Digital born PDF/A – tough nut or (un)recognized potential?

PDF/A archiving, document formats, conversion, validation

Scanned documents have been archived successfully in PDF/A format for more than six years. However, the attitude towards archiving digitally created documents is more reserved. What are the reasons? Some are obvious: scanned documents are easier to convert into PDF/A format, whilst converting digitally created documents usually represents a technical challenge. Slightly less obvious are errors in the reproduction of the converted document, functional limitations of the PDF/A standard, and other reasons. However, these challenges can be overcome with the right strategies.

A large proportion of archived electronic material consists of scanned documents such as business correspondence, book-keeping records, contracts, paper-based archives and documents worth keeping that should be migrated in an electronic format. The number of electronically created documents is, however, quickly catching up; they are usually invoices generated by ERP systems, emails, office documents in the outbox and a more special type of document such as design drawings from CAD systems.

Reproduction fidelity – technical challenge for scanned documents

The fact is: scanned documents are largely raster images. For years, it was perfectly acceptable to file them as TIFF images, usually in black and white, to save memory space. However, requirements have become more demanding. Today, the ISO standard PDF/A has made colour, metadata and full text searching a matter of course without requiring significantly more memory. The technical challenges in relation to these raster images concentrate on image analysis and processing. For instance:

- Images are processed with a text recognition machine (OCR). Empty pages are detected, the image is straightened and any smears removed. This is followed by the recognition of text and barcodes.
- Segmentation and compression: the colour image of every page is broken down into its individual components such as background, text and photos. These components are reduced in size by processing them with specifically designed compression methods. This Mixed Raster Content (MRC) method makes it possible for colour documents to reach file sizes that can compete with black-and-white files.

Software manufacturers had learned how to control this process before the era of PDF/A. However, PDF/A returns a standardized result in contrast to TIFF. As a subset of PDF, PDF/A can do a lot more. Its color spaces, fonts, vectors, fill patterns and transparency mixes make PDF one of the most powerful 2D graphic models; as such it is predestined for the reproduction of digitally created documents. All one has to do is to convert the digital source to PDF/A. However, this step is a greater technical challenge than might appear at first glance.

Firstly, there is the large number of document formats that need to be converted: ASCII texts, Word, Excel, PowerPoint, PDF, emails, HTML and XML from various sources such as file directories, ZIP archives, mailboxes, file attachments and data flows from applications. Additionally, the quality of these digital sources rarely reaches that of raster images. Files are often either damaged during transmission or poorly created in the first place. This is particularly often the case with PDF files created using freeware. The "bad PDF" problem is the cause of high costs not just for software producers, and is often the reason for problems affecting document-based business processes.

The greatest challenge facing the conversion of documents from digital sources to PDF/A is, however, reproduction fidelity. Even if the converted file formally complies with the ISO standard, it can still happen that the visual result does not correspond to the original. These kinds of imaging

www.pdf-tools.com

Dr. Hans Bärfuss is founder and managing director of PDF Tools AG and delegate of the Swiss Association for Standardization (SNV) to the ISO. He is also an initiator and founding member of the PDF Association and is currently chairman of the Swiss Chapter. PDF Tools AG is a manufacturer of software solutions and programming components for generating, processing, reproducing and archiving PDF and PDF/A files.

DOK.magazin-2013-3-Digital-born-PDFA-EN

errors can have many causes. It is usually because the source documents have complex graphic elements such as fill patterns or transparency and the conversion software is unable to map every graphic function or all of the possible combinations in PDF/A. The many virtual print drivers that are used to create PDF/A files via the Print function are prime examples. The majority of these drivers are based on the PostScript driver provided together with the operating system that actually only implements a part of the defined graphic interface.

Strategies for error-free PDF/A documents

Today, it is no longer a question of principle: PDF/A is suitable as a long-term storage format for both scanned and digitally created documents. Users are, however, still cautiously reserved due to the technical difficulties affecting the conversion of digital sources to PDF/A. Nonetheless, these challenges can be overcome. The choice of conversion software plays an important role – but the choice of the right system architecture is the determining success factor.

It has proven to be beneficial if scanned images are converted into a searchable, possibly digitally signed document with metadata using special software (scan server). All of the steps in the process are optimally aligned with each other. It is important that the scanner provides the raw image only to enable the best possible compression. The result is usually less than ideal if processing is distributed between the scanner, the scanning computer and the server. There are various methods of converting digitally created documents to PDF/A at a professional level. The easiest of them all is to create the document – quotes, for instance, or invoices and reports – in PDF/A format. All that is then needed is a tool (PDF/A Validator) to check whether the document complies with the rules of the standard.

If the document is not in PDF/A format, it will need to be converted. In the best case scenario, the native application, for instance a product from the Microsoft Office range, will incorporate the appropriate function ("Save as PDF/A"). Experience shows, however, that these functions are affected by reproduction errors and minor non-compliances with the PDF/A standard. A tried and tested strategy is therefore to use the less precarious function for directly creating a normal PDF file ("Save as PDF"). The result is subsequently converted to PDF/A using a specialized converter.

The print function route is often the only option in the absence of a direct function for PDF/A creation. The document is "printed" as a PDF/A file via a virtual print driver. In this case, it is recommendable to use a specially developed PDF/A print driver to avoid the reproduction errors that occur with conventional, PostScript-based PDF print drivers.

Central PDF/A conversion – the reliable method

To put it in a nutshell: a central PDF/A conversion solution for both scanned and digitally created documents is worthwhile even for just a small number of users. The reasons are simple:

- **Quality:** the server's protected runtime environment ensures that every step of the conversion process is always carried out in exactly the same way with the tools selected for the best results.
- **Supported formats:** central solutions can support a wide range of document formats, including formats for which the corresponding software is not installed on the client. This saves the costly roll-out of software on workstations.
- **Robustness and stability:** the applications for conversion are run in an automated and controlled runtime environment. This makes it possible to ensure that the conversion service is always reliably available. The server monitors the correct functioning of each application and automatically restarts them in the event of a problem.
- **Validation:** the server checks the created data for conformity with the standard. Additionally, the server can run an automatic image comparison as an extra assurance feature to rule out any reproduction errors.
- **Scalability:** conversion servers can be scaled by multiprocessor machines or by way of distribution across a number of machines.
- **Centralization:** a centrally managed server and lean clients with less software help cut operating costs.

All in all, these are convincing arguments in favor of PDF/A conversion with professional tools.

Conclusion

The PDF/A standard will be continuously improved and adapted to meet new requirements. The conversion of scanned and digitally created documents to PDF/A will enable many companies in different segments to comply with the growing demands for secure digital archiving whilst ensuring that documents remain accessible at all times in the long term. The application of proven strategies will enable the implementation of a successful digital archiving project that fully satisfies every technical, legal and economic aspect.