Webinar

www.pdfa.org

PDF/A for scanned documents

Paper becomes digital

Carsten Heiermann, LuraTech, Managing Director
Armin Ortmann, LuraTech, CTO

Carsten Heiermann
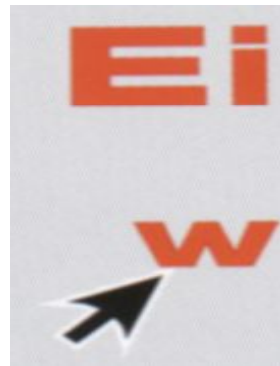Managing Director

Existing Solutions for Scanned Documents

- **black/white: TIFF G4**

- **Color: JPEG. Randomly used PNG, BMP and other raster graphics formats**

- **Often special version formats like „JPEG in TIFF"**

- **Disadvantages:**
  - **Several formats already for scanned documents**
  - **Even more formats for digital born documents**
  - **Loss of information, e.g. with TIFF G4**
  - **Bad image quality and huge filesize, e.g. with JPEG**
  - **No standardized metadata spread over all formats**
  - **No full text searchability (OCR) inside files**

Black/White: **Color:**
- **TIFF FAX G4**   - **TIFF**
  - **TIFF LZW**
  - **JPEG**
  - **PDF**

Carsten Heiermann
Managing Director
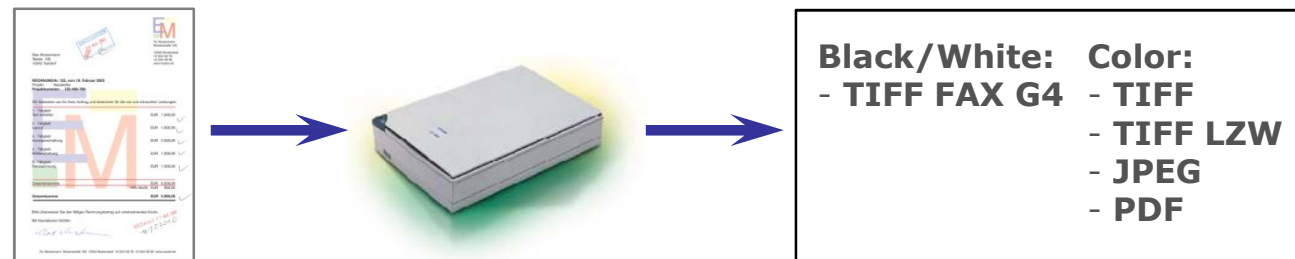
Existing Solutions for Scanned Documents

▷ **Bad image quality vs. file size**

TIFF/BMP	JPEG	TIFF G4
23,8 MB	**180 kB**	**60 kB**
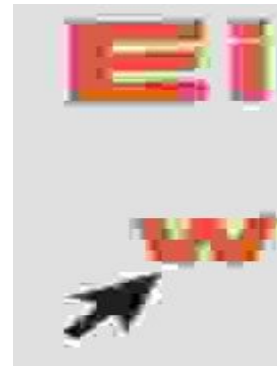
Alternative Solution: PDF

> **Already used widely:**
>> **Either to unify file formats**
>>> Image2PDF
>>> but also "Office" and other formats
>> **Or to have full text enabled/searchable files**
>> **Or to have access to new compression schemes (e.g. the JPEG2000 file formats family)**
>> **Or to harmonize metadata**
>>
>> **Conclusion: PDF has none of the disadvantages of the legacy formats**

> **"If the decision for PDF as archiving file format is made, why not PDF/A – with all its advantages over PDF in that application area then?"**
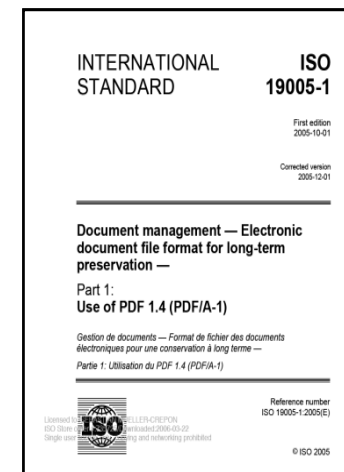
Carsten Heiermann
Managing Director

PDF/A

What is PDF/A?

- **ISO 19005-1, Document Management**

- **Electronic document file format for long-term preservation**

Goals of PDF/A

- **Maintain static visual representation of documents**

- **Consistent handing of Metadata**

- **Option to maintain structure and semantic meaning of content**

- **Transparency to guarantee access**

- **Limit number of restrictions**

Carsten Heiermann
Managing Director

INTERNATIONAL STANDARD

ISO 19005-1

First edition
2005-10-01

Corrected version
2005-12-01

Document management — Electronic document file format for long-term preservation —

Part 1:
Use of PDF 1.4 (PDF/A-1)

Gestion de documents — Format de fichier des documents électroniques pour une conservation à long terme —

Partie 1: Utilisation du PDF 1.4 (PDF/A-1)

Reference number
ISO 19005-1:2005(E)

© ISO 2005

PDF/A – full text searchability (OCR)

▷ **Benefit: searchability on file level**

   ▷ **e.g. digital library, „after book download"**

   ▷ **e.g. large manuals or multi page construction files**

   ▷ **e.g. documents fetched from the archive database and send to customers, suppliers, lawyers,...**

Carsten Heiermann
Managing Director
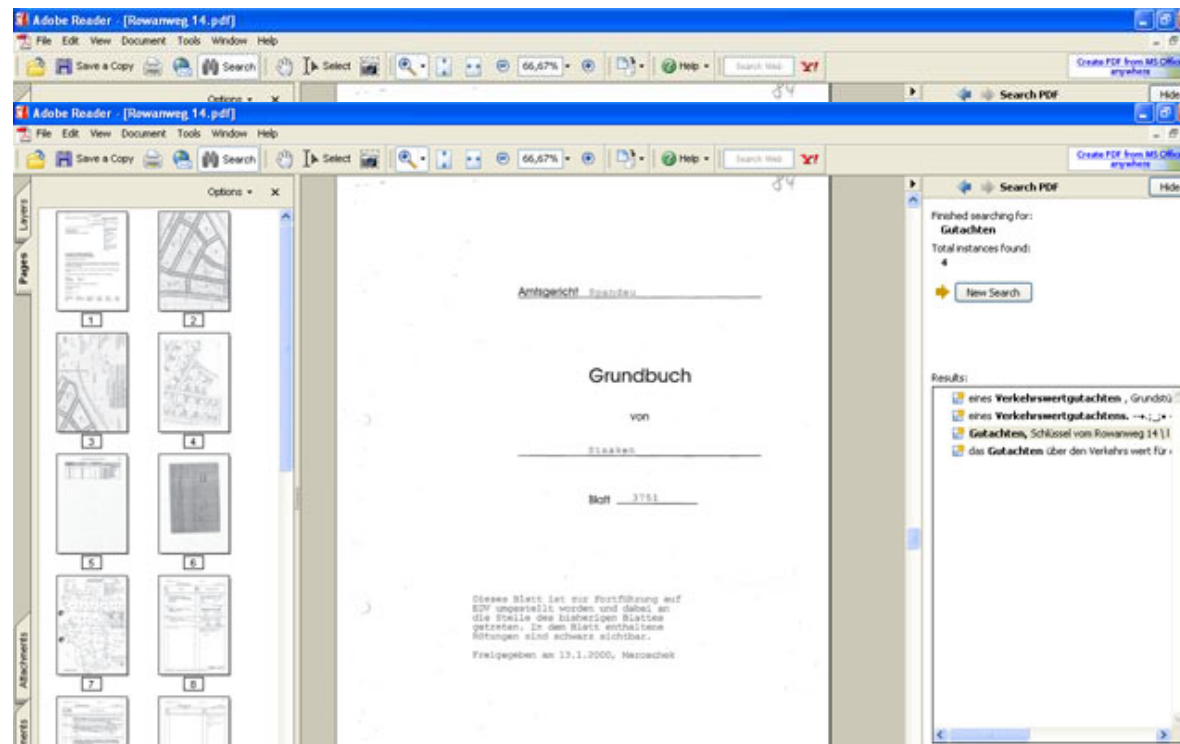
PDF/A – Enhanced Compression

- **For black/white documents**
  - **JBIG2 - ISO/IEC 14492**
    - Used as alternative to TIFF G4
    - Full and visual lossless mode
    - Embedded in PDF/A, available in Acrobat Reader
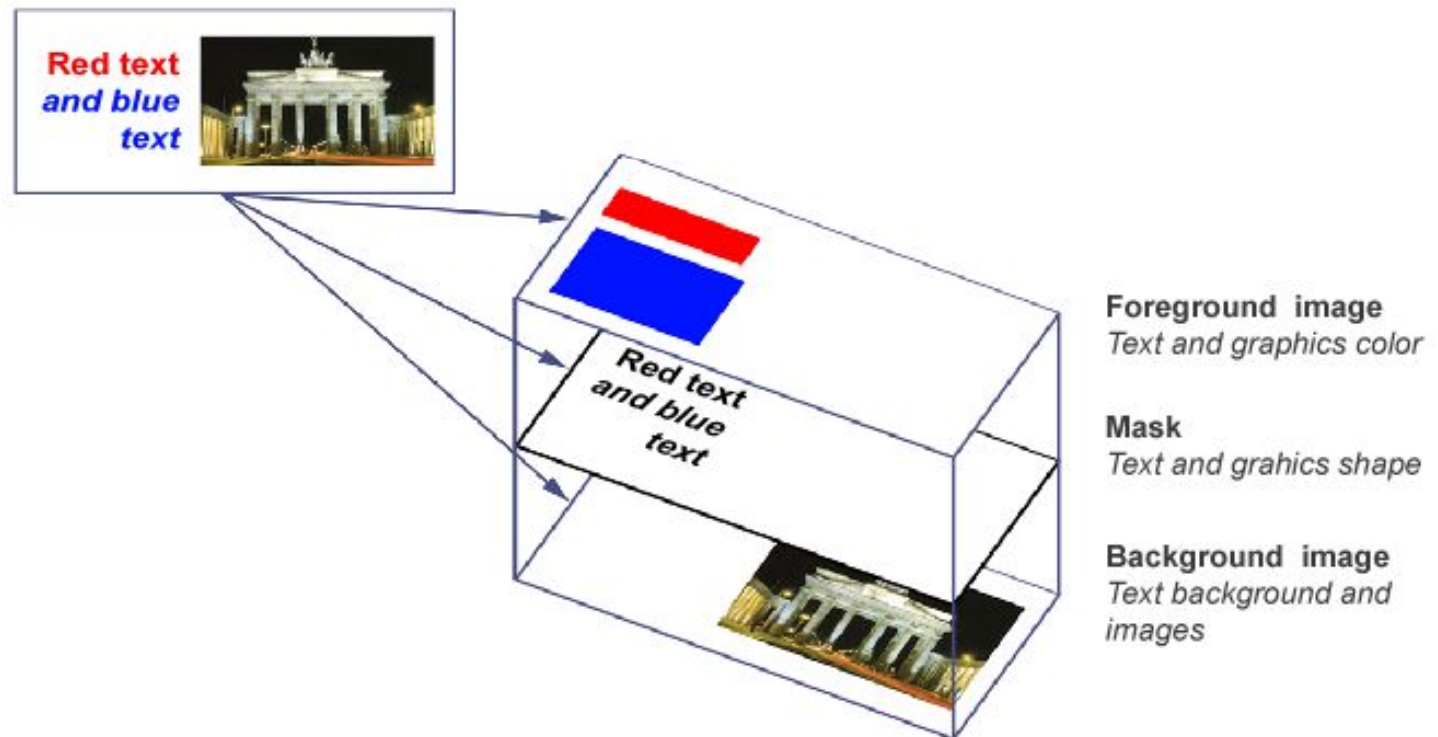
FAX G4	JBIG2/lossless	JBIG2/lossy
60 kB	46 kB	29 kB

PDF/A – Enhanced Compression

▷ **For color documents**

  ▷ **MRC-compression, also known in JPEG2000 (JPM)**

  ▷ **Splitting documents in three layers, to be compressed independently and stored in PDF/A**

Foreground image
Text and graphics color

Mask
Text and grahics shape

Background image
Text background and images

PDF/A – Enhanced Compression

▷ **For Color Documents**
  ▷ **Extreme compression, fully legible**
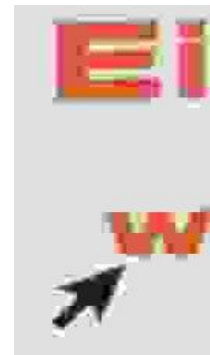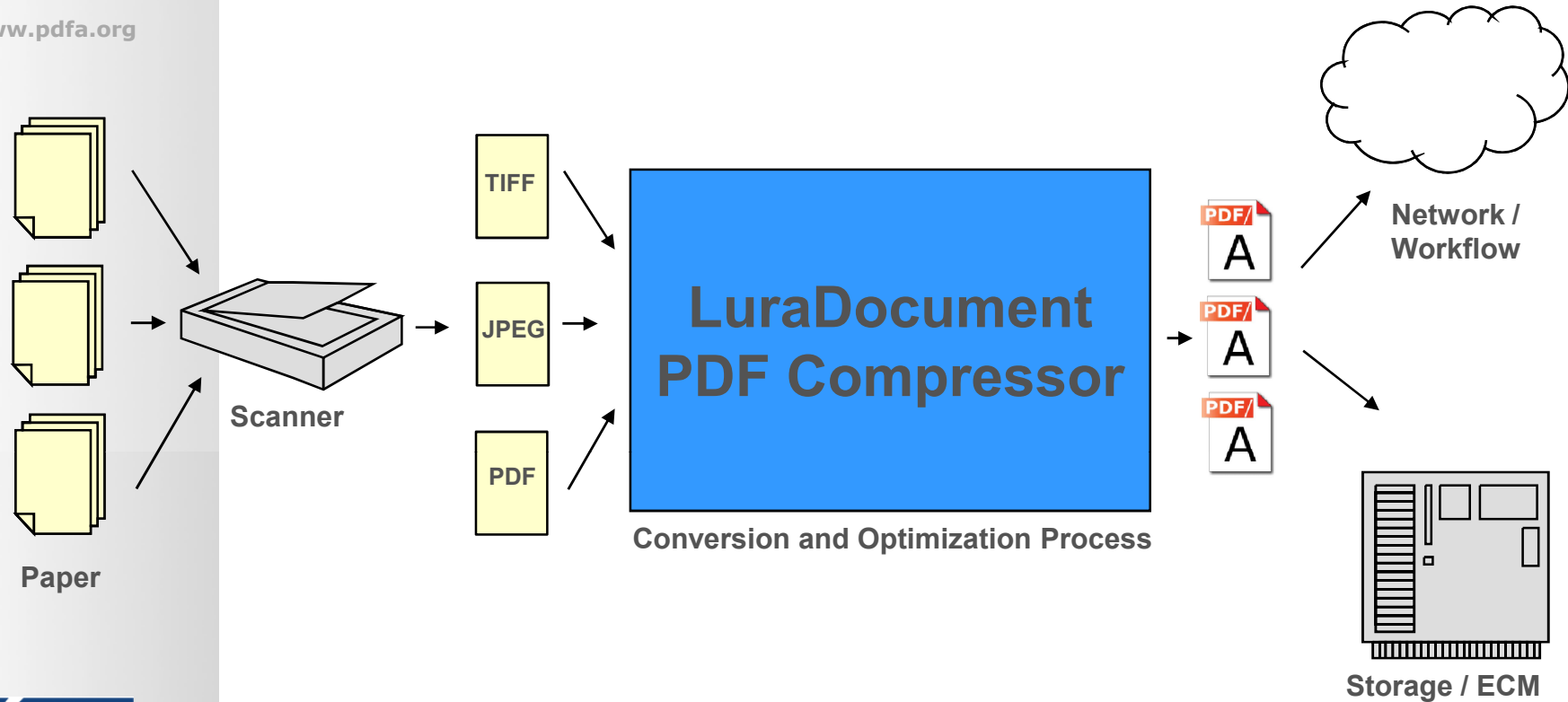 ▷ **Saves the color and the visual quality**

TIFF	TIFF G4	JPEG	PDF/A
23,8 MB	**60 kB**	**180 kB**	**65 kB**

PDF Compressor Basics: How it works

Paper → **Scanner** → TIFF / JPEG / PDF → **LuraDocument PDF Compressor** (Conversion and Optimization Process) → PDF/A → Network / Workflow, Storage / ECM

www.pdfa.org

Carsten Heiermann
Managing Director

▷ **Convert Scanned documents**
  ▷ **Batch conversion**
  ▷ **„unattended"**
  ▷ **Fully automated**

Demo

▷ **Armin, let's have a look!**

Question:

▷ **PDF/A: hype or the future archiving format?**

Carsten Heiermann
Managing Director

PDF/A – Example Credit Files

▷ **Mailroom for credit files and international checks**

▷ **Example: HeLaBa (German State Bank) Mailroom**
  ▷ **168 Billion Euro balance sheet total**
  ▷ **5.700 employees**

▷ **Project Outline**
  ▷ **Convert 20 Mio. Pages paper based archive to PDF/A**
  ▷ **Convert all daily incoming mail to PDF/A**
  ▷ **Create complete electronic credit files**
  ▷ **Used tools: LuraTech PDF Compressor, Kofax Ascent, EMC Centera, Wincor Nixdorf archive:net (Taxnet)**

▷ **Results**
  ▷ **Full color scans in electronic archive**
  ▷ **High compressed PDF/A files**
  ▷ **Full text searchable credit files**
  ▷ **Long term readability of credit files**
  ▷ **First step on the way to single archiving format**

Carsten Heiermann
Managing Director

PDF/A – Example eGovernment

- ▷ **Resident registration files and construction files**

- ▷ **Example: Long term archiving at City of Erlangen**
  - ▷ **103.000 citizens, more than 70% internet access**
  - ▷ **eGovernment-Center initiative**

- ▷ **Project Outline**
  - ▷ **First: Convert paper resident registration to PDF/A**
  - ▷ **Second: Convert Construction files to PDF/A**
  - ▷ **Third (plan): Use PDF/A for all digital files**

- ▷ **Results**
  - ▷ **PDF/A is suitable for mass wise (smaller) documents and also for large documents like technical drawings (serveral 100 MB raw data)**
  - ▷ **High compressed PDF/A files reduce storage costs and bandwidth needs**
  - ▷ **Long term readability of all files**
  - ▷ **Files are now available quickly for daily research**

Carsten Heiermann
Managing Director

Serveral Billion pages to be preserved between 10 and more than 100 years

- APO-Bank (D)
- West LB (D)
- Bank Julius Baer (CH)
-
- Sparkassen Informatik (D)
- Swiss RE (CH)
- DAK Versicherung (D)
- Universa Versicherungen (D)
- AOK (D)
- OCE (NL/D)
- Southern Nuclear (USA)
- RWE Energie (D)

- Siemens (D)
- Vattenfall (D)
- Airbus (D)
- Blohm & Voss (D)
- Bosch Rexroth (D)
- International Labor Organization (CH)
- British Library (UK)
- Harvard Library (USA)
- Het Utrechts Archief (NL)
- Internet Archive (USA)
- Only Luratech's reference customers...

Carsten Heiermann
Managing Director

www.pdfa.org

Thanks a lot for your interest!

Please fill out our questionnaire.

Demo software or more information?

c.heiermann@luratech.com

Carsten Heiermann
Managing Director