PDF/A for Digital-Born Documents – Archiving MS-Office Documents, E-Mails and Websites

Dr. Hans Bärfuss,

CEO of PDF Tools AG and Vice-Chairman of the PDF/A Competence Center

1 Introduction

When compared with the preservation of data in its original format, there are many advantages to archiving documents and data from digital sources into PDF/A. The source applications are rapidly being developed further. As a result of this, after only a few years, the readability and the authentic display of data can no longer be guaranteed. Furthermore, a company must maintain all of the applications that are used and all of the platforms on which they operate. This incurs considerable costs. Even for documents and files that are created digitally, PDF/A is an excellent choice for long-term archiving and comes with great advantages with regard to uniformity, searchability and cost-effectiveness.

2 Development of digital documents as archive materials

The ECM model from AIIM distinguishes between five major processes in the management of business information: Capture, manage, deliver, preserve and store the documents. These processes can be easily assigned to the following PDF/A functions:

The ECM model from AIIM and the associated PDF/A functions

Digital documents are created in all of the mentioned processes and PDF/A is also important in all of these processes, although in different ways, as explained in the following.

What are the typical sources of digital documents that are later archived, and in which processes do these emerge?

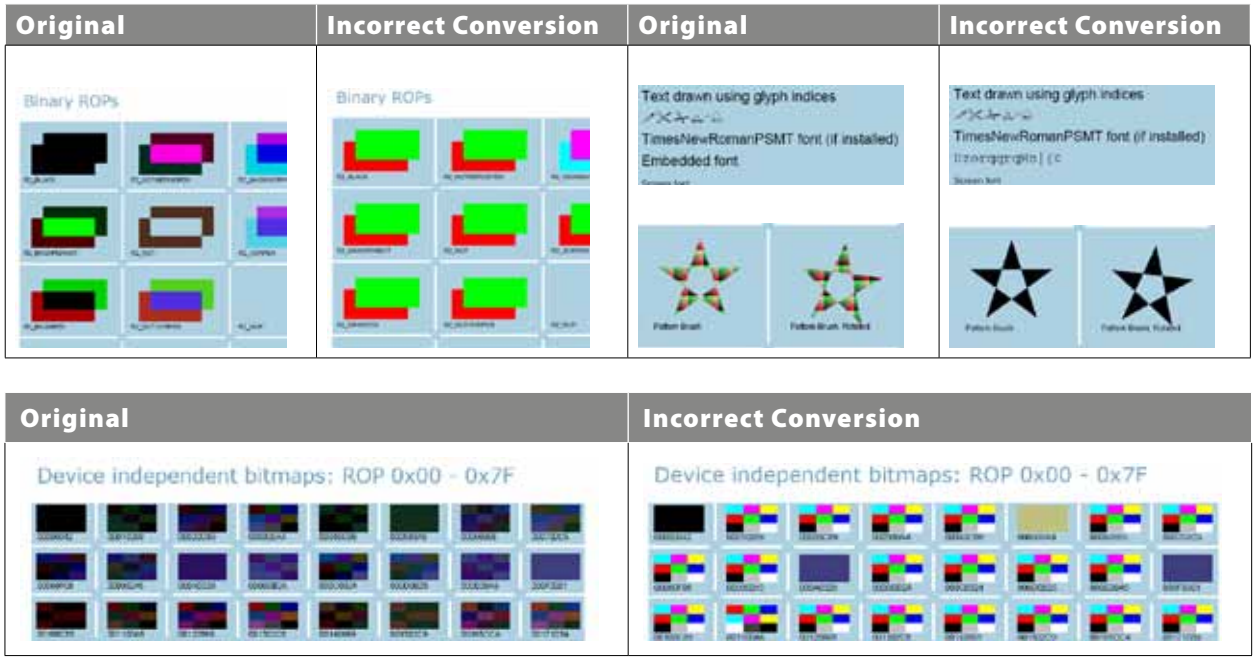
Process AIIM ECM Model	Use case	Applications/Examples
Capture	Inbox	- Scans with or without OCR - E-mails with or without attachments
Manage	Office, graphics and construction	- MS Word, Excel, PowerPoint, Visio, etc. - Illustrator, Indesign, Photoshop, etc. - CAD: Autocad, 3D Studio Max, etc.
Deliver	Outbox	- Print data streams: PostScript, PCL, AFP, etc.
Preserve	Archive migrations	- Masses of TIFF and other files, including source data (metadata, object relationships, etc.)
Deliver/Capture	Electronic data exchange	- SWIFT, EDIFACT, etc.

3 Attributes of analog and digital sources

Digital documents can emerge from analog and digital sources. Some parameters are relevant for their subsequent long-term archiving:

Attribute	Analog	Digital
Sources	Scanner, raster images	Standard and proprietary formats from applications and data streams, in file storage, mailboxes and attachments
Quality of the source	Good	Large differences
Complexity of the source	Low	Can be very high
Product differentiation	Compression rate, performance	Quality
Biggest challenge	OCR recognition rate	Loss of information during the conversion

From these differences, it is clear that we require different strategies for handling different sources, both in the general outline and in detail. These strategies are required both for the employees of IT departments, the records manager and for manufacturers of conversion products. The challenge here lies not only in creating a document that conforms to the PDF/A standard but in interpreting the source in such a way that the visual appearance corresponds to the original document. The following diagram shows the results of conversions to PDF/A whose form conforms to the standard, but whose visual appearance does not sufficiently correspond to that of the source:



Correct and incorrect conversions: In both cases, the result was a document that conforms to PDF/A, but, in the case of an incorrect conversion, does not correspond to the original document in any way.

4 Converting digital sources to PDF/A

4.1 Why convert?

Long-term archiving of digital data to PDF/A offers great advantages:

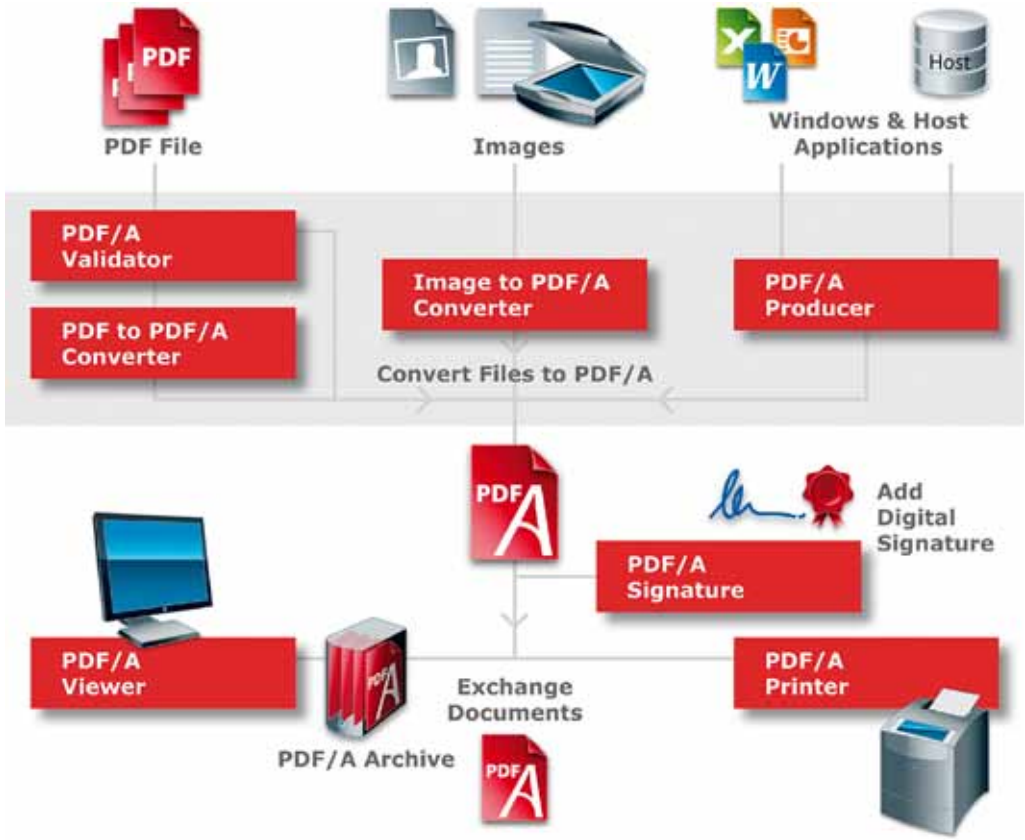
- The user does not have to maintain the original "native" applications and the platforms on which the applications operate.
- Users depend less on software manufacturers because all of the relevant information is saved in one ISO-standardized format and this format is manufacturer-independent.
- Simplified processing due to the fact that the archived data is standardized into one format.
- Option to perform a full-text search in all of the stored data.

These advantages also involve an economic benefit that must not be underestimated.

Of course, when compared to the native formats, archiving in PDF/A also has a few disadvantages, for example, due to the loss of interactivity or the built-in "functionality" of the native format. MS Excel should be used as an example here. MS Excel offers calculation formulas for content and these are lost during the conversion. Therefore, for these formats, it always makes sense to also archive the original document and to use the archiving in PDF/A as a fallback variant. With "interactive" files, the time for archiving can be chosen so that there is hardly any need for further changes (Document Lifecycle Management). In certain formats (for example, e-mails), the original document may have to be saved due to compliance reasons.

4.2 Overview Development and conversion processes

In the following complete overview, the development of digital documents (above) is particularly relevant:



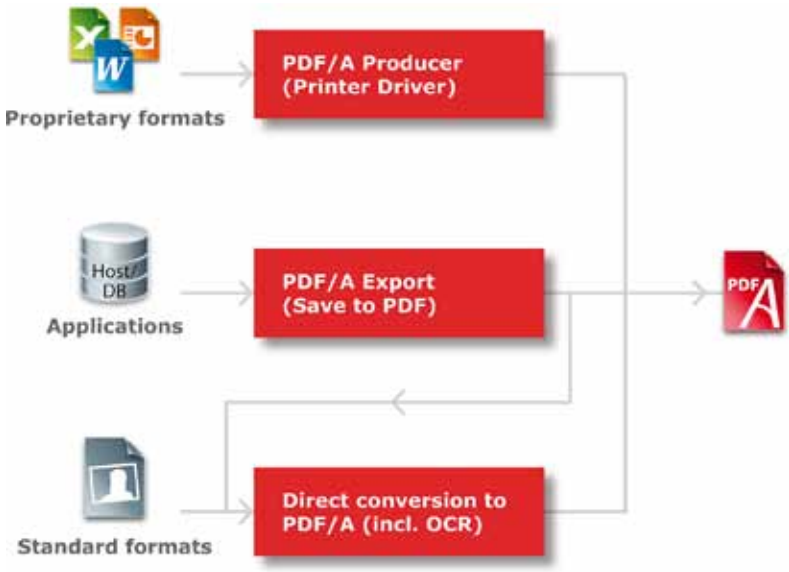
Complete overview of the PDF/A processes, with particular focus on the development of digital documents.

The easiest way to create PDF/A from proprietary formats such as Office documents, CAD drawings, etc. is to use an effective printer driver, also known as PDF Producer, PDF Creator or PDF Converter (for example, Adobe Distiller etc.). This "detour" via a printer driver is required because, so far, most native applications do not have a Save to PDF function. This function is now available for MS Office 2007 but it must be downloaded as a separate add-in.

The process of archiving e-mails, including attachments, to PDF/A (for example, from MS Outlook) is more complex. There are currently only a few providers of this type of functionality, for example, PDF Tools AG with their 3-Heights Document Converter Service, which converts an e-mail and its attachments into a single PDF/A document.

From databases, ERP systems, etc., PDF/A is usually controlled using an export function (Save to PDF). Often, these files must be post-processed because they do not completely conform to the standard. Another option here is the direct, programmatic creation of PDF and PDF/A files. In this process, the contents from any sources can be merged, for example, for processing personalized printed materials. PDFLib GmbH is one of the leading providers of these tools.

Specific tools are usually used to convert images and, in this process, an OCR function is important for the creation of metadata and for the searchability of the texts. In spite of this, even in scanned documents, we cannot underestimate the complexity of such applications, particularly in the areas of multiple formats (for example, dozens of variants of TIFF), colours, fonts and compression and segmenting procedures (for example, Mixed Raster Content). LuraTech offers the leading products in this area.



Converting digital sources to PDF/A using various conversion procedures

All conversion software in all of the areas must take into account the specific obligations and prohibitions from PDF/A, for example, the embedding of fonts, colour profiles and metadata (as XMP).

4.3 General Challenges

From a general perspective, when creating PDF/A from digital sources, we are confronted with the following challenges:

Area	Challenge
Colours	If the colour profiles from the sources are missing, assumptions are made about the colour space.
Fonts	If fonts (or glyphs) are missing, replacement fonts must be selected. To do this, the text must be a Unicode text.
Transparency	The flattening of transparency is complex and may lead to the loss of information (fonts, vectors, etc.)
Levels, interactive and multimedia elements	Only the "Print Preview" is retained
Actions	Functionality (JavaScripts etc.) is lost
Digital signatures	Check, document, sign again

4.4 Converting e-mails

An e-mail can contain all types of documents, interlaced archives and much more (executable files etc.). In addition, the e-mail can contain internal or external references (e.g., HTML mails) and different systems, interfaces, file systems and data streams are involved. The process of archiving e-mails, including attachments, is therefore effectively the "supreme discipline" of archiving in PDF/A, since all of the challenges in connection with converting sources that were originally analog or digital must be solved using one single product.

To solve this, a different conversion strategy must be selected for each individual element of an e-mail: The e-mail body and attachments are converted individually and, only then, are merged into a single document. In this PDF/A document, each attachment can then be identified using a so-called bookmark entry. By doing this, the structure of the e-mails can also still be traced at a later point. In addition, information, such as tables of contents from Word documents, is not lost, because these are mapped as a second level of hierarchy in the bookmarks and are linked accordingly in the PDF/A. Even the handling of digital signatures poses a challenge when archiving e-mails.

4.5 Converting websites

The topic of archiving websites is relatively new. This basically involves retaining the contents and state of one's own website in a way that is legally trustworthy so that the required evidence can be provided in legal or other procedures.

The difficulty when archiving websites is that the output using a print driver does not normally represent the authentic appearance of the website, because websites are usually specially prepared for printing. To be able to bring forward trustworthy evidence, this "true to the original" is crucially important.

Therefore, from the website, a "Capture" function is used to create an image that is merged with the relevant text and other information (fonts, colour spaces, etc.) to effectively produce a "vectorized, searchable screenshot". Another complex issue is the handling of external links and the internal link structure of a website. In addition, it is necessary to decide on one browser and one browser version because different browsers and browser versions display websites differently.

4.6 Converting on the client or on the server

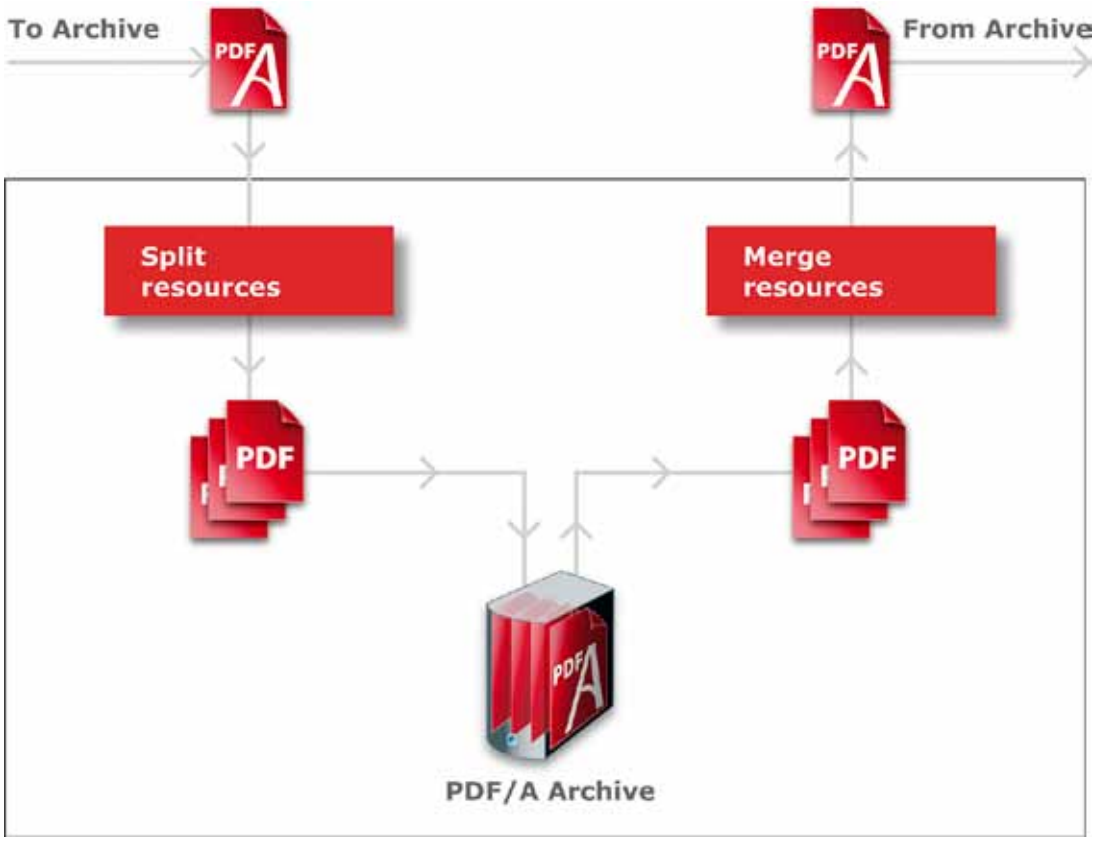
We must consider the following aspects with regard to the question of whether conversion software should be installed on individual clients or on a central server:

Attribute	Client	Server
Scaling workstations	Small amount	Large amount
Distribution	Complex	Simple
Robustness for the users	Depends on the creator-application	Independent
Performance for the users	Restricted by the client	Scalable
Supported source formats	Restricted by the installation	Scalable
Application support	Local	Central

4.7 Font handling in mass archiving

Single, individual PDF/A documents can be directly archived. When archiving large quantities of similar PDF/A documents (for example, telecom invoices etc.), the situation often arises in which the documents contain the same fonts, logos or other corporate identity elements that must also be archived for each individual document. The repeated saving of collective resources (fonts, images) is undesirable and reduces the acceptance of PDF/A.

To solve this, the archive system can be upgraded using an add-in that separates the shared resources and saves them in only one instance for all documents when performing mass archiving of PDF/A documents. When a document is accessed, the shared resources are again merged with the document to produce a complete PDF/A document. This procedure can also be used for digitally-signed documents, but, during the signing process, the document must already be prepared for the separation of resources.



Concept for preventing superfluous saving of resources (e.g., fonts) in mass archiving.

4.8 Legal security with digital signature

The process of digitally signing PDF/A files that derive from digitally-created documents brings greater legal security. Depending on the application, the user must be clear about what the signature really provides. In any case, in a qualified electronic signature, it is absolutely clear at what time the conversion occurred and whether the document has been changed since the conversion. It is also clear who performed the conversion process in a company.

However, the uncertainty that arises from the "dynamic" source (e.g., a database) of such a PDF/A document cannot be dispelled. Nor is it possible to verify whether the created PDF/A document actually corresponds to the appearance of the original document (e.g., a Word document) or whether all of the information that is contained in the document (e.g., contents and e-mail attachments) actually exists in the PDF/A file. To increase the credibility of such documents, the entire process must be certified. This is therefore a topic that transcends the simple use of digital signatures. However, such certifications require a certain volume of data so that this is worthwhile for service providers, manufacturers of software and systems and large companies.

4.9 Quality assurance by validators

"Trust is good, control is better": This, of course, also applies for PDF/A documents and products that create PDF/A. Or that claim to create PDF/A. Not all the products that are labelled as PDF/A are actually PDF/A products. In extreme cases, the archiving of company data can be crucial for the existence of a company. For example, in a court case, if the exonerative records have not been prepared or have not been prepared correctly.

It is therefore important to use tools that ensure the highest standards of quality. Validators exist to determine if a tool fulfils this prerequisite. These validators also need to be checked. For this task, the PDF/A Competence Center created a freely-available test suite that systematically breaches the standard and then checks that a validator can identify all of the breaches.

The use of a validator is not only important when evaluating a tool, but it is also important in the operational processes. A validator should therefore be used regularly to check the conformity of the created PDF/A documents – as a permanent quality check. This is because different sources, application versions, etc. may lead to different conversion results.

5 Summary

PDF/A is beneficial as a format for archiving digital documents and can lead to considerable cost savings in comparison to archiving in the native format. However, the devil is in the details with this and we must not underestimate the complexity that arises depending on the source of the digital documents. It is therefore essential to collaborate with specialists in this area and this collaboration can protect users from unnecessary costs accrued through incorrect processes etc.

For both day-to-day business and from a strategic point of view (e.g., in legal cases), it is very important that information can be accessed quickly and securely. Discrepancies in this area can result in damage to a company's image or in substantial financial consequences. Processes for archiving directly from digital data are therefore given top priority.