! the font :if available;, or else replacin! the font with a loo)+ali)e. #ither approach can cause problems, but no worse than the problems of convertin! to "I\*\* or fiche.

>oo)in! the problem ri!ht in the eye remains a )ey challen!e for many or!ani8ations. /PD\*2, is directed at avoidin! the loss of information,0 says PD\*lib President and PD\*2, "echnical Wor)in! Broup chair "homas ?er8. /In a sense, conversion to PD\*2, provides a benchmar) 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 %& years... so it ma)es sense to do it today.0 ?er8 says.

## Why are the &!r #eans leadin\$ the ' ( n PDF/A?

"I\*\* remains the dominant di!ital archival document format in the \$\$, but PD\*2, has already proven persuasive in #urope. "he dense interrelated profusion of super+national :#\$ level;, national, re!ional and local !overnments contributes to a drive towards the most broadly capable document archivin! technolo!y, and in #urope, that technolo!y is PD\*2, .

In contrast to the #uropean view, in the \$\$, PD\*2, is often seen as a cost, not a profit !enerator. "his view may be chan!in!, however. Scalable search technolo!y means corporate archives are increasin!ly seen as treasure+troves of information. If the ris)s of data+loss in conversion are more+or+less e uivalent between PD\*, "I\*\* and electronic+source fiche, the superiority of PD\* over "I\*\*s 7 searchability, navi!ability, hi!h fidelity and an effective unification of paper and electronic sources in a sin!le format 7 become obvious.

DrCmmer says that for his #\$ customers /...nobody will blame you for choosin! PD\*2, ... but you mi!ht be blamed for any other choice.0 In #urope, it seems, !oin! with PD\*2, today is li)e buyin! IB ? ' & years a!o.

## Where is PDF/A \$ in\$ "r m here?

While !overnment archivists were amon! the first to as) the sorts of uestions that eventually led to PD\*2, , awareness of the need to ensure that today's documents are future+proofed has !rown to include heavily+re!ulated industries, ma!or corporations, law+firms, and others with an interest in assured lon!evity.

In the \$nited States, corporate interest in PD\* is led from the pharmaceutical, ban)in! and financial sectors, accordin! to the leadin! 'rd party \$\$+based PD\*2, solutions providers, , pa!o and 4 5 ISI ( 6 "echnolo!ies. ,dobe's 3osenthol cites the e9ample of ,irbus, who !ave a presentation about their usa!e of PD\*2, at the -<sup>st</sup> [International PD\\*2, conference](#) in ,msterdam last , pril.

"here's little uestion that !overnments and businesses are ready for an archival format that improves on "I\*\*. Is PD\*2, that format? It's hard to thin) of a better option, even in theory. PD\* facilitates the unification of hetero!eneous paper and electronic document sources in a sin!le, standardi8ed format, alon! with the relevant semantics. Serious lon!+term stora!e depends on such attributes, and PD\* can deliver.

"he ne9t few years will see more vendors provide /Save as PD\*2, 0 functionality from within their applications 7 a crucial step to address the problems raised by the wide variety of formats. We can e9pect more and better batch+archivin! tools as well.

?ana!ers should e9pect to see PD\*2, on the a!enda in I" and document+mana!ement and retention policy meetin!s in the near future.

A partial list of organizations in the PDF/A space

[ISO: Publishers of PDF 2, 3, 4](#)

[Adobe Systems](#)

[ISO/TC 461 Committee](#)

[The PDF 2, 4 Competence Center](#)

[Paoli, Inc.](#)

[Callas Software, Bmb@](#)

[45 ISO 6 Technologies](#)

[Datalogics](#)

[PDFlib, Bmb@](#)

[PDF Tools, B](#)

[OpenOffice.org](#)