White Paper

3-Heights™ Document Converter – Basics and Applications

PDF-TOOLS.COM
Premium PDF Technology

Contents

Introduction

What does a central conversion service do?

The 3-Heights™ Document Converter is a central service that converts corporate documents to a uniform, standardized file format PDF, PDF/A, or TIFF.

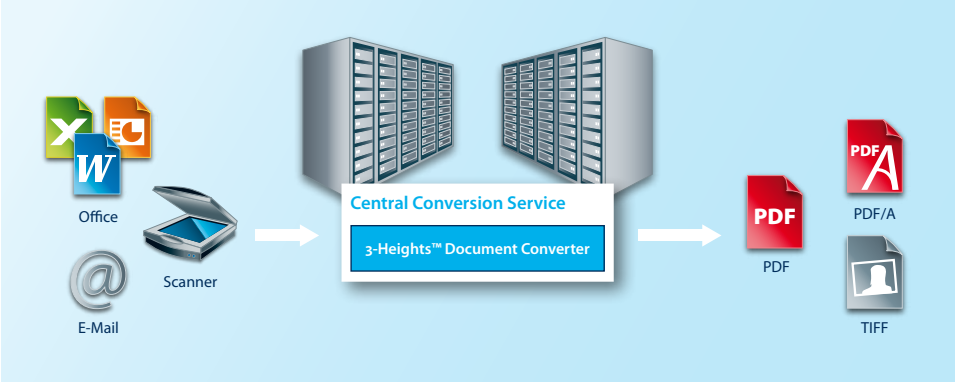

Figure 1: A central service for converting documents into a standardized format

Documents for conversion may exist in a wide variety of file formats. Most common are files produced by applications such as Microsoft Word, Excel, PowerPoint, and Visio, as well as emails, text, and scanned image files such as JPEG, TIFF, PNG, GIF, and BMP. Long-term retention and interchange with business partners are the main reasons for converting such documents.

How is the service used?

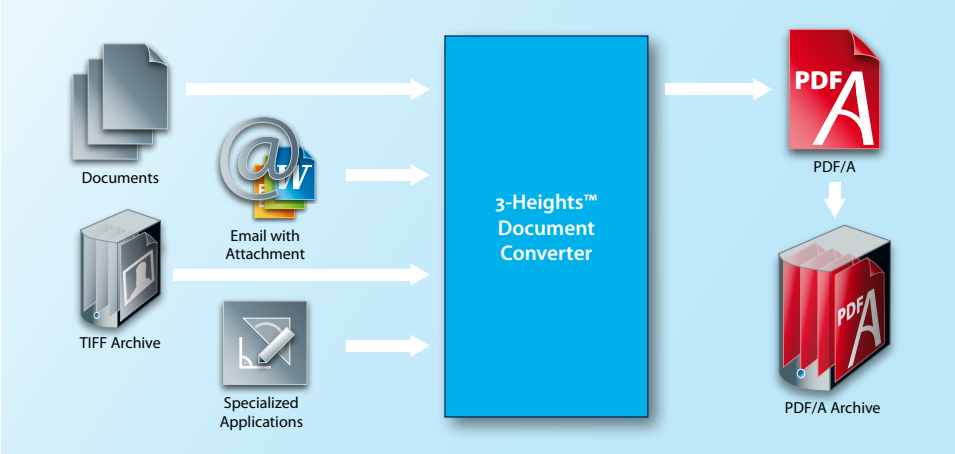

Figure 2: The 3-Heights™ Document Converter creates PDF/A files from many different file formats and applications for long-term storage

The 3-Heights™ Document Converter can be used to:

- Make PDF/A copies of all incoming and outgoing documents for the corporate archive
- Archive documents which are produced to support business processes
- Archive all email traffic between the organization and its business partners, including email attachments
- Migrate archives containing digital documents in an obsolescent or proprietary format to a new archive that uses the ISO standard PDF/A format
- Generate PDF documents from business applications centrally, via a web service or a programming interface (API)

What are the benefits of a central service?

The 3-Heights™ Document Converter makes sense even with just a small number of workstations since there are economic benefits and qualitative gains.

The following table summarizes the benefits of central service vs. local software:

	Client Based	Server Based
Scalability	Workstation computing power is the limiting factor in conversion performance	Load balancing enables performance to be scaled as desired
Installation and maintenance	Conversion software needs to be deployed to every client, and to be configured by the users	One centrally configured and maintained software installation
User support	Hotline and remote assistance to fix problems on the workstation	Problems can be reproduced in the test environment, then a lasting fix can be implemented in production
Unattended operation	Workstation-based conversion often involves manual user actions (responding to dialogs, acknowledging messages, etc.)	The conversion process runs with no manual intervention; the service itself controls native applications as required
Variety of input formats supported	Native applications must be installed on the workstation to handle each proprietary input format, limiting the number of supported formats	Standardized formats are converted directly; proprietary formats need just one centralized installation for all users
Application versions and configuration differences	Workstations may have different application versions installed, with configuration variances	Centrally installed and configured software ensures output documents with uniform and consistent quality
Robustness	Users must deal with possible disruptions caused by installation and configuration problems, or conflicting applications	The service runs each application in its own protected environment, monitors and automatically restarts them when problems arise

What additional functions does the service offer?

The primary purpose of the 3-Heights™ Document Converter is converting digital documents to a uniform, standardized file format such as PDF/A.

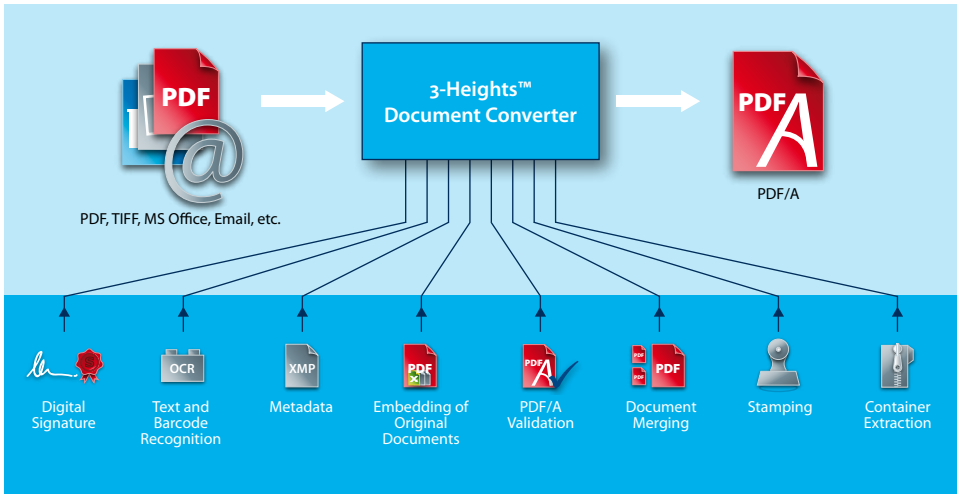

Figure 3: Main and ancillary functions of the 3-Heights™ Document Converter

The service also provides the following supplemental functions:

- Digital signature
  Applying a digital signature assures the document's integrity and authenticity. A signature may be created according to a specific signature law (e.g. Qualified Electronic Signature QED), meet the needs of long-term archiving or just the needs of a straightforward document exchange. A time stamp can be applied with or without a digital signature. The system generates digital signatures via a cryptographic infrastructure (USB token, HSM) and using a standard interface (PKCS#11).
- Text and barcode recognition
  Scanned image files need to be made searchable. The service can use the 3-Heights™ OCR Enterprise Add-On text recognition service to identify text in an image file and embed it in the converted version, thus making it searchable.
- Embedding of metadata
  The PDF/A ISO standard requires that metadata is embedded in the form of XMP packets into the document. The service offers this feature.
- Embedding of the original documents
  The original file will be embedded into the converted file, e.g. an Excel file into the PDF/A file. This is required if the original file contains information that is otherwise lost in the conversion process, such as Excel formulas. Another example is the embedding of mandatory XML invoice data (e.g. ZUGFeRD). The service implements this feature of the PDF/A-3 standard.
- PDF/A validation
  For quality assurance purposes, special software is available to check the conformance of PDF/A files with the ISO standard.
- Document merging
  Documents of the same business case can be merged into a single file or a collection of files, for example merge correspondences into a single dossier.

- Stamping
  Output documents may sometimes require a stamp or watermark. The service receives the stamp data from an XML file and applies the required stamps to the document.
- Container extraction
  Files may be packaged in TAR, ZIP, RAR and other containers, especially if those are email attachments. Such containers are often nested, i.e. the files inside are themselves containers. The service is capable of extracting content from nested containers to any depth, and sends the unpacked files for conversion.

Further useful supplemental functions are documented in the User Manual.

Architecture and performance features

The features of the 3-Heights™ Document Converter is a result of the system architecture chosen by the development team. From the outset, they based their work on the following requirements:

- High-quality document conversion with emphasis on conformity with ISO standards and reproduction fidelity
- Robust, hands-off operation
- High throughput capability
- Performance scalability
- Interfaces for application integration
- Extensible for further file formats and additional functionality

Quality: reproduction fidelity and ISO conformance

The 3-Heights™ Document Converter is comprised of the following:

- Native applications for performing conversions, in particular Microsoft Office
- Post-processing software for files directly after being produced by those applications
- Virtual printer driver specifically designed for this purpose: 3-Heights™ PDF Producer
- Built-in conversion programs for standard formats
- Verification software: 3-Heights™ PDF Validator

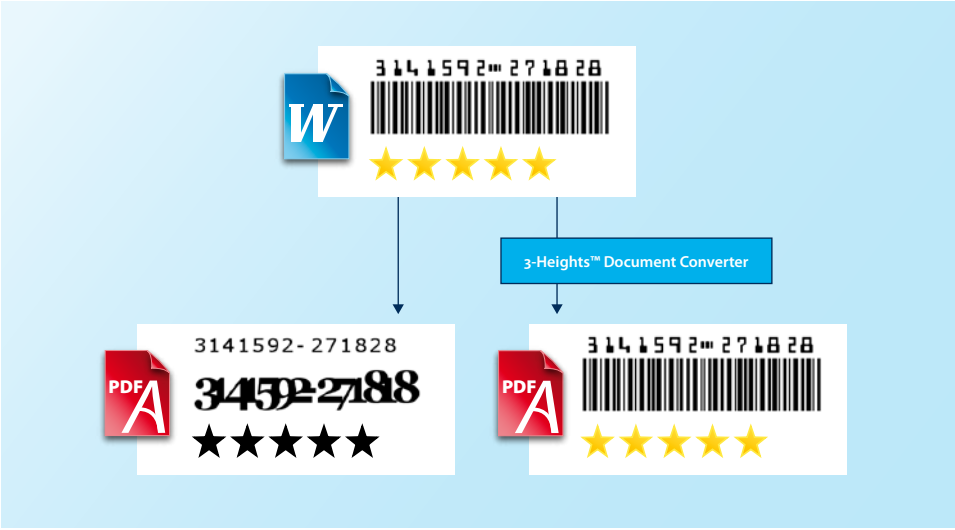

Figure 4: Ensuring high visual fidelity is a key feature of document conversion.

For high reproduction fidelity, the service uses the native application designed for a given file format – Microsoft Office in particular. Alternatives such as Open Office frequently produce rendering discrepancies that fall short of what users expect from converted documents.

Although target formats can often be generated directly from the native applications (‚Save As PDF/A'), results tend to lack fidelity of reproduction and standards conformity. When there is benefit to using an application's built-in functionality, the service does so and then post-processes the application output file for quality assurance purposes. Typically, it will use ‚Save As PDF …' and then convert the resulting PDF to PDF/A format.

With many applications, the print function is the only way of producing a PDF or TIFF file. For optimal document conversion via this route, the service employs a suite of virtual printer drivers specifically developed for this purpose. Those are capable of converting directly to the target TIFF or PDF format, rather than via a PostScript driver with its inherent graphical limitations.

The service uses built-in programs such as the 3-Heights™ Image to PDF Converter, 3-Heights™ PDF to PDF/A Converter, etc., for the conversion of standard formats such as the raster image formats TIFF, JPEG, PNG, GIF, and BMP as well as EMF and other vector graphic formats, and for converting PDF to PDF/A. Optimal and consistent quality conversion to those formats is thus assured.

Extra functionality for PDF/A validation can be used in situations where verifying conformity with the ISO standard is necessary. Conformity checking is often a prerequisite for linking the service with an archive system that does not have its own validation facility.

Robust, unattended operation

The 3-Heights™ Document Converter runs native applications like Microsoft Office in the isolated, controlled environment of a Windows Terminal Server session. Although

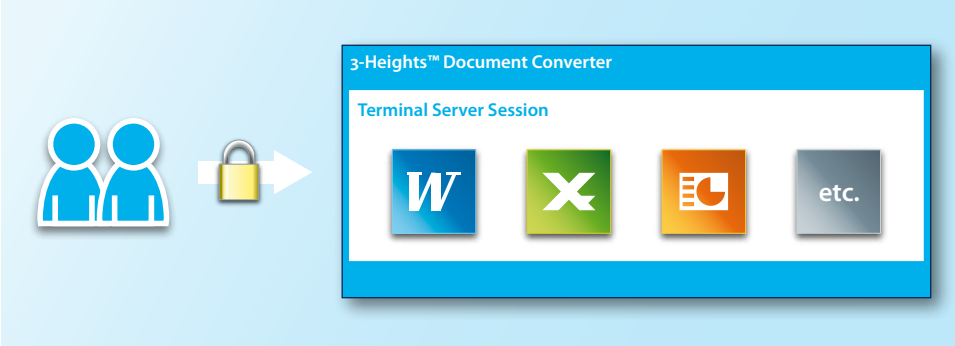

Figure 5: The original applications are opened in a Windows Terminal Server session to guarantee robust, stable operation

this may seem an exaggerated measure, it is necessary for robust operation. The main reasons for this approach are:

- Preventing interferences with interactive users
- Running multiple instances of an interactive application
- Automatic responses to interactive messages from applications
- Monitoring, starting and stopping interactive applications and sessions

Native applications such as Microsoft Office are generally designed for interactive operation. Controlling these via a programming interface comes with the risk that an interactive user may step in during a command sequence. This can seriously interfere with conversion and ultimately cause a failure. This kind of disruption cannot occur within a protected Terminal Server session.

Interactive native applications are usually able to convert only one document at a time. The service accommodates this behavior by serializing the conversion jobs. While this impacts throughput, the system can make up by starting multiple application instances in separate Terminal Server sessions.

Native applications designed for interactive use may display dialog boxes in the course of processing a document (e.g. opening a file, or printing). These dialogs require dismissal before processing will continue, so the service watches the application and steps in automatically to close open windows. An interactive session is the sole way to implement this functionality, ideally hosted in a Terminal Server environment.

A corrupt document may cause an interactive native application to freeze or crash. Therefore, the service keeps watching for this kind of event and automatically closes and restarts the application should a problem arise. The service also manages the Terminal Server sessions used for hosting instances of the interactive native applications.

Distributed architecture and scalability

With its distributed architecture specifically oriented to document processing, the 3-Heights™ Document Converter is broadly scalable by distributing service tasks among several subsystems.

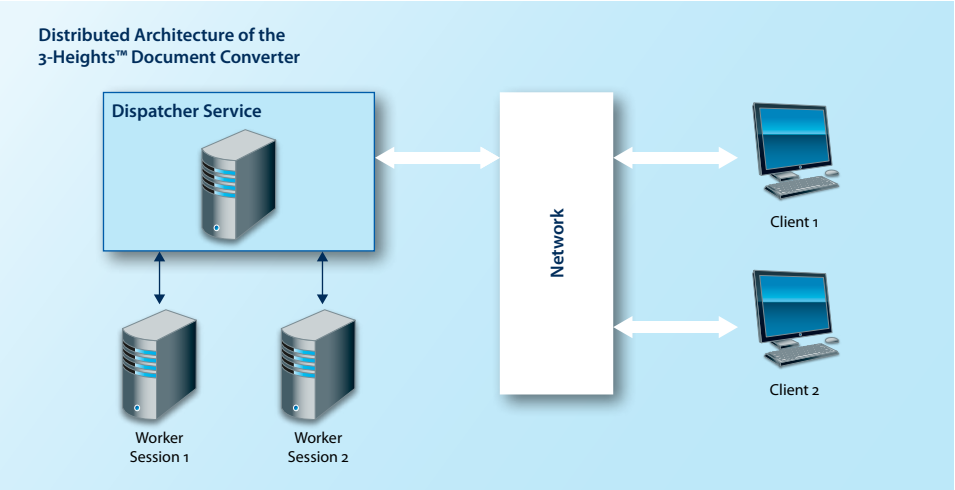

Figure 6: Distributed architecture of the 3-Heights™ Document Converter

The key subsystems are:

- Dispatcher
  The dispatcher is a multi-threaded Windows system service that runs as a single instance on each installation. Its tasks are:
  - Accepting conversion jobs and splitting them up for distribution among Worker Sessions that do the conversion work
  - Starting and managing Worker Sessions
  - Starting, monitoring and controlling interactive applications
  - Performing conversions that do not require a Worker Session
  - Performing supplemental functions such as text recognition, digital signatures, etc.
- Worker Session
  The Worker Session is a Windows Terminal Server session; one installation may run several sessions concurrently. A session:
  - provides a runtime environment for interactive applications
  - isolates multiple instances of an interactive application from one another
  - generates output files using the virtual printer driver
- Watched Folder Service
  This service monitors file system directories, sends conversion jobs to the dispatcher, and saves converted files to an output directory.
- Mail Folder Service
  This service monitors mail server mailboxes, sends conversion jobs to the dispatcher, and returns converted files as email or stores them in the file system.
- Webservice
  This service is an Internet Information Server (IIS) extension that accepts conversion jobs from the network and returns converted files to the sender.
- OCR Service
  A service used by the dispatcher for text recognition. It accepts image files and returns the recognized text.

The distribution of tasks among subsystems makes the 3-Heights™ Document Converter broadly scalable. Various configurations are feasible, depending on the required throughput. In a basic configuration, all of the subsystems operate within a single session on one computer. In a complex set-up, subsystems may be hosted on separate hardware. Practical production environments often use Windows Server with Terminal Server sessions, all running on high-performance hardware with multi-core processors.

Performance

The benchmark configuration for measuring the performance of the 3-Heights™ Document Converter was as follows:

- Hardware: HP Server DL 380p Gen 8
- Virtualization: ESXi/VMWare
- Operating system: Windows Server 2012
- Microsoft Office: Office 2013
- Converter Service: 4 Worker Sessions
- Client: Windows 7

Here are the resulting performance figures:

	2 pages		10 pages		50 pages		250 pages	
	PDF	Print	PDF	Print	PDF	Print	PDF	Print
DOC	0,42	0,55	0,77	0,78	2,02	1,72	8,5	6,4
DOCX	0,4	0,53	0,73	0,87	2,2	1,93	12,8	14,3
XLS	0,5	0,6	1,3	0,9	6	2,1	64	8,7
XLSX	0,45	0,6	1,2	0,85	5,9	2,1	63	8,6
PPT	1	1,2	1,3	1,6	2,3	2,7	6,8	7,5
PPTX	0,9	1,1	1,2	1,5	2	2,5	5,7	6,5
PDF	0,11		0,17		0,65		2,2	
tagged PDF	0,2		0,8		5,3		60	
TIFF bilev	0,2		0,4		4,1		20,5	6,9*
TIFF OCR	10,5		67		329,5			
TIFF color	1,4		1,9		10,3		51,6	

* From Doc
All numbers: seconds per document

Figure 7: Performance Figures

Various findings are arising from the measured processing times:

1. A significant portion of processing time is consumed by:
   - Starting Office applications
   - Opening documents in Office applications
   - Sending documents over the network
2. Twice as many pages do not necessarily mean twice the processing time; each document has a certain overhead, and content is also a factor. Complex, cross-referenced documents have a somewhat longer overall processing time.
3. Scaling with multiple Worker Sessions is gainful when multiple clients are able to use the service concurrently.
4. Faster hardware is beneficial for long documents with numerous pages.

Application integration interfaces

There is a series of interfaces for integrating workstations and servers that host applications used by the 3-Heights™ Document Converter. The major ones are:

- Webservice
  The web service presents a SOAP/XML interface. The web service integrates easily with applications using a WSDL file.
- Programming Interface (API)
  These are components for integrating the service with applications at the programming level. Java, C, COM and .NET interfaces are all provided. The same components are also available for other platforms such as Linux, Sun OS, AIX, HP-UX, Mac OS/X, etc.
- Command Line Tool
  The tool is a standalone program that can be run directly from the command line. A command language (Shell Command) can then be used to automate processes without requiring a development environment. The command line program is also available for other platforms such as Linux, Sun OS, AIX, HP-UX, Mac OS/X, etc.
- File Explorer add-on
  This component is a Windows file explorer extension for users to convert single files.

Extensibility:
Document formats and supplemental functions

Depending on its line of business, an organization may use many document formats. Moreover, a modern archiving system may be seriously challenged by files from older digital archives that were created by specialist applications using proprietary formats (e.g. project planning, design), or obsolescent formats (text processing). Every document format has its own peculiarities, so converting everything into one standard format is not always easy.

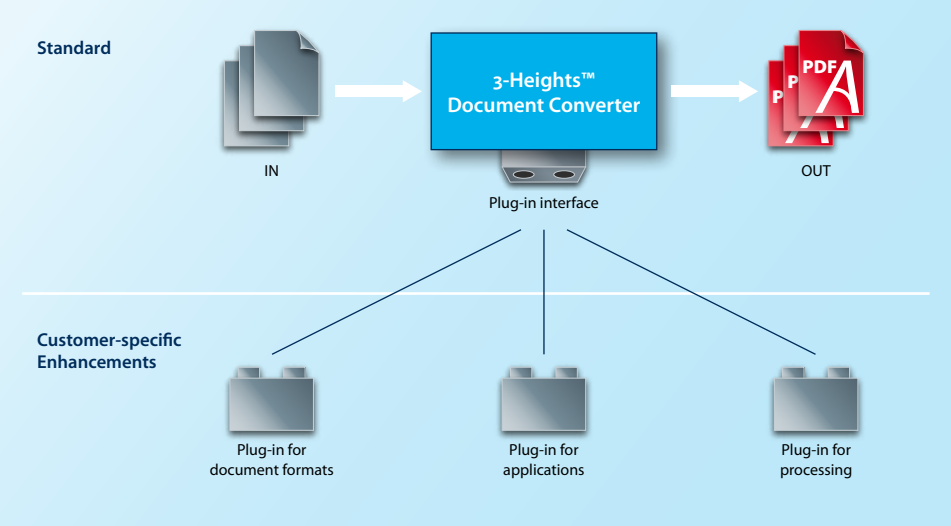

Figure 8: Plug-ins can be used to expand the number of formats and functions supported by the 3-Heights™ Document Converter

Sending documents like these to a central service is made possible by plug-in extensions, which are a feature of the 3-Heights™ Document Converter architecture. A plug-in is a program that performs a certain conversion process. There is an open interface dedicated to communication between the service and plug-ins. Two examples (COM and .NET) are provided to help developers get up to speed with plug-in programming.

Plug-ins are suited to integrating a specific conversion function with the service, extending native system functionality.

Product Editions

The 3-Heights™ Document Converter comes in two different editions. Each edition is intended for a certain purpose. Here are the differences between the editions:

Interfaces	Enterprise	Small-Medium Enterprise
Watched file folders	■	■
Watched email folders (via IMAP)	■	□
Command line (batch processing)	■	□
Programming interface (API)	■	□
Web-service	■	□
Shell extension for Explorer (right-click to convert)	■	■
Functions		
OCR	optional	optional
Merge files	■	■
Compliance validation	■	■
Digital signature	■	■
Encryption	■	■
Linearization	■	■
File compression	■	■
Support for meta data	■	■
Custom (e. g. CAD)	■	□
Load balancing	■	□
Script Plugin	■	□
Input Formats		
MS Word, Excel, Powerpoint, Visio	■	■
MS Outlook (MSG)	■	□
Simple text	■	■
Word Perfect	■	■
Open Office	■	■
Image formats (BMP, GIF, JPEG, PNG, TIFF etc.)	■	■
Nested containers (ZIP, TAR)	■	■
Websites (URL)	■	□
HTML	■	□
Email, email with attachment	■	□
Links	■	excl. URLs
Output Formats		
TIFF	■	■
PDF	■	■
PDF/A-1a, PDF/A-1b	■	■
PDF/A-2a, PDF/A-2b, PDF/A-2u	■	■
Zipped (TIFF or PDF)	■	■

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.

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

PDF-TOOLS.COM
Premium PDF Technology

Whitepaper-DocumentConverter-EN-20141027