PDF/A for Scanned Documents

Overview of topics:

From analog to digital
- Previous solutions for scanned documents
- The PDF alternative
- If generating PDF, it makes sense to create PDF/A straight away

Full-text search options in PDF/A

Improved compression in PDF/A

PDF/A – usage examples
- PDF/A for personnel records in a service company
- The DAK: Migration of knowledge base to PDF/A
- PDF/A for the decentralized scanning of credit files

PDF/A documents are not only generated from digital sources – a large percentage of documents are created from scanned hard copies received by mail or from files that are being converted to digital form. In such cases, the company has no access to the original files, and the documents that need to be converted into electronic documents are merely paper copies. PDF/A is preferable to other electronic formats because it is an ISO standard and a target format that provides a range of benefits with regard to archiving and reusing content.

From analog to digital

The digitalization of paper documents (letters, files, invoices, photographs, and many more) is part of everyday life in many companies and institutions. There are many common processes depending on the various intended uses of the documents concerned.

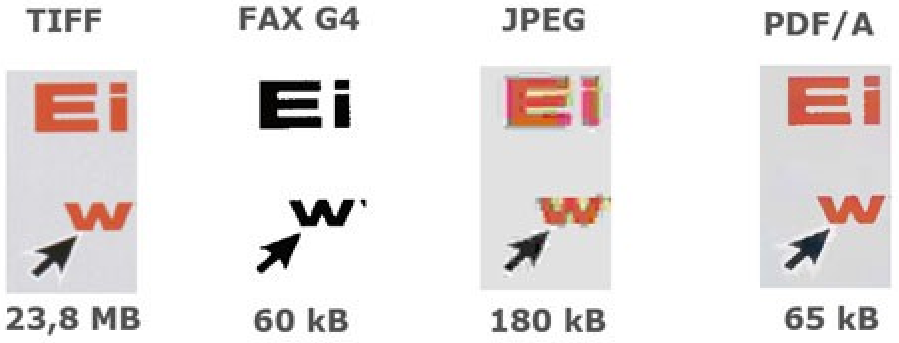
Previous solutions for scanned documents

In the case of documents that only exist in black and white, such as invoices, TIFF G4 has often been used, a format that is still in use today. This format was developed for fax transmissions. If original color documents exist, the JPEG image format is a popular choice. Other less common formats include PNG and BMP. In certain cases, special formats such as ‘JPEG in TIFF’ are preferred, in order to reduce the file size or create multi-page files, for example.

Disadvantages:

These older methods are subject to a series of disadvantages in comparison with the digitalization of documents using the PDF/A format. Users who still work with these older formats today will be confronted with problems such as the following:

- **Format variety:** Because different file formats are required for different tasks, the older procedures do not result in a uniform format for scanned documents. In certain circumstances, users have to use a different viewer for each format. As a rule, each display program is operated differently. Only one viewer is required to display PDF and PDF/A – one of them, the Adobe Reader, is even available free-of-charge.
- **Loss of information:** PDF/A is capable of adopting content in a one-to-one fashion. Other, older file formats cause the loss of detailed information. One example is TIFF G4, which can only display content in black and white.
- **Image quality versus file size:** When using image file formats, users are often faced with a choice between bad quality or large files. For example, if using JPEG, the size of a file can only be reduced if the user accepts a consequential reduction in its quality. This disadvantage is particularly irksome when displaying text, and it can impede readability.



File size versus display quality: The images are derived from a page in DIN A4 format and the file sizes also refer to DIN A4 with a resolution of 300 dpi.

- **The myth of revision-safe TIFF:** The commonly held opinion that storing documents and data in TIFF is sufficient to make them revision-proof is a falsity. Every format can be manipulated and it is especially easy to make slight changes to the TIFF format. Archive formats can only form a revision-proof solution in the overall context of the system in which they are being used, whereby the systems themselves – for example, a DMS or financial accounting system – provide the required security, rather than the format used.

- **Non-uniform metadata:** If a file archive comprises a large number of documents in different formats, it is not possible to achieve standardized metadata for all the formats used. Each file format tends to build on its own proprietary solution, making standardization impossible. PDF/A provides a uniform metadata system. The standard XMP (Extensible Metadata Platform) integrates any additional information directly into the PDF file itself, making it permanently accessible. This means that users can call up specifications such as the author, access permissions, keywords, and copyright directly and without resorting to the use of a database.
- **Full-text search:** Most image formats do not support text recognition (OCR) for files. As a result, they – unlike PDF – do not permit a full-text search.
- **Laborious data recall:** Image formats only allow data to be recalled via databases, not at file level. Example: You want to find "Simon Sample"'s personnel files. The database can localize all documents that name this person, but cannot highlight the exact location of the hit on the correct page. In the case of large documents, this can result in extremely time-consuming searches and – in consequence – high costs.

The PDF alternative

PDF is a modern, standardized alternative. Digitalization via conversion to PDF is already a popular choice for users who wish to standardize document formats (Image2PDF) or enable full-text searchability. PDF also permits the use of newer, more powerful compression formats, such as JPEG2000. Many users have switched to PDF in order to achieve metadata uniformity.

Using PDF eliminates all the disadvantages of the older formats, but – even so – the traditional PDF format is not the best solution for every single usage area.

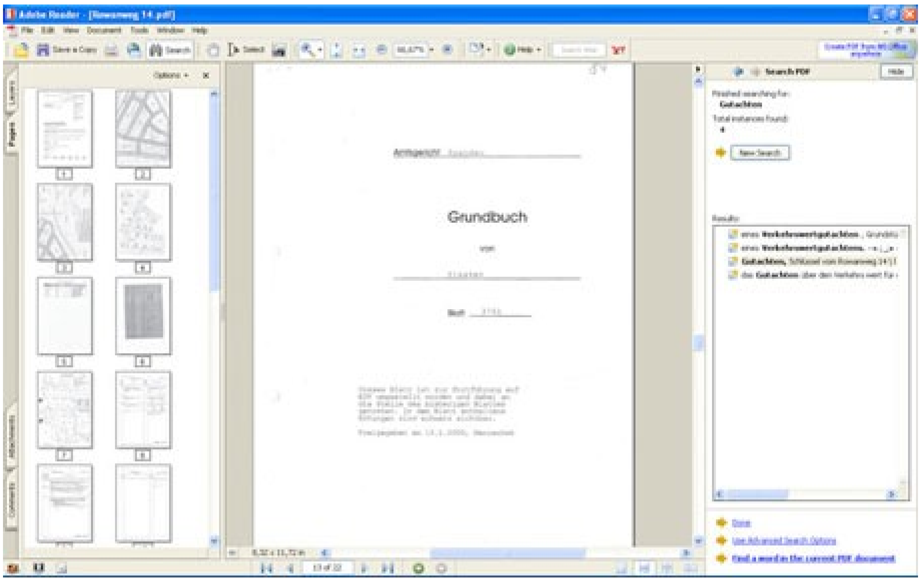
If generating PDF, it makes sense to create PDF/A straight away

If you decide to use PDF as your archive format, it makes sense to use the PDF/A variant, since this is the only format that was developed as an ISO standard for long-term archiving.

Full-text search options in PDF/A

PDF enables text searches at file level. This improves the usability of the documents concerned in many areas, such as the following:

- Electronic libraries – after download
- Manuals, design documents, and construction files in archives for product liability purposes
- Documents that are sent to customers, tax consultants, or attorneys



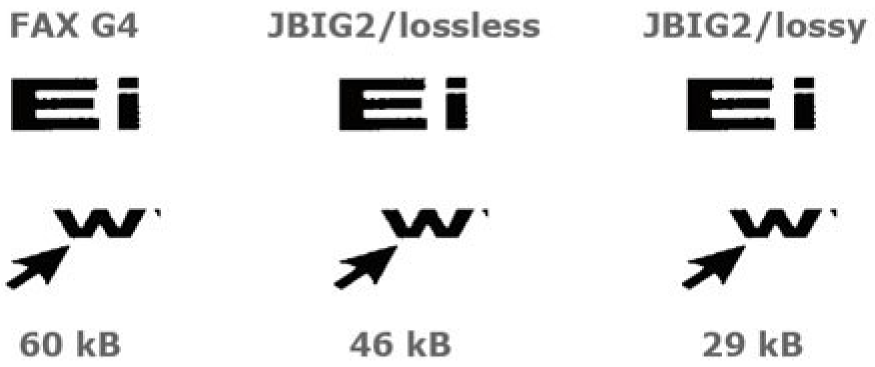
Text search in a PDF file, in this case in a plat book.

Improved compression in PDF/A

For documents in black and white

An increasing number of customers who process black-and-white documents recognize the advantages provided by PDF/A.

In the case of black-and-white documents, the JBIG2 compression format (standardized in ISO/IEC 14492) is particularly effective. This compression format is positioned as an alternative to TIFF G4. JBIG2 allows users to choose between lossy and lossless compression. This technology, which is – as yet – not well-known, has been implemented in PDF/A-1 and is available in Adobe Reader.



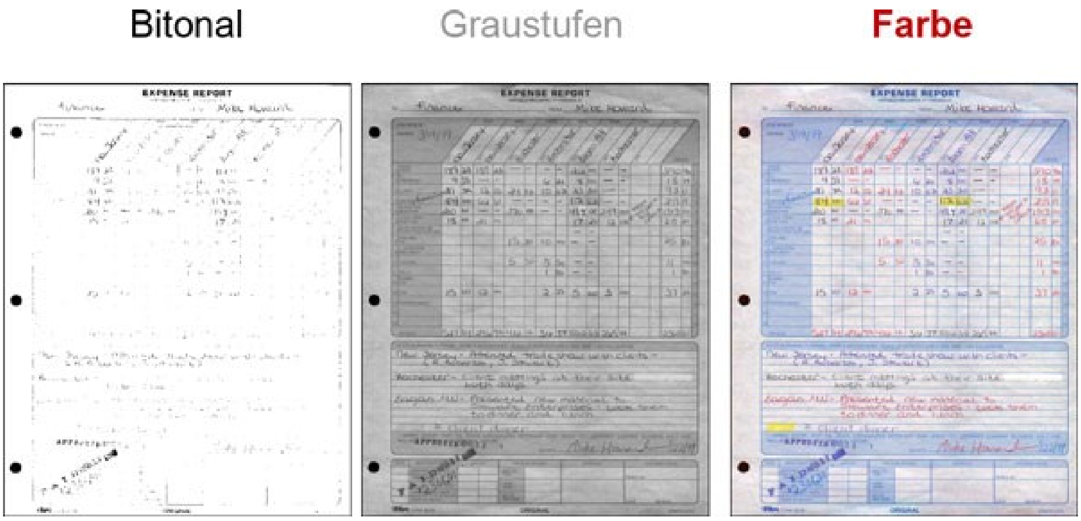
The JBIG2 compression format significantly reduces file sizes for best-quality text (these values refer to a scanned page of DIN A4 in 300 dpi).

For color documents

Color is an important bearer of information. It can have both content-related and semantic significance. The processing of color documents increases the productivity of employees and thereby helps companies to reduce costs.

A study that was instigated by **Kodak** found that employees work better with color documents, which bring the following advantages:

- Around 14% better comprehension of documents
- Around 70% improvement in decision-making ability
- 80% improvement in reading accuracy

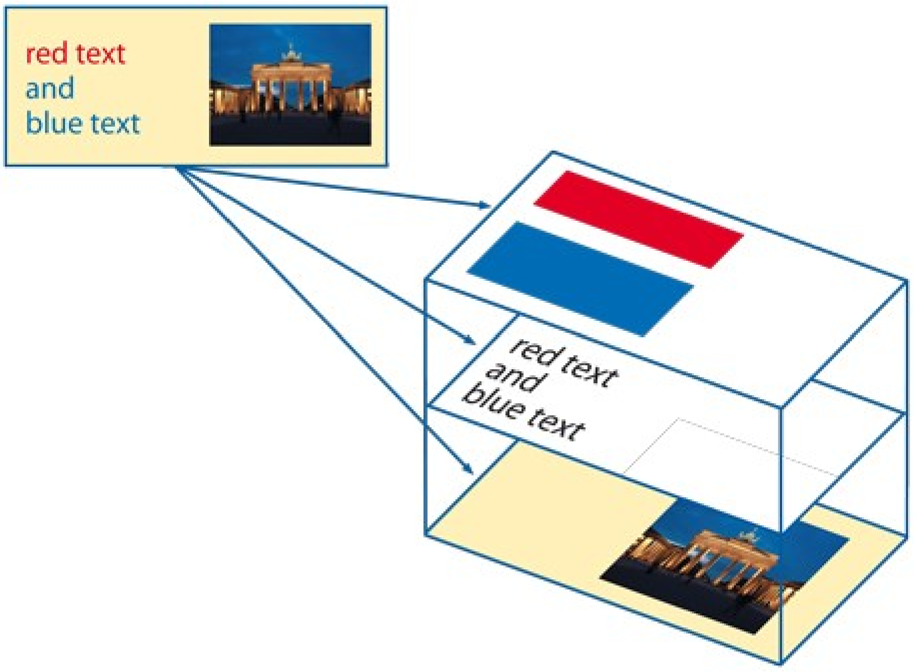


Color helps employees to understand content. Many color documents lose important details when they are processed. For example, highlighted text can become illegible when scanned in black and white.

If all documents are scanned in color rather than being separated into color documents and black-and-white documents, the pre-sorting effort (which accounts for around 75% of the costs) is drastically reduced. This method also means that there is no need for changed scanner settings or rescans of a single document.

In the case of color documents, powerful compression of the image data can reduce file sizes significantly. MRC compression – which is also known as **JPEG2000** (JPM) – can drastically reduce file sizes without causing a visible decrease in the display quality.

LuraTech uses a procedure that efficiently solves the problem of file size reduction in its Scan-to-PDF/A solutions. The division of each document into three layers that are converted entirely separately from each other enables the separate compression of text, colors, and images.



*The **LuraTech** layer procedure combines the benefits of the sharp display of color images and text with a particularly small file size.*

Three-layer technology produces optimum quality by digitalizing a compressed original that splits the content into text, image, and color layers using modern MRC procedures.

PDF/A – usage examples

The three case studies below show the advantages of the digitalization of documents via conversion to PDF/A for personnel records, knowledge bases, and credit files.

PDF/A for personnel records in a service company

This company is a services company that has a global turnover of 7.1 billion euros and a turnover of 420 million euros in Germany alone. A total of 220,000 employees work for the company worldwide, 14,500 of them being based in Germany.

The project

The task definition was as follows: 14,000 personnel records of around 150 pages each needed to be digitalized. This corresponds to a total processing volume of 2 million pages. These documents must be available to 200 authorized employees with access from 70 locations. The paper documents existed in black and white, grayscale, and color. The solution was the conversion of the original documents into the ISO future-proof PDF/A variant of the PDF format along with effective compression to reduce file sizes as much as possible. The OCR (Optical Character Recognition) process prepared the scanned text for full-text searching.

The results

The uniform conversion of the document set into PDF/A enabled all personnel files to be safely retained in digital form. The ISO PDF/A standard guarantees the suitability of data for long-term archiving. It makes it significantly easier to use the data, since employees now have access to documents that support full-text searching. The electronic search function replaces visual searching, resulting in a high accuracy of hits at the same time as saving time. Choosing the PDF/A format also results in files that are up to 60% smaller than if using TIFF or JPEG. Lastly, the smaller file sizes cause a significantly lower network load and permit direct access to data.

The advantages at a glance:

- Safe data storage for decades
- PDF files that support full-text searching
- Small file sizes (up to 60% reduction in size)
- Lower system load and quick access

The DAK: Migration of knowledge base to PDF/A

The DAK's INFO services needed to be digitalized to provide uniformity. The DAK (*Deutsche Angestellten-Krankenkasse*) is the second largest health insurance company in Germany, with 6.2 million members and 12,000 employees working in 750 branches.

The project

The internal information archive, which contains around 300,000 pages of text, took the form of image files before the migration. Most of the text was stored in TIFF format, with more recent additions in PDF. The stored information – originally stored on microfilm – was already partly digitalized, but using a mix of formats. TIFF, for example, neither saves space nor provides full-text search options. This archive is growing constantly, with around 3,000 new documents each year. Each file can have 50 or more pages. The aim of the project was to create a uniform archive with as low a file volume as possible while enabling digital data recall.

In order to optimize the possibilities provided by the info service and to make them future-proof, the DAK decided to archive the knowledge base in PDF/A format. The DAK used LuraTech's PDF/A solutions to carry out the migration. During this initial project, the DAK was able to gain early experience of the new PDF/A format that will be of use in later projects.

The results

The employees of the DAK's INFO service can now enjoy the advantages of easy and quick full-text search functions. The smaller file sizes allow information to be accessed more quickly. Naturally, a program for displaying the data must be installed on employee's PCs. The DAK uses Adobe Reader, which can be downloaded from the Internet free-of-charge. Thanks to PDF/A, the DAK's data is now suitable for long-term archiving in accordance with the ISO standard. Lastly, the DAK has gained practical experience from this reference project with regard to further data archiving using PDF/A.

The advantages at a glance:

- Files that support quick full-text search options
- Reduction in required disk space
- Quicker, easier access to documents for users
- Long-term readability
- Only one viewer required (Adobe Reader)
- Acquisition of practical PDF/A experience

PDF/A for the decentralized scanning of credit files

In Tennessee, the headquarters of an American finance company,‵check into cash' procedure documents were digitalized and stored in a data archive in PDF/A format. The financial service provider concerned has 1,200 payday advance centers in 30 US states.

The project

The service provider required the decentralized scanning of credit files. Documents were to be processed in color throughout. Lastly, the switch to the new system was to improve the transmission of data to headquarters.

The results achieved for the centers:

The centers now benefit from quick data transmission thanks to the implementation of the LuraDocument PDF Compressor, which creates PDF/A documents via scanning and data conversion procedures. All documents can be processed in color. This means that the centers do not need to sort documents into color documents and black-and-white documents before digitalizing them. This has resulted in a considerable decrease in processing time.

The results achieved for the headquarters

The headquarters, where the PDF/A documents are stored, have benefited from a reduction in the required disk space since the modern data compression procedure used yields significantly smaller file sizes. Smaller file volumes also cause a noticeable reduction in administration costs. Last but not least, the company's headquarters benefit from the long-term readability of data and safe archiving in accordance with the ISO standard.

The advantages at a glance:

- No need to pre-sort documents into color documents and black-and-white documents
- All credit files can be read in a single process
- Reduction in file sizes
- Quicker transmission of data
- Safe long-term archiving of credit files

Conclusion: PDF/A is the optimum format for scanned documents

PDF/A is the format for scanned documents. It can be implemented in every single company and institution without any major technological problems. Anyone considering digitalizing paper documents today should choose the modern, standardized PDF/A solution straight away. In an environment where other formats have been used to archive scanned paper documents up until now, clearly defined, well arranged projects provide an opportunity to experience the advantages of PDF/A and gain practical experience of this new format.

Carsten Heiermann, LuraTech/aoe