WHITE PAPER

3-Heights™ Scan to PDF Server
Basics and Applications

PDF-TOOLS.COM
Premium PDF Technology

The 3-Heights™ product family from
PDF Tools AG stands for:

High Quality – High Volume – High Performance

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

Whitepaper-Scan-Server-EN-20160126

Content

Functions and benefits

Why is a scanner not enough?

In most companies, scanning paper documents has become a routine task when handling incoming mail. Multifunction printers (MFP) or high-performance scanners are used for this purpose, depending on the type and volume of paper documents received. In most cases, the scanned images are created as black-and-white TIFF files, the typical format used by fax machines. In special cases, such as when scanning checks or ID photos, the file is generated in color. However, color scanning is usually avoided, since the created TIFF files are either too large or the JPEG compression visibly reduces the image quality. But good image quality is an important requirement for a good text recognition rate. Achieving good image quality at a high compression rate requires a level of processing power that local multifunction printers do not usually possess. Separate scanning software can offer considerable advantages in this respect.
The PDF/A standard is now widely established in incoming mail applications. The PDF/A standard offers the following important advantages in comparison to conventional document formats, such as TIFF and JPEG:

- **Standardized format:** PDF/A is suitable for storing both scanned and digitally created documents.
- **High compression rate:** The PDF/A standard supports more modern and powerful compression processes, and thus small file sizes for color images.
- **Text recognition:** The created PDF/A documents can be made searchable by embedding text from an OCR engine.
- Embedded metadata: In order for the document and the associated metadata to form an inseparable whole, the metadata is embedded in the file in PDF/A. For saving, PDF/A uses the Extensible Metadata Platform (XMP) format, which, like PDF/A, is also defined as its own ISO standard.
- **Digital signature:** In order to ensure the integrity and authenticity of the created documents, a digital signature can be applied to the PDF/A document in accordance with the PAdES standard. The digital signature is a kind of electronic signature that can serve the same purpose as a handwritten signature, provided that the corresponding legal requirements (national signature laws) are met.

In principle, TIFF documents offer all these advantages, but only as proprietary extensions, since the TIFF standard itself does not offer solutions.

Requirement	TIFF	PDF/A
Long-term readability	+	+
Clear rendering	+	+
Data consistency	Proprietary tags for metadata	+
Authenticity / Integrity	With detached signatures	+
Required storage space	Black/white: + Colour: –	+
Searchability	Proprietary tags for OCR text	+
Long-term experience	+	+

Illustration 1: Advantages of PDF/A over TIFF

Usually, the individual processing stages, such as text recognition, compression, PDF/A generation and digital signature, cannot be performed by the scanner alone, as metadata is often added retroactively by an index station. However, this work stage breaks the seal of the digital signature and makes it worthless. Here, too, separate software can offer a decisive advantage.

What does a central scan server do?

The 3-Heights™ Scan to PDF Server is a central service that converts locally scanned files and associated index files into the standardized PDF/A file format within a company. To this end, the service performs all tasks that can be delegated to it by the local scanning station. The solution is particularly suitable for processing stages that do not require any user interaction or which impair the efficiency of the local scanning station with CPU-intensive functions (OCR, compression).

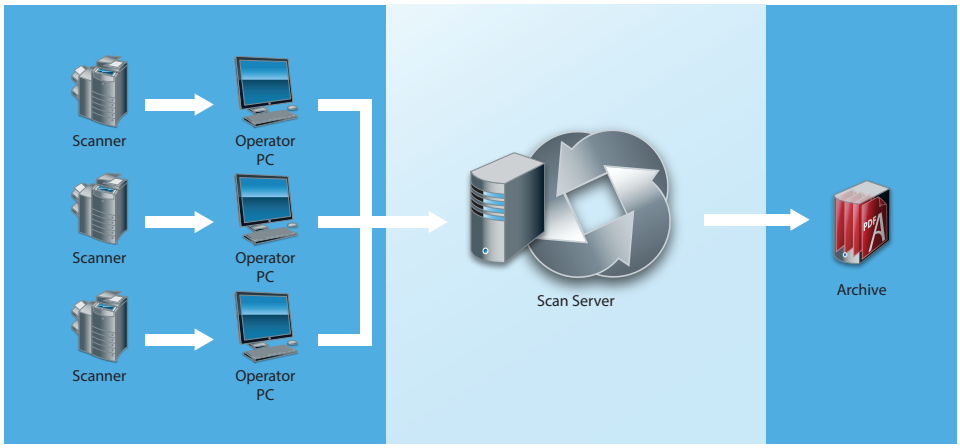


Illustration 2: A central service for creating PDF/A files from scanned documents

The main functions of this service are:

- **Text and barcode recognition:** Scanned image files need to be made searchable. The services can use the 3-Heights™ OCR Service to identify text in an image file and embed it into the converted file in a way that makes it searchable. The recognized barcodes can be used in several ways: in the text search, as part of the embedded metadata, or to control the processing (name of the output file, page separation, etc.) within the service.
- **Compression:** Color images are broken down into several elements. Using the Mixed Raster Content (MRC) process, they are then heavily compressed with no visible losses.
- **Embedding of metadata:** The PDF/A standard requires metadata to be embedded in the document in the form of XMP packets. This function is offered by the service.
- **PDF/A creation:** The service creates single or multi-page output documents in accordance with the ISO 19005 series of standards. All published parts of the standard – PDF/A-1, PDF/A-2 and PDF/A-3 – are supported.
- **Digital signature:** The signature can be advanced or qualified, suitable for long-term storage or simply for exchange. It may also contain a time stamp. Only one time stamp can be applied in place of the personal signature. The service can use a cryptographic infrastructure (USB token, HSM) via a standard interface (PKCS#11) to create a digital signature.

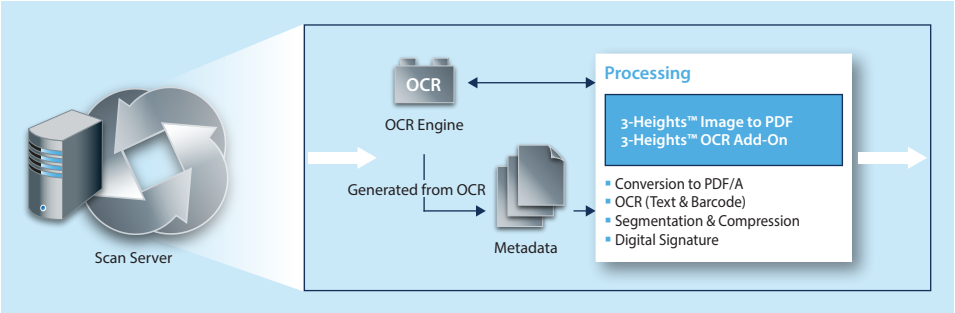
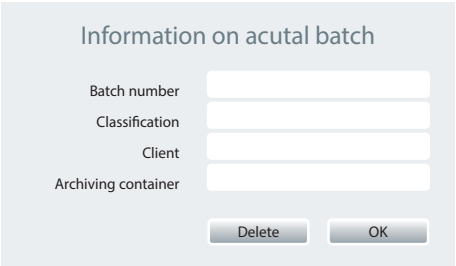


Illustration 3: The main functions of the 3-Heights™ Scan to PDF Server

A typical sequence would look as follows:

- **Image acquisition:** The scan operator starts the scanning process and creates a color TIFF file. The scanner usually stores files in a file folder. Facsimile documents are received by the fax machine and stored in a special folder as black-and-white TIFF files.
- **Manual classification:** Depending on the process, the scan operator can perform a manual classification. They control the scanner so that the images are stored in different folders (e.g. invoices and delivery notes), or special barcode sheets are added that help to separate and classify the documents, or a minimum set of index files is created.



- **Segmentation and compression:** The color image of each page is broken down into its different elements, such as background, text and pictures. The size of the individual elements is then reduced by subjecting them to compression processes specifically designed for that type of element. This MRC process makes it possible to achieve competitive file sizes for color documents.
- **Text and barcode recognition:** The images are processed further by an OCR engine. The image is cleaned up and deskewed, and text and barcode recognition then takes place.
- **Metadata:** Information from the manual classification, recognized barcodes and other sources is assembled into standardized XMP metadata.
- **PDF/A creation:** The prepared images of each page, the recognized text and the metadata are assembled into a PDF/A document together with the ICC color profile of the scanner. Optionally, an index file containing only the metadata can be created.
- **Digital signature:** If desired, the PDF/A files can be digitally signed in order to preserve the traceability and revision integrity of the documents.
- **Validation:** As an additional option, the PDF/A conformance of the created document and the validity of the digital signature can be verified.

The service also offers a range of additional functions (see further down).

Where can the service be used?

The 3-Heights™ Scan to PDF Server is used for the following purposes:

- **Paper capture:** Electronic archiving of paper documents received as incoming mail within a company.
- **Facsimile capture:** Electronic archiving of all fax transactions between the company and its business partners.
- **Archive migration:** Migration of paper archives to an electronic archive with the standardized PDF/A format.
- **Web/mobile capture:** Use of the central service in client/server applications via a web service.
- **Enterprise application integration:** Use of the central service for PDF/A document creation via a programming interface (API) from specialist applications that create TIFF or JPEG files.

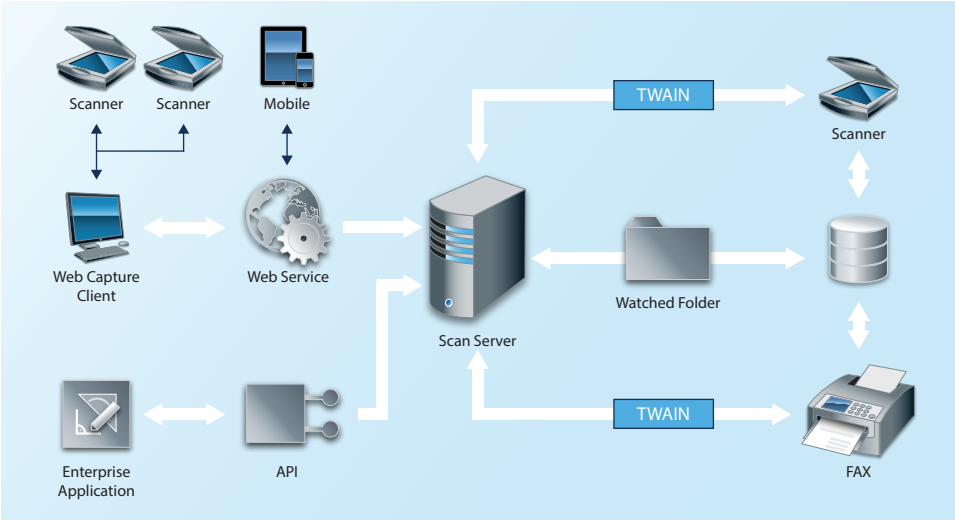


Illustration 4: Uses of the 3-Heights™ Scan to PDF Server

What are the advantages of a central service?

The 3-Heights™ Scan to PDF Server is worthwhile even for a small number of multifunction devices, but especially if high-performance scanners are used. In addition to economic benefits, it also offers qualitative advantages:

	Client	Server
Performance scaling	The scanning performance is limited by the performance of the workstation	CPU-intensive functions (compression, OCR) can be delegated to the server. Scalable performance through load balancing
Quality	The quality of the compression and text recognition in particular is limited by the technical features of the scanning station	Scalable performance optimizes the quality of the compression and text recognition
Installation and maintenance	Scanning software that includes all functions must be rolled out to every client and configured by the user	Most of the software can be installed, configured and maintained centrally on the server. Only the operator software needs to be installed on the client
User support	Problems can be resolved via a hotline or remote maintenance on the workstation	Problems can be simulated on the test infrastructure and resolved permanently in a live environment
Different versions and configurations	The installed version and its configuration can differ from one workstation to another	The central installation and configuration of the software ensures that documents are of standardized and consistent quality

Illustration 5: The advantages of central scanning software

What additional functions does the service offer?

The main function of the 3-Heights™ Scan to PDF Server is to convert scanned documents into a uniform, standardized file format, such as PDF/A.

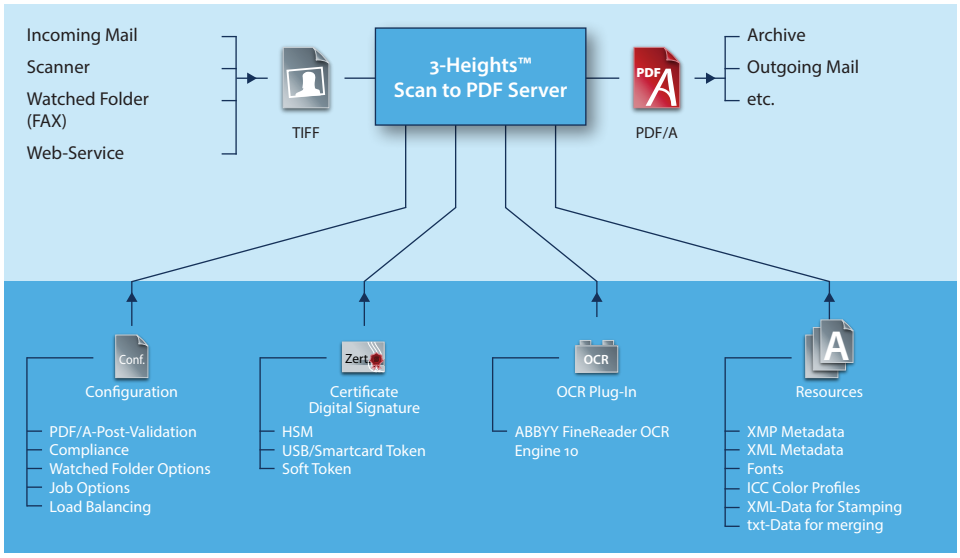


Illustration 6: Main and additional functions of the 3-Heights™ Scan to PDF Server

The service offers the following additional functions:

- **Embedding of XML files:** If TIFF files are created from specialist applications, it is often desirable to embed XML invoice data; for example, in accordance with the ZUGFeRD standard. The possibilities of PDF/A-3 can be used for this purpose.
- **PDF/A validation:** For quality assurance purposes, validation software can be used to check that the created PDF/A conforms to the ISO standard.
- **Document merging:** Single-page images need to be merged into multi-page files. Or documents belonging to the same business case need to be merged into a single file or, for example, a collection of files that corresponds to a folder. The service can read text files that control the merge for this function.
- **Stamping:** A stamp or watermark can be added to the created documents. The service processes an XML file containing the stamp data.

Additional functions can be integrated into the service by means of extensions (see further down).

Architecture and features

The features of the 3-Heights™ Scan to PDF Server are the result of the system architecture chosen by the development team. The development process was based on the following requirements:

- High-quality document conversion; in particular, compliance with ISO standards and image fidelity
- Robust, hands-off operation
- High through-put capability
- Performance scalability
- Interfaces for application integration
- Extensible for additional functionality

Compression and image quality

To convert the raster image formats TIFF and JPEG into PDF/A, the 3-Heights™ Scan to PDF Server uses integrated programs from the 3-Heights™ TIFF Tool Suite. This collection of image processing programs also performs the MRC compression. The image segmentation and compression processes used in the programs ensure both a high compression rate and high image quality.
The image segmentation breaks down the color images into background image (e.g. the color of the paper), image mask (e.g. text and lines of tables) and a foreground image that contains the fill color of the image mask. The size of the broken-down images is reduced using different compression processes. The background and foreground image can be compressed with JPEG200, for example, and the image mask with JBIG2. The compression procedure and the relationship between compression rate and image quality can be configured for the service and adapted to the requirements of the application.
The image segmentation can be improved considerably by using an OCR engine. The OCR engine provides the coordinates for different areas of the image, such as photos, text and remaining content. These coordinates are used by the segmentation algorithm; e.g. to isolate photos and compress them separately.

Text recognition

For text recognition, the 3-Heights™ Scan to PDF Server uses the 3-Heights™ OCR Server and the relevant OCR plug-in for the 3-Heights™ TIFF Tool Suite's programs. The OCR server itself uses the ABBYY FineReader engine. The engine also contains optional modules for handwriting recognition (ICR/IWR).
This architecture allows the compression and conversion function and the OCR function to be distributed across different servers, thereby enabling a more flexible load balance.

ISO-conformance and conformance check

To check that the generated PDF/A files conform to the ISO standard, the additional function for PDF/A validation can be used. The 3-Heights™ Scan to PDF Server uses the integrated program from the 3-Heights™ PDF Validator for this purpose. The program checks the following:

- Correct physical structure of the file (syntax)
- Correct logical structure of the document (semantics)
- The do's and don'ts for PDF/A
  - Accessibility (no encryption)
  - Clarity (calibrated colors, no invisible, dynamic or alternative content)
  - Self-reference and consistency (embedded scripts and color profiles, no references to external content)
  - Metadata (XMP standard and embedded schemes for extensions)
  - Searchability of the text (only for conformance levels a and u)
- Customer-specific checks
  - Resolution of scanned files
  - Compression process used
  - Existence of OCR text
  - Existence of required metadata
  - Corporate identity colors and fonts (only digitally created documents)

The conformance check is often a prerequisite for connection of the service to an archiving system that does not have this checking function.

Distributed architecture and scalability

The 3-Heights™ Scan to PDF Server is a scalable and freely configurable service. The service accesses a separate program for each work stage, such as compression, OCR recognition, conversion into PDF/A, etc. It receives the result of the previous work stage as its input and makes the output available for the next work stage. The work stages are linked by means of an XML configuration file. This architecture allows the work stages of the service to be structured in a highly flexible way, and enables almost any number of extension possibilities (see further down) by adding additional work stages.
To increase the level of parallel processing, the documents can be broken down into individual pages and sent through the processing stages simultaneously, after which they are then merged back into a single document. This option can improve the use of computer resources considerably (processor cores, memory, input and output, OCR engine, etc.).

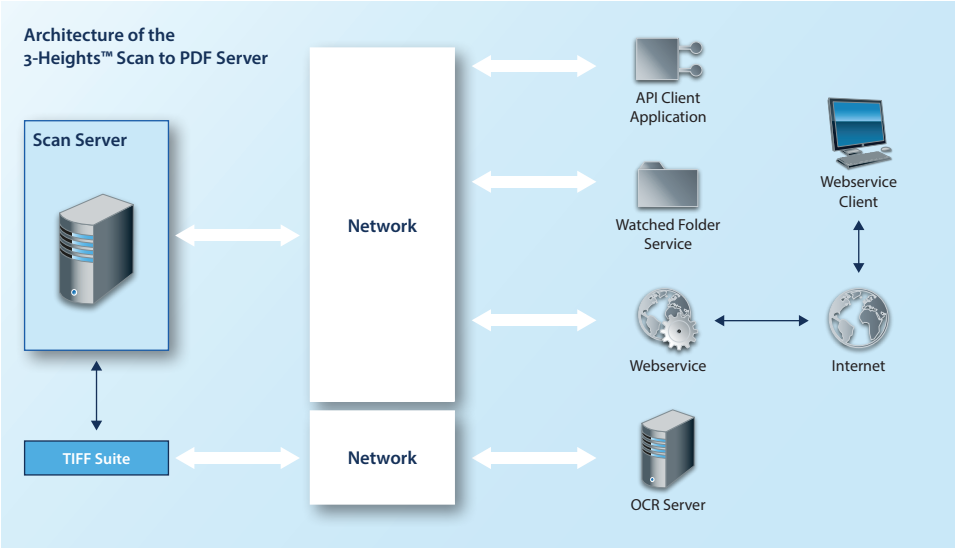


Illustration 7: Distributed architecture of the 3-Heights™ Scan to PDF Server

The service's tasks are distributed across several subsystems. The key subsystems are:

- **Scan Server:** The Scan Server has the following tasks:
  - Receive the conversion jobs, break them down into work stages and execute them
  - Perform the main and additional functions on the basis of the configuration
  - Delegate the text recognition to the OCR server and process the results of the OCR recognition
- **OCR Server:** This service has the following tasks:
  - Accept text and barcode recognition jobs
  - Pre-process the image (clean-up, deskewing, etc.)
  - Segment the image into photos, text and remaining content
  - Perform the text and barcode recognition using an OCR engine
  - Send the recognized text and barcodes back to the initiator. If desired, the pre-processed image and segmentation data can also be sent back.
- **Watched folder service:** This service monitors file system directories.
- **Web service:** This service accepts conversion jobs from the network and returns the converted files to the sender.

The 3-Heights™ Scan to PDF Server is scalable by distribution of the tasks to the subsystems. Different configurations are possible depending on the output requirements. In theory, the subsystems could be operated on one computer for basic configurations and on separate computers for complex configurations.
However, in live operating environments, it is usual to use just one computer that provides a large amount of memory through powerful multi-core processors.

Performance

The benchmark configuration for measuring the performance of the 3-Heights™ Scan to PDF Server is as follows:

- Hardware: HP Server DL 380p Gen 8
- Virtualization: ESXi / VMWare
- Operating system: Windows Server 2012
- Converter Service: 4 Worker Sessions
- Client: Windows 7

This configuration resulted in the following performance figures:

Test	Time
Conversion of TIFF to PDF/A	0,05–0,07 s/page
Overhead OCR	0,5 s/page
Overhead web service	0,2 s/page

Illustration 8: Performance values of the 3-Heights™ Scan to PDF Server

Interfaces for application integration

A range of interfaces are available for the integration of workstation and server computers that run applications wishing to use the 3-Heights™ Scan to PDF Server. The most important ones are:

- **Web service:** The web service enables scanning via internet/intranet from a web client or an application for a mobile device.
- **Programming interface (API):** This component enables the programmatic integration of the service into applications. It offers interfaces for Java, C, COM and .NET technologies. The component is also available for other platforms, including Linux, Sun OS, AIX, HP-UX, and Mac OS/X.
- **Command line tool:** This tool is a stand-alone program that can be run directly from the command line without any other requirements. A command language (shell command) can then be used to automate processes without the need for a development environment. The command line program is also available for other platforms, including Linux, Sun OS, AIX, HP-UX, and Mac OS/X.
- **File Explorer add-on:** This component is a Windows File Explorer extension that enables the user to convert individual files interactively.

Extensibility with additional functions

The 3-Heights™ Scan to PDF Server can be easily extended. To do so, it is necessary only to create an executable program that implements the desired work stage. The program can be integrated into the 3-Heights™ Scan to PDF Server through the configurations. The prerequisite is that it must be possible to start the program from the command line and provide the input and output files and the control options as parameters from the command line.

Here are a few examples of such additional functions:

- **Automatic classification:** The automatic classification of documents based on their content – after scanning suppliers, customer addresses and invoice numbers, for example – can speed up large volume document processing considerably. This process makes the index stations redundant for a large part of the scanned documents.
- **Splitting and merging of page content:** The content of a page can have several logical sections that may be divided by barcodes, for example. A desirable function may be to isolate these sections and distribute them as separate pages.
- **Conversion of color into gray-scale:** If color is not required for a specific use, this will free up additional storage space.
- **Importing other file formats:** Some scanners provide PDF files that can be imported and optimized directly by the 3-Heights™ Scan to PDF Server.
- **Automatic control of work stages:** Based on the content and formats, the 3-Heights™ Scan to PDF Server can control the type and sequence of the work stages.

Overview of the implemented features

The following table describes which features have been already realized in the current release. The remaining features will be implemented in future versions of the software.

Feature	Release 4.6
Input Formats: JPEG, TIFF, PDF	■
Output Formats: PDF, PFD/A	■
Mixed Raster Content (MRC) Compression	■
Optional OCR Processing	■
Optional Barcode Processing	■
Embedding of XMP Metadata	■
Optional Digital Signature	■
Optional PDF/A Validation	□
Watched Folders	■
TWAIN Interface	□
Web Service	□
Mail Server Interface	□
Embedding of ZUGFeRD Invoice Data	□
Plug-Ins: Merging PDF Documents, Stamping, etc.	□
Application Programming Interface (API)	□
Command LIne Interface (CLI)	□

PDF-TOOLS.COM
Premium PDF Technology

About PDF Tools AG

PDF Tools AG counts more than 4,000 companies and organizations in 60 countries among its customers, making it one of the world's leading producers of software solutions and programming components for PDF and PDF/A products.

Dr. Hans Bärfuss, founder and CEO of PDF Tools AG, began using PDF technology in customer projects more than 15 years ago. Since then, the PDF and PDF/A format have evolved into a powerful, widely used format and ISO standard that can be used for almost any application. During this time, PDF Tools AG has developed into one of the most important companies on the market for PDF technology, and has played a significant part in developing the PDF/A ISO standard for electronic long-term archiving.

As the Swiss representative on the ISO committee for PDF/A and PDF, the company's knowledge flows directly into product development. The result is high quality, efficient products based on the 3-Heights™ philosophy of the development team, which consists of experienced engineers.

The portfolio of PDF Tools AG ranges from components to services through to solutions. The products support the entire document flow, from raw materials to scanning processes through to signing and storage in a legally compliant long-term archive. An advantage of the components and solutions is the broad range of interfaces, which ensure smooth and easy integration into existing environments.

Due to the growing demands of the market, the products are enhanced and refined continuously. Support is provided by the developers themselves, allowing them to identify trends and customer requirements quickly and use this knowledge when planning enhancements and components.

All development activities are performed in-house at PDF Tools AG in Switzerland. The company does not outsource any programming, so that the entire development process can take place centrally in a single location. This helps to ensure the high standards expected by the company, particularly with regard to the 3-Heights™ technology.

The effectiveness of this approach is confirmed by the success of the products on the market. Our customers include well-known global companies from every industry. That is the greatest compliment of all – and the perfect motivation to continue shaping the world of PDF and PDF/A.

Further information

on our webiste via www.pdf-tools.com

- Manual 3-Height™ Scan to PDF Server
- PDF Expert Blog: blog.pdf-tools.com