PDF Tools: Document digitization in PDF/A

How to make a success of complex projects

Public and private enterprises like to keep up with the times; they launch projects and ride on the crest of the digitization wave. Infrastructures for centralizing archives and worldwide online research are created to raise productivity and lower costs, as is always the case with these projects. But do we actually have digitization under control? Are we not creating new risks? What do we need to know to prevent projects from turning into nightmares? A contribution by Dr. Hans Bärfuss, CEO of PDF Tools AG.

Dr. Hans Bärfuss, founder and CEO of PDF Tools AG:

„The question is how to combine all the different needs of an enterprise to create a single uniform scan strategy."

Suitable scanners and corresponding software are prerequisite to prevent digitization projects from turning into nightmares. A good consultant to define the ideal workflow would also be advantageous. But basic knowledge of the digitization process is doubtlessly helpful when it comes to making projects a success. This article provides an insight into some aspects of specialized digitization software and is intended to facilitate its selection and use in concrete projects.

What does scan software do?

Scan software carries out a number of steps along the path from a paper-based to archivable document, independently of architecture and scope; some of these steps are optional, whilst others remain invisible to the user.

Image acquisition: The scanner creates a black-and-white or color raster image of the scanned paper and hands it over to the scan software via a TWAIN, ISIS or FAX interface. The format and resolution of the raster image are selected at this point. Documents received by fax hardly differ from scanned documents and can usually be processed using the same software.

Automatic image processing: Images can be prepared for a quality inspection: blotches and empty pages are removed and the brightness and contrast adjusted to achieve optimum legibility, to name but a few of the steps in this process.

Quality inspection: The scan operator can carry out a visual inspection, intervene where necessary and repeat the scanning process for individual pages or the complete batch. Simple classification data such as the batch number are often entered at this point (operator workstation).

Text recognition and barcodes: The conditioned images are now processed by OCR software (OCR = Optical Character Recognition). The pages are first rotated to the reading direction, after which the text and barcodes are recognized and allocated to the images.

Classification: The text and barcodes recognized by the software can be used to classify the document. It can differentiate between invoices, delivery notes and other transaction documents, for instance, or assign a tax declaration to the declaring person. This step in the process can be carried out manually (index workstation) if automatic classification is partially or entirely impossible.

Metadata input (indexing): Information from the manual classification of barcodes and other sources is summarized as metadata (index data) and assigned to the documents.

Segmentation and compression: The memory space requirements for scanned raw image data are considerable (45 MB for one A4 page in color with 400 dpi). Efficient compression processes significantly reduce the amount of data (to around 200 kB). Additionally, a special process known as MRC (Mixed Raster Content) can reduce the data further still (to around 20 kB). This process is based on segmentation: splitting the image into individual components such as background, text and photos.

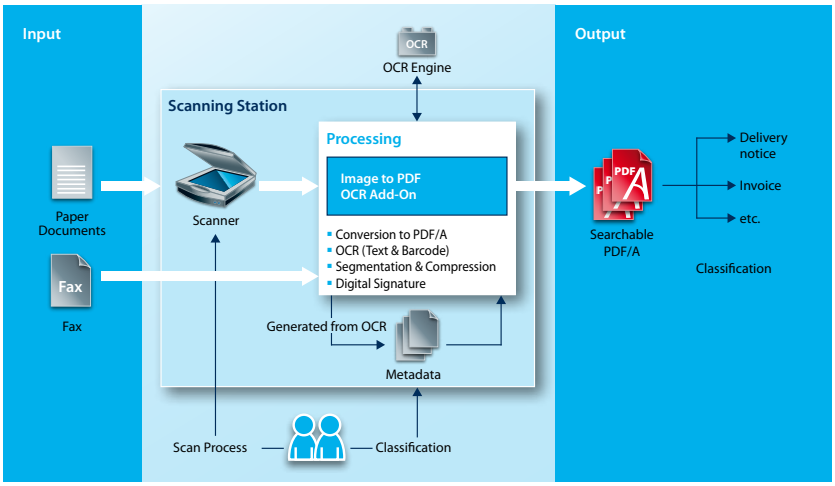
PDF/A generation: The processed and compressed images of each page, the recognized text and the metadata are combined with the scanner's color characterization (ICC color profile) to generate a PDF/A document. Metadata is often sub jected to additional separate processing (index file).

Digital signature: A digital signature can be applied to ensure the legal comprehensibility of the document's condition at the time of receipt.
Validation: The conformity of the generated document with the PDF/A standard and the validity of the digital signature can be verified and the results documented in a log.

The product of digitization: PDF/A

PDF/A is an ISO standard for the use of the PDF format in the long-term storage of electronic documents. It was first published on October 1, 2005 as ISO-19005. The PDF/A standard defines "a file format based on PDF called PDF/A that offers a mechanism that represents electronic documents such that the visual appearance remains preserved for an extended period, independent of tools and systems for producing, saving and reproducing it." The PDF/A standard is not a new format, but rather defines the requirements that documents created on the basis of the PDF format need to fulfill for reliable long-term storage. Parts 2 and 3 of the standard have since been published to ensure the format stays abreast of developments.
Doesn't the popular TIFF format offer the same features? Yes, at first glance. Both formats can store scanned raster images. However, PDF/A is the more up-to-date format and offers numerous advantages. The most important are:

- Efficient compression processes;
- Standardized process for full text searching;
- Standardized digital signatures (PAdES: PDF Advanced Electronic Signature) can be embedded in the document to protect its integrity;



High-performance scanners are the preferred choice for centralizing functions such as image processing, segmentation and compression, PDF/A generation, etc.

- Metadata are stored in a standardized format (XMP: Extensible Metadata Platform) and embedded in the document;
- As a universal document format, PDF/A is ideal not only for scanned but also for electronically generated documents.

Architecture: Local or central?

The choice of architecture depends greatly on the type, scope and regularity of processing. A simple multi-function scanner with integrated scan software is sufficient for occasional personal use. These circumstances hardly call for comprehensive digitization projects. The question is rather how to combine all the different needs of an enterprise to create a uniform scan strategy. Multi-functional devices (MFP) located in each department cater for the personal needs of employees, whilst scanning lanes with batch scanners in service centers regularly process high document quantities.
The specialized software for each scanner is usually installed locally, often as a part of the device itself. This could be a reason for the growing popularity of multi-functional devices. However, local solutions are not as popular with regard to high-performance scanners because they are expensive and their decentralized architecture can slow down processing. Hence the centralization of complex and costly scan software functions to counteract these problems. The functions are often split as follows:

- Local: image acquisition, image processing and quality control;
- Central: character recognition, segmentation and compression, PDF/A generation, digital signature and validation.

This distribution increases the scalability of the architecture, which in turn results in lower acquisition and operating costs and greater throughput for high document volumes.

PDF Tools AG
Kasernenstrasse 1
8184 Bachenbülach
Switzerland
Tel: +41 43 411 44 51
Fax: +41 43 411 44 55
pdfsales@pdf-tools.com
www.pdf-tools.com