A PDF/A Solution for Scanned Documents

PDF-TOOLS.COM
Premium PDF Technology

A PDF/A Solution for Scanned Documents

Scanning paper documents has become a daily ritual in the mail receiving room of many businesses. This task is often performed by a third-party scanning service provider. In most cases the scanned images are saved as black & white TIFF files, the format synonymous with faxes. In special cases, for example checks, identification papers with photos etc., the documents are scanned to color files. One must be cautious, however, since colored TIFF files can quickly become extremely large.

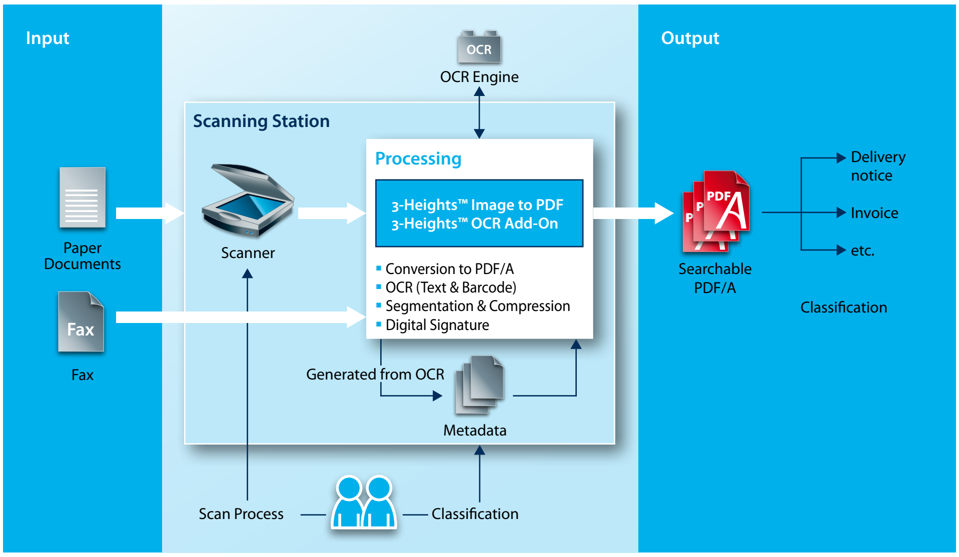
The PDF/A standard has now also established itself in incoming mail applications, especially when dealing with color scans. However, individual processing steps like text recognition, compression and digital signatures are generally not optimized with one another or integrated into one single solution. There are, for example, scanners that can create PDF/A files and also sign them. However, the subsequent compression of the file invalidates the digital signature, making it worthless.

Advantages

- Scalability through configurable options
- PDF/A files in color with small file size
- OCR recognition is possible
- Application of digital signatures is possible
- Validation of PDF/A conformance

A New Approach

PDF Tools AG offers a solution for creating PDF/A files from scanned documents and fax images that fulfills all the vital requirements like small file size, searchable files and embedded metadata. The following diagram illustrates the principle.



The workflow is managed in the following manner:

1. **Image capturing:** The scan operator begins the scanning process and generates a TIFF file in color. The scanner usually stores the file in a file folder. A direct interface to the scan software (via TWAIN) is also possible. Facsimile documents are received by the fax machine and automatically stored as black & white TIFF files in a special folder.

2. **Manual classification:** Depending on the process and, if desired, the scan operator can conduct a manual classification, controlling the scanner so that the images are saved in different folders, e.g. for invoices or delivery notices.
3. **Segmentation and compression:** The color image of each page is separated into its different elements, like background, text and pictures. Each individual element is then reduced in size by subjecting it to compression processes specifically designed for that type of element. This process, known as MRC (Mixed Raster Content), makes it possible to achieve competitive file sizes for color documents.
4. **Text recognition and barcodes:** The images are further processed through an OCR engine (Optical Character Recognition). The image is first cleaned up and deskewed (straightened out), then the text and barcode recognition takes place.
5. **Metadata:** Information from the manual classification, recognized barcodes and other sources are gathered and entered as standardized XMP metadata (eXtensible Metadata Platform).
6. **PDF/A creation:** The prepared images of each page, the recognized text and the metadata are assembled together with the ICC color profile of the scanner and saved as a PDF/A file. Optionally, a separate index file can be created containing just the metadata.
7. **Digital signature:** If desired, the PDF/A files can be digitally signed in order to preserve the traceability and revision integrity of the documents.
8. **Validation:** As an additional option, the PDF/A conformance of the created documents and the validity of the digital signatures can be verified.

Advantages

- The solution is scalable from basic up to full functionality. Configurable options are provided for MRC, OCR and digital signatures.
- The MRC process makes the creation of color PDF/A files possible that are of a comparable size to black & white TIFF files (approximately 40 Kbytes).
- The ABBYY FineReader can be applied as an OCR engine. Alternatively, if requirements are more basic, the free OCR recognition software from Tesseract can be used.
- An HSM (Hardware Security Module) from SafeNet can be used for applying large numbers of digital signatures
- The PDF/A conformance of the created files can be verified with a validation protocol.

Installation and Use

The scan server from PDF Tools AG can be installed as a service on a computer with a Windows operating system. Certain process steps like OCR recognition and the application of digital signatures can also be carried out separately and on different computers.

More complex architectures, for example the recognition of text and insertion of metadata after signing, multiple digital signatures etc., are possible within the scope of a defined project.

PDF-TOOLS.COM
Premium PDF Technology

About PDF Tools AG
PDF Tools AG is a world-leading manufacturer of programming components and software solutions for generating, processing, displaying and archiving PDF and PDF/A files. The architecture of this powerful software is geared specifically towards high-performance processing of large volumes of documents. The company's versatile products are especially suitable for supporting corporate work processes and as components for integrators and OEM customers.
Visit www.pdf-tools.com for more information.