

PDF Tools: High image quality at a low data volume

From scan to information

A central scan server service enables large volumes of paper documents to be quickly and efficiently converted into electronic documents, prepared for processing and stored in a long-term archive. A scan server, such as the 3-Heights Scan to PDF Server offered by PDF Tools AG, converts scanned files and accompanying index files into the standardized PDF/A file format.



Nadine Schuppisser
of PDF Tools AG:

“A central scan server is an efficient, multifaceted solution for the processing of large volumes of scanned documents from various sources.”

Even in the age of electronic invoices, online shops and e-commerce, paper has not yet become obsolete: documents such as invoices, tax forms, service reports and contracts are still prepared on paper, sent in the post and received in one's letterbox. Once the paper documents reach the company or agency, IT systems are responsible for processing the information – everything on paper has to be scanned, prepared in a format that is machine-readable, stored and archived. Documents are usually scanned in the individual departments using multifunctional devices (an MFP with additional printing and fax function) or centrally using a high-performance scanner.

For the majority of companies, scans accumulate in various locations: at the central office, at scan stations in office departments and on mobile devices, e. g. when visit-

ing clients. Fax messages received are nothing more than an image of scanned information.

From an image to a standardized document

When scanning a document, a facsimile is created as an image file first in raster formats such as TIFF and JPEG. However, a raster document is simply an image without any additional information. Texts and information contained in barcodes must be extracted from the image via text recognition (OCR – optical character recognition) after it has been scanned. Ideally, the text and image are then saved together in the same document. This makes data storage simpler and preserves both the appearance and information contained in the original document.

PDF/A has established itself as a standardized storage format for the long-term archiving of scanned and electronically generated documents. The PDF/A standard supports the storage of image and text information in the same document. The documents can be perused using the full-text search feature. PDF/A uses a powerful compression technique for the image information, thereby significantly reducing the original file size without losing any information. This is especially

important if the document contains color images in addition to grey-scale images and the color information is intended for further use. PDF/A also permits metadata such as classification information to be saved directly in the document. XMP (extensible metadata platform) is used for this – as with PDF/A, it is defined as its own ISO standard. PDF/A also has a digital signature option to guarantee the authenticity of the documents and integrity of the contents. Overall, PDF/A offers the security of an international document standard that guarantees long-term stability and features an exhaustive range of functions.

Scan locally, process centrally

Scanning places little demand on hardware and software in terms of performance. In principal, scans could be carried out using a simple digital camera. The steps that follow, however, require much more computer processing power and intelligence – image compression, OCR and conversion to PDF/A require time and effort. Above all, there are two opposing needs to consider: reliable text recognition requires the highest possible image quality. This increases the amount of space required for storage. Of course, the aim is to keep data vol-

PDF Tools AG

PDF Tools AG is a producer of software solutions and programming components for generating, processing, reproducing and archiving PDF and PDF/A files.

umes to a minimum when storing files. Software that caters for both requirements places great demand on the computer processing power, especially when a large volume of scanned documents needs to be processed.

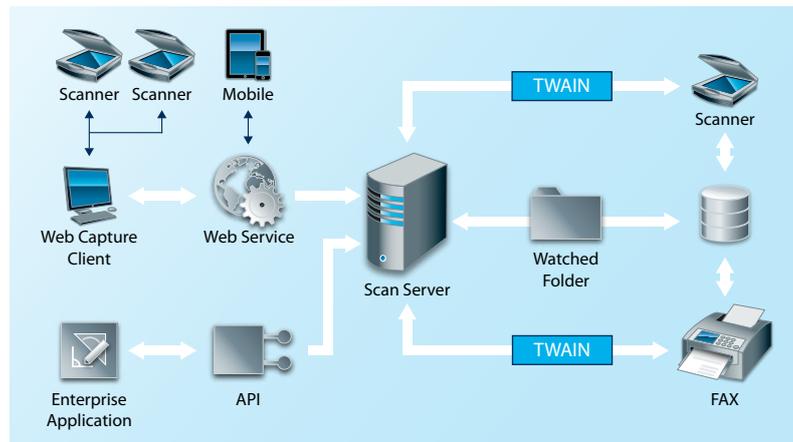
Another aspect to consider is that information from other workstations and different IT systems is required for embedding index data, classification data and other meta-data and digital signatures. The decentralized data must be combined to create the PDF/A document.

Performance-intensive operations

The solution for both problems is a central scan server – one example is the 3-Heights Scan to PDF Server by PDF Tools AG. This server receives the scanned image files, analyzes the documents and generates a PDF/A document with all text and image information compressed to the right size. The document can also be tagged with a time stamp or a digital signature. The consolidated information is now available in a standardized, high-quality format that is suitable for human readers and for automated processing with IT applications.

A central scan server also simplifies software distribution and maintenance. Comprehensive scan software with integrated OCR function does not have to be individually rolled out, configured and maintained at the scan stations. An elemental operator application is sufficient for image acquisition.

Problems encountered during more complex processing steps do not have to be individually resolved at the respective workstation. The scan server service instead utilizes test infrastructure to analyze all problems and rectify any errors. The service is then transferred into pro-



The scan server operates as a central PDF/A-conditioning instance and processed scans from different sources.

ductive operation. To ensure that the scan server is tailored to the respective environment and can, if required, be scaled when shared among more than one computer, the 3-Heights Scan to PDF Server distributes the tasks over multiple sub-systems:

- The scan server receives jobs for conversion into PDF/A format, delegates text recognition responsibility to the OCR server, and combines the OCR results, scanned image and meta data into a complete PDF/A document.
- The OCR server receives jobs from the scan server for text and barcode recognition, prepares the image information through processes such as straightening texts and removing flaws to provide the best possible conditions for identifying the text, divides the document into text, barcode and image fields, and carries out the text recognition process.

The server offers two additional services for locally generated scans: a watched folder service transfers all files stored in certain directories to the scan server for automatic processing. The scan server utilizes a web service to receive jobs created via a web-based application and

then sends the converted documents back to the job provider. The scan server can also take on other useful jobs, including validating the generated PDF/A documents for conformity with the ISO standard, tagging the documents with a watermark and combining individual documents belonging to the same business case into an overall document.

A central scan server is an efficient, multifaceted solution for the processing of large volumes of scanned documents from various sources. It converts the scanned image data into standardized, searchable PDF/A documents that are packed with information, alleviates the amount of work to be carried out by scan stations when processing information, supports the integration of other IT systems, and helps to maintain a consistent, company-wide document standard.

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