PDF-TOOLS.COM
Premium PDF Technology

WHITE PAPER

PDF/A

the standard for long-term archiving

WHAT IS PDF/A?

Introduction to the PDF archiving format

PDF/A is basically a type of PDF that was designed specifically for long-term digital archiving. It combines the benefits of the PDF format with other specific requirements for long-term archiving. The PDF/A standard is a set of rules that defines which criteria a document must meet to be PDF/A-compliant. It is much more limited in scope than PDF, because PDF itself is already the underlying standard

Archiving formats vary from country to country. Traditional archiving methods (paper, microfilm, microfiche), while guaranteeing reproducibility, no longer comply with the latest technology. Traditional PDF types are sharply declining because they are less and less able to meet companies' requirements in the context of the digital transformation. There are also legal conditions, industry-specific regulations and internal guidelines that must be considered for any given archiving concept.

Many organizations set up TIFF archives as a first step towards electronic archiving. TIFF can now be transmitted quickly and easily in globally connected organizations; however, searching is still difficult.

The regular PDF format cannot fully meet the requirements of an archiving format either, which is why it was used as a solid foundation for developing the PDF/A standard.

A number of reasons make PDF more attractive than TIFF:

- PDF saves structured objects (such as texts, vector graphics, raster images) that support efficient searching in the entire archive. TIFF, on the other hand, is a raster format and must be processed with an OCR machine to enable a full-text search.
- PDF files are more compact and often require a fraction of the storage space of a corresponding TIFF file, often even with better quality. The small file size is especially beneficial for electronic data exchange (FTP, email attachments, etc.).
- Metadata such as title, author, date of creation and modification, content, keywords, etc. can be embedded directly in the PDF document. Thus, they can be classified automatically without any human intervention.
- The page contents in a PDF document are usually device-independent, i.e. independent from the raster resolution, color code, etc. The pages are not displayed on the raster until the reproduction (rendering process). PDF documents therefore benefit from the technological progress of output equipment, such as printer, monitor etc.

The creator of PDF de facto standards, Adobe Systems, has published several versions of its ‚PDF Reference Manual'. Each new version expanded the format with numerous new features and modified some of the old features. It was therefore necessary to develop a stable, internationally accepted standard for long-term archiving, built on Adobe's proprietary PDF specifications. The outcome: PDF/A.

Comparison of formats

Requirements	TIFF	PDF/A	XPS	Office
Long-term readability	(+)	+	(+)	-
Clear rendering	+	+	(+)	-
Data constistency	Proprietary tags for metadata	+	+	-
Authenticity / Integrity	With detached signatures	+	+	+
Required storage space	Black/white: + Colour: -	+	0	+
Searchability	Proprietary tags for OCR text	+	+	+
Long-term experience	+	+	-	-

Comparison between PDF and PDF/A

The normal PDF format does not guarantee long-term reproducibility or complete independence from the software and the output device. In order to guarantee both principles, it was necessary to both limit and expand the existing PDF specification. It was clear from the outset that PDF/A-1 had to be based on an existing version of PDF in order to achieve the acceptance of a wide audience.

ISO TC 171 uses Adobe’s PDF reference 1.4 (Acrobat 5) as the basis for the PDF/A standard (ISO 19005). It states that PDF/A “must meet all requirements of the PDF reference which additionally include this part of the ISO 19005 standard.”.

Der Standard beschreibt also nur die Unterschiede zur Referenz. Um PDF/A vollstänIn other words, the standard describes only the differences from the reference. To fully understand PDF/A, it is necessary to understand the PDF reference 1.4 as well. Certain functions that are supported by PDF 1.4, such as transparency or audio/video reproduction, were excluded from PDF/A. At the same time, PDF 1.4 includes optional elements that are mandatory in PDF/A. In the case of PDF/A, for example, all fonts that are used must be embedded. In short,

PDF/A primarily defines the specific properties set out in PDF reference 1.4 that are mandatory, recommended, restricted or forbidden.

THE PDF/A STANDARD

GOALS OF PDF/A

ISO Standard 19005 defines a file format based on PDF called PDF/A. The format 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.

This Standard specifies neither the methods nor the intention or the purpose of preservation. The Standard is thus intended to guarantee that electronic documents can be viewed in their original appearance, even in the future. For this reason, the document may not refer, either indirectly or directly, to an external source, for example an external image or a font that is not embedded in the document itself.

Where does the PDF/A format come from?

On September 28, 2005 the International Standards Organization (ISO) formulated a new Standard governing archiving of electronic documents – the official formulation is:

ISO 19005-1 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1)

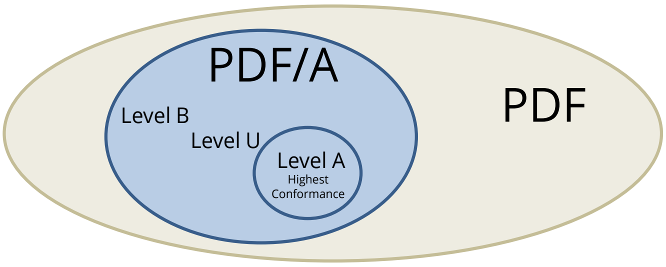
The Standard is the result of more than 36 months of collaboration among companies and organizations around the world. The initial impetus for this initiative occurred in May 2002 in the USA. The stated goal was to create a standardized format for electronically archived documents.

The Association for Information and Image Management (AIIM), the National Printing Equipment Association (NPES) and the administrative body for the US courts were all involved. The kick-off meeting took place in October 2002. Renowned PDF manufacturers participated, including: Adobe Systems, the Library of Congress, Surety Inc., Quality Associates Inc., Appligent, Merck, EMC, PDF Sages, and the National Archives & Records Administration (NARA). These were joined at a later time by others, including Xerox, Honeywell, EDS and Glaxo Smith Kline, among others.

The founders of the project put together a first version and submitted their recommendation to the ISO in order to have it registered as an international Standard.

The project was referred by the ISO to a Technical Committee designated TC 171 (Document Management Applications). This committee is composed of 15 member states and each has one vote, which is cast by their respective representatives. The committee is supplemented by an advisory commission representing another 21 countries. The Standard was improved over multiple stages until it was finally approved in September 2005.

PDF/A versions



The PDF/A-1 Standard is divided into two levels of conformance: PDF/A-1a and PDF/A-1b. PDF/A-1a (Level A Conformance) defines conformance with all requirements of the PDF/A-1 standard.

Part 1. The description of the minimum requirements for PDF/A conformance is

contained in PDF/A-1b (Level B Conformance). PDF/A-1b requirements should be adequate for ensuring visual, long-term reproduction.

In July 2011, the technical committee approved the second part of the standard: ISO 19005-2, Part 2 (PDF/A-2). While PDF version 1.4 serves as the basis for PDF/A-1, PDF/A-2 offers useful features that became available only in later PDF versions – up to and including PDF version 1.7. Most importantly, however, PDF/A-2 is not based on a certain version of Adobe PDF, but rather on the ISO 32000-1 standard.

In October 2012, the ISO committee already approved the third version of the standard (PDF/A-3, ISO 19005-3).

- **PDF/A-1** was introduced in 2005 as the ISO standard and is based on the Adobe PDF 1.4 (2001) standard. It is limited to the PDF features that were available at the time (e.g. no transparency)..
- **PDF/A-2** was released in 2011 and is based on ISO PDF 1.7.
- **PDF/A-3** is identical to PDF/A-2 and also supports a range of file attachments (e.g. used by Factur-X).
- **PDF/A-4** is expected to be released in 2019/20 and is based on ISO PDF 2.0 (2017)..

Conformity levels A, B, U

PDF/A-1a PDF/A-2a PDF/A-3a	[ACCESSIBILITY]	Accessibility – semantic correctness and structure (tagged PDF). Archive PDF, including complete accessibility to all types of content
PDF/A-1b PDF/A-2b PDF/A-3b	[BASIC]	Visual integrity – clear, long-term visual reproduction of static content
PDF/A-2u PDF/A-3u	[UNICODE]	Searchability of text and copying of Unicode text for digitally created PDF documents and documents scanned using optical character recognition (OCR)

PDF/A-1 versus PDF/A-2

PDF/A-2 does not replace or supersede PDF/A-1 in any way. PDF/A-1 conform documents that were already created will remain valid PDF/A files for long-term archiving. Archived PDF/A-1 documents can remain unchanged in the storage archives, so an „upgrade" to PDF/A-2 is not necessary.

For organizations that find the features introduced with PDF/A-2 useful, converting the original source documents to PDF/A-2 will have an advantage. This includes a higher rate of successfully converted documents and smaller file sizes thanks to compressed object and XRef streams. But likewise, for organizations that do not see a benefit of the features introduced with PDF/A-2, converting source documents to PDF/A-1 will continue to work fine. Both - PDF/A-1 and PDF/A-2 fully support the long-term archiving of PDF documents.

PDF/A-2 includes numerous features::

JPEG2000 Compression	The JPEG2000 Compression was introduced with the PDF 1.5 specification which was past the release time of the PDF/A-1 standard. Adding the JPEG2000 compression benefits particulary scanned documents.
Embedded PDF/A Files via Collections	Acrobat allows users to create collections (sometimes also referred to as „portfolios") where multiple PDF/A documents are combined into one „container PDF" document. A Possible use of a PDF/A collection is for instance the archival of email attachments can be converted to PDF/A and stored as „collections" inside a converted PDF/A email text body. PDF/A collections can also benefit security applications where a signature can be applied to individual single pages. The PDF/A collection then combines the signed single page. Individual pages can be subsequently be removed without affecting the validity of the signatures of the remaining pages.
Transparency	Although transparency is part of PDF 1.4, at the time of the PDF/A-1 standard release it was not defined well enough to be included in the PDF/A-1 standard. The specification has substantially matured since then, and transparency has become a common characteristic of PDF documents. Transparency is often found in the form of drop shadows, cross fades and highlight mar-ups for example.
Optional Content (Layers)	Optional content - sometimes also referred to as layers - is useful for mapping applications or engineering drawings where individual layers can be shown or hidden according to the information requirements of the viewing person. Another are of use is in user manuals of products that are sold internationally - where different languages can be implemented on different layers.

New Conformance Level PDF/A-2u - „u“ for Unicode	PDF/A-1b and PDF/A-2b concentrate on visual integrity, where „b“ stands for „basic“. PDF/A-1a and PDF/A-2a concentrate on accessibility - hence the „a“ notation. New to PDF/A-2 is the conformance level PDF/A-2u. It simplifies the text searching and copying of Unicode text for digitally created PDF documents and PDF documents that were scanned with subsequent optical recognition (OCR).
Object Level XMP Metadata	PDF/A-2 specifies the requirements for custom XMP metadata.
Comment Types and Annotations	Some of the newer comment types were added to the list of prohibited annotation types, and at the same time some of the newer comment types such as text editing comments are now acceptable to the PDF/A-2-standard.
Digital Signatures	While PDF/A-1 already allows for digital signatures, PDF/A-2 defines the rules that need to be applied to guarantee interoperability. Interoperabilität zu gewährleisten.

PDF/A-2 versus PDF/A-3

PDF/A-3 meets an important user requirement, namely an option to embed file formats that do not comply with the PDF/A standard. Because this amendment is desirable but controversial, it remains the only change to the PDF/A-2 standard. The user can therefore choose between a pure PDF/A collection and a mix of the various standards that is easily differentiated thanks to the „PDF/A-3“ label.

The purists among the experts opine that this amendment is contradictory to the original idea behind the PDF/A standard. However, pragmatists in companies from various segments, such as the pharmaceutical industry or the banking and financial sector, have a concrete need to keep the original file format alongside the converted PDF/A file. Files that belong together are compiled to form a „collection“. This construct has been known since the days of PDF/A-2. Typical applications include archiving emails and their attachments that can consist of many different file formats.

The standard only assures the representation of PDF/A documents viewed via a conform viewer. The presentation of non-conform embedded documents is implemented via a separate action using the tools that support the document formats in question.

PDF/A-3 should only be used if you plan to embed documents that do not comply with the PDF/A standard. PDF/A-2 is the right choice in all other cases as it makes it quite clear that no other formats are embedded. PDF/A-1 is still good enough for anyone who does not need all the functionality offered by PDF/A-2. There is no need to migrate existing archives as a PDF/A-3 conform viewer can display all PDF/A conform files.

CONVERSION FROM PDF TO PDF/A

A format conversion always poses certain challenges - for example regarding:

- Conversion without loss of information and quality
- Effort
- Authenticity (What happens to the signatures, for example?)
- Treatment of confidential data
- Non-convertible content
- Legacy

The conversion of a document into a PDF/A is a hybrid conversion. This means that not only the PDF/A specification influences the conversion parameters, but also those of the PDF standard itself. Typical examples are that the embedded fonts and the colors used must be calibrated. Less known is that the PDF/A standard contains additional, stricter rules. An example of this is that the text characters must not refer to the .notdef glyph.

PDF/A has been designed with document creation in mind not conversion. Nevertheless, a PDF to PDF/A converter must generate a new a PDF file, which follows the rules of the standard. Here are some examples:

- Uncalibrated color spaces can be easily replaced with calibrated ones by choosing an ICC color profile for each of the device dependent color space DeviceGray, DeviceRGB and Device CMYK.
- It is not necessary to introduce an output intent if it is not present in the input file. However, if the input file already has an output intent profile, e.g. a CMYK profile, then it is advised to keep it and the device dependent colors that refer to it.
- Embed missing font programs is only easy if the original font is available which is often not the case. If the font program is not available then it has to be replaced by a font program which has the nearest possible appearance. The viewer applications have built-in font replacement strategy. A converter should follow these strategies as well since the resulting file should look the same independent whether the fonts are embedded or not.
- If transparency is prohibited, such as with PDF/A-1, then the converter must perform some sort of transparency flattening or refuse the file if it cannot.
- With prohibited features such as JavaScript, multimedia content, some kind of actions etc. the converter has the option of removing the features or refuse the file if the user does not want it.
- Text characters that map to the .notdef glyph can be remapped to a new glyph which is a copy of the .notdef glyph.

Further tasks that a converter must perform:

- Pre-validation: If the input document does already conform to the requested standard there is no need to perform the conversion. This is in particular useful with digitally signed documents.
- Post-validation: After the conversion the user wants to verify that the converted result conforms to the requested standard.
- Repair: The user expects that the converter repairs the input file if it contains minor corruptions such as missing mandatory dictionary entries, damaged cross reference tables etc.
- Return status: A fine-grained return status value allows for a well designed, user controlled conversion process.
- Log file: An inevitable means to locate and eliminate conversion problems.

Validation of PDF and PDF/A

For businesses, it is essential that they know that the PDF and PDF/A documents passing through the business-relevant processes actually meet the respective standard. Not everything labeled as PDF/A is actually PDF/A – PDF/A is a quality criterion that supports the standard-compliant archiving in a long-term digital archive. Yet how can it be ensured that PDF/A documents generated externally as well as internally meet all of the standard’s criteria?

It is a good idea to check all incoming documents and repair them if needed. They will then be converted again and sealed/signed if needed. Documents that cannot be repaired or converted are labeled “invalid”. The documents should be validated again before they are finally stored in the digital long-term archive.

A PDF validator checks the conformance of a PDF document with a certain specification, e.g. PDF 1.7, PDF 2.0, or PDF/A. The tool includes several sets of rules – mostly in the form of profiles – that are used to analyze the documents accordingly. Caution: There can be significant discrepancies between the quality of a PDF or PDF/A document. The same applies to validators, for example, in terms of which aspects of the ISO standards are checked and treated as mandatory, what is clearly based on the standard, and what rules were ignored or interpreted incorrectly.

PDF/A AND DIGITAL SIGNATURES

In today's world, digital documents are closely intertwined with business processes. Electronic signatures play a key role in this respect. Knowledge in this area is sorely lacking, however. Electronic signatures have four main functions:

- **Replacing handwritten signatures:** Electronic signatures can satisfy the same requirement as a hand-written signature, provided that they meet the applicable legal requirements.
- **Protecting integrity:** Electronic signatures are a "seal" for digital documents.
- **Guaranteeing authenticity:** Electronic signatures make it possible to verify that the natural or legal person are identifiable entities.
- **Ensuring authorization:** Rights and authorities can be defined and managed in the certificate and therefore traced back to a certain person. Neither the electronic signature nor a specific document format (PDF, TIFF) can prevent a digital document from being altered by technical means. An electronic signature can, however, ensure that the change is always identifiable and traceable. This greatly improves the legal relevance of the digital document in legal processes.

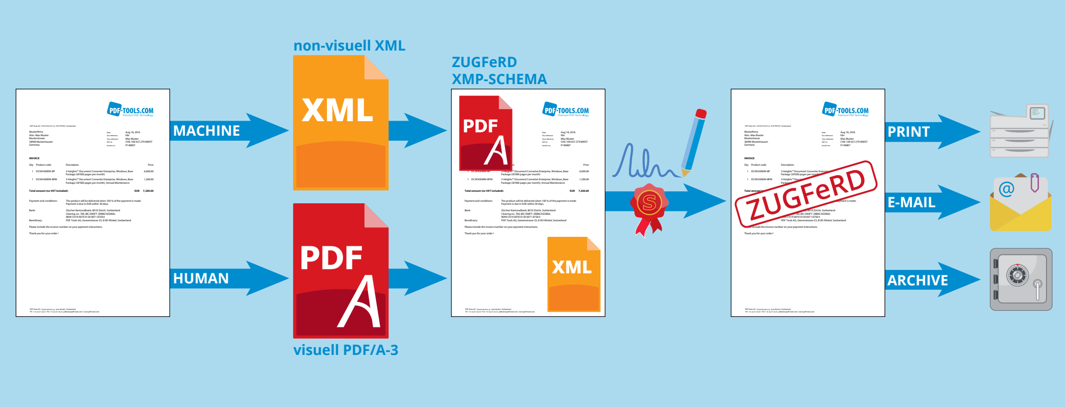
How electronic signatures are actually used in business processes depends on the particular situation. A qualified signature is required to create electronic invoices, for example. In the case of signed documents, the PDF/A format is recommended along with digital signing software that meets all requirements for valid signatures and long-term archiving.

PDF/A AND ZUGFERD / FACTUR-X

A standard for electronic invoices in PDF/A format

ZUGFeRD invoices are based on the ISO standard PDF/A. They combine machine- and human-readable data in the same document, which gives users the choice of handling invoices manually or through automated processes.

ZUGFeRD is a German invoice document format based on PDF/A-3 with embedded XML data (see www.ferd-net.de). The specification was developed by industry partners together with the PDF Association. The EU has also recognized the need for a standard format for electronic invoices and has created the European standard EN 16931 for this purpose.



This EU standard aims to standardize electronic invoicing throughout Europe and create a corresponding legal basis, with a view to supporting public administration processes and correct taxation in particular. It makes sending, receiving, and processing invoices electronically just as easy as in paper format – or even easier, thanks to automated processes.

The previous ZUGFeRD standard had to be adapted to the new one. ZUGFeRD 2.0, internationally referred to as “Factur-X”, was created as part of the Franco-German Digital Agenda. The new ZUGFeRD version meets all the requirements of EN 16931 and is also based on PDF/A-3.

ZUGFeRD is suitable for organizations of all sizes, and “Factur-X” also has greater international significance thanks to EN 16931 compliance. Moreover, starting from November 2018, electronic invoices will become mandatory in Europe for B2G transactions.

USING THE PDF/A STANDARD

Where do I get a copy?

The PDF/A Standard is distributed directly from the ISO Website (www.iso.org). Both paper copies and electronic versions (as PDF) are available. As is the case for all other ISO Standards, the document is copyright protected. It is therefore illegal to offer free copies via the internet.

To whom is the Standard addressed?

The objective of the PDF/A Standard is to optimize archiving methods. The Standard is purely technical in nature. For this reason, it is essentially only fully comprehensible to specialists with extensive knowledge about page description languages such as PostScript and PDF.

The main document itself is small, however the scope of the basis document is very large. PDF Reference 1.4 alone consists of almost one thousand pages – and this does not include all information associated with the Reference, such as font and compression formats, XML specifications, ICC color profiles, digital signatures, RFCs, etc.

In addition, the Standard alone cannot guarantee long-term preservation. A strategy for developing company-wide archiving is generally the result of a comprehensive project.

Collaboration with experts who understand the requirements of the PDF/A Standard and can apply them is recommended. Only in this manner can a consistent strategy be produced that ensures long-term document preservation goals.

CONCLUSION

PDF/A – the archiving standard

PDF/A is the standard for archiving electronic documents. The PDF format is widespread globally. It is used in both the public and private sectors for a wide range of purposes. The PDF/A Standard is the perfect instrument to ensure long-term preservation and reproducibility of documents over extended periods.

The PDF/A Standard also influences the future development of the PDF format itself. Independent of it, Adobe will continue to develop new functionalities. For example, 3-dimensional models or XFA for dynamic PDF forms. Conversely, these developments will influence the PDF/A Standard.

The PDF/A Standard will not be short-lived. Demand has existed for years for a standardized framework for archiving with PDF. The format is already used for precisely this purpose, even if many users must define specific guidelines in order to do so.

The fact that Microsoft is responding to customer demand by making it possible to create PDF/A documents directly from the most recent Office palette is a clear signal: the internationally valid PDF/A Standard for long-term archiving is here to stay.

PDF/A as a component of a comprehensive long-term archiving concept

In itself, the PDF/A Standard is merely a component of a comprehensive solution. In isolation, the Standard does not guarantee long–term preservation or reproduction parameters. Moreover, it is not the ideal solution for every project. PDF/A defines the specific requirements for electronic documents so that they can be archived over the long-term.

To build an archive that is conformant to the PDF/A Standard, other aspects must be taken into consideration. These include, among other things, in-house company standards and processes, quality management, reliable data sources and dedicated requirements tailored to the specific application purpose. In particular, the transfer of existing paper or TIFF archives to a PDF/A conformant archive requires careful planning.

FURTHER INFORMATION

PDF/A Competence Center

The PDF/A Competence Center was founded in 2006. The aim of this international association is to promote the exchange of information and experience in the field of long-term archiving in accordance with ISO 19005 - PDF/A. PDF Tools AG is a founding member of the association. In less than two years more than 85 companies and organisations as well as numerous experts from more than 20 countries have joined the association.

www.pdfa.org

Additional White Papers

If you would like to read more about specific PDF technologies, there are numerous web portals and white papers that could help you out. PDF Tools AG (http://www.pdf-tools.com) is publishing a complete series of White Papers dealing with a variety of PDF technologies.

PDF Expert Blog

For information from the PDF Experts - out from our develoment team - visit our Blog on blog.pdf-tools.com.

At your disposal

If you wish more information about standards, comparions and product information inclusive a tailored quote to your requirements, please don‘t hesitate to contact us.

Components & Solutions for PDF and PDF/A document processing

Get your own fully functional trial version of our PDF software for 30 days. Just visit our website for further details.

ABOUT PDF TOOLS AG

PDF Tools AG counts more than 5,000 companies and organizations in 70 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 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 andPDF/A.

SWISS
MADE

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

High Quality – High Volume – High Performance

Names and trademarks of third parties are legally protected property. Rights may be asserted at any time. The representation of third-party products and services is exclusively for information purposes.

PDF Tools AG is not responsible for the performance and support of third-party products and assumes no responsibility for the quality, reliability, functionality or compatibility of these products and devices.

PDF Tools AG
Kasernenstrasse 1, 8184 Bachenbülach, Switzerland

Tel.: +41 43 411 44 51 , www.pdf-tools.com