www.pdfa.org

Paper To PDF

- Paper becomes digital

Leonard Rosenthol, PDF Architect
Adobe Systems

Adobe

Leonard Rosenthol
PDF Architect

Existing Solutions for Scanned Documents

- Wide Variety of Formats
  - black/white: TIFF G4
  - Color: JPEG, JPEG2000 or PNG
  - Special version formats like „JPEG in TIFF“

- Disadvantages:
  - Too many formats – most not standardized
    - BE AWARE: TIFF is proprietary! (from Adobe!)
  - Loss of information
  - Bad image quality and huge file sizes
  - No full text searchability (OCR) inside files
  - Inconsistent use of metadata standards

Leonard Rosenthol
PDF Architect

Existing Solutions for Scanned Documents

- **Bad image quality vs. file size**

TIFF or PNG

23,8 MB

JPEG

180 kB

TIFF G4

60 kB

Adobe

Leonard Rosenthol
PDF Architect

Alternative Solution: PDF

- Open International Standard (ISO 32000-1)
- Supports B&W and Color images
  - Lossless and lossy compression incl. modern options such as JBIG2 and JPEG2000
- Text can be included for search/index
- Open Standard metadata (XMP: ISO 16684)

- Conclusion: PDF has none of the disadvantages of the legacy formats – and many advantages!

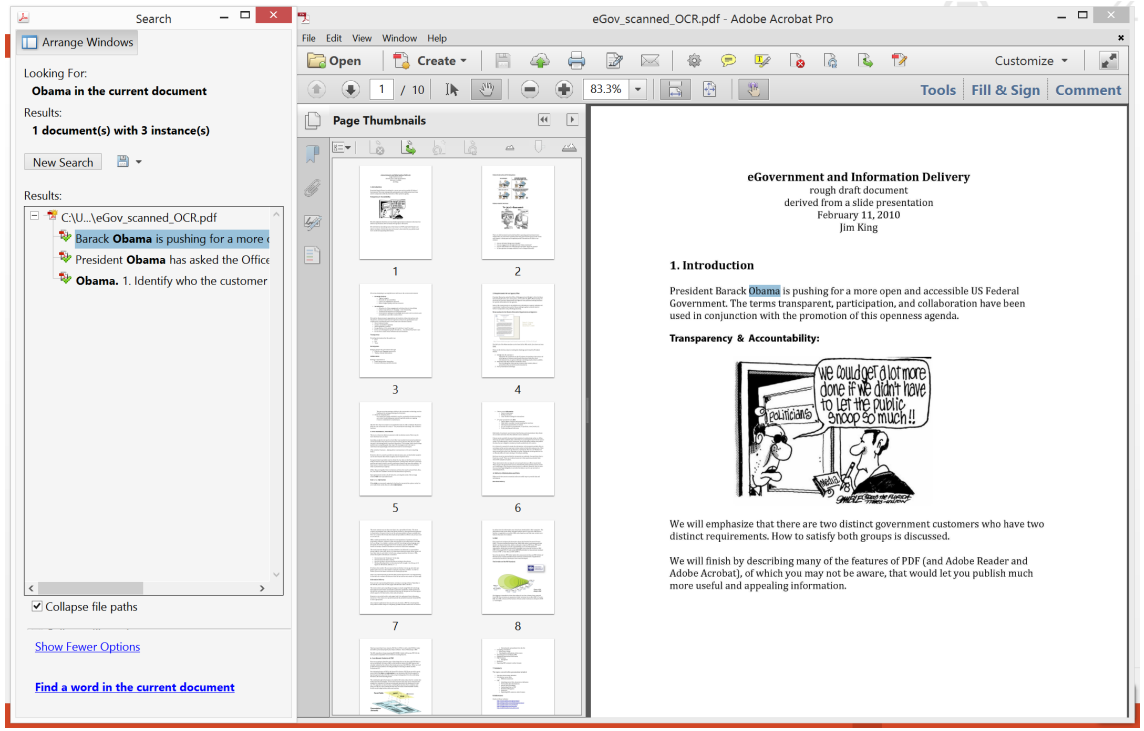
Adobe

Leonard Rosenthol
PDF Architect

PDF – full text searchability (OCR)

www.pdfa.org

- **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,...**



Adobe

Leonard Rosenthol
PDF Architect

PDF – Enhanced Compression

- **For black/white documents**
  - **JBIG2 - ISO/IEC 14492**
    - Created as alternative-successor to TIFF G4
    - Full and visual lossless mode
    - Also supported by PDF/A, available in Adobe Reader

FAX G4	JBIG2/lossless	JBIG2/lossy
		
60 kB	46 kB	29 kB

Adobe

Leonard Rosenthol
PDF Architect

PDF – Enhanced Compression

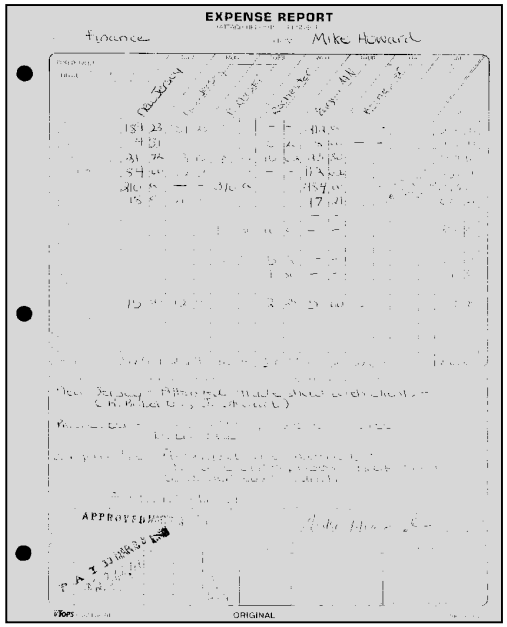
- **Color is not just "nice to have", but enhances employees productivity and saves money**
  - **Employees work smarter with color documents**
    - 14% better recognition of documents
    - 70% faster decisions
    - 80% enhanced reading correctness
  - **Saves 75% costs for scanning**
    - No need for manual sorting prior to scanning
    - No different software and procedures
    - No different scanner settings
    - Less Rescans

Adobe
Leonard Rosenthol
PDF Architect

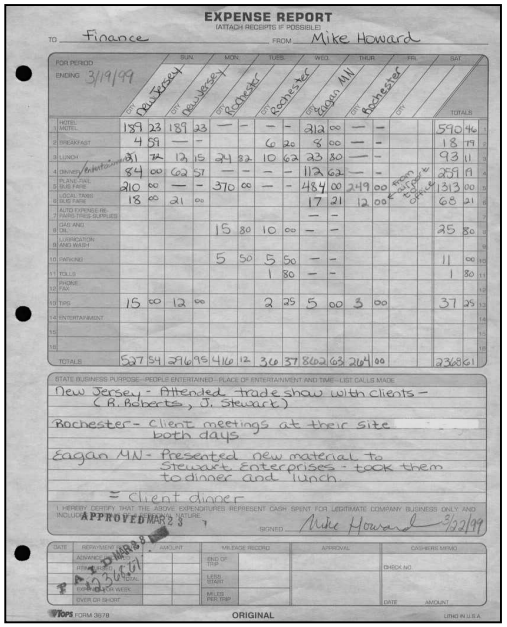
PDF/A – Enhanced Compression

- **Compare the results:**

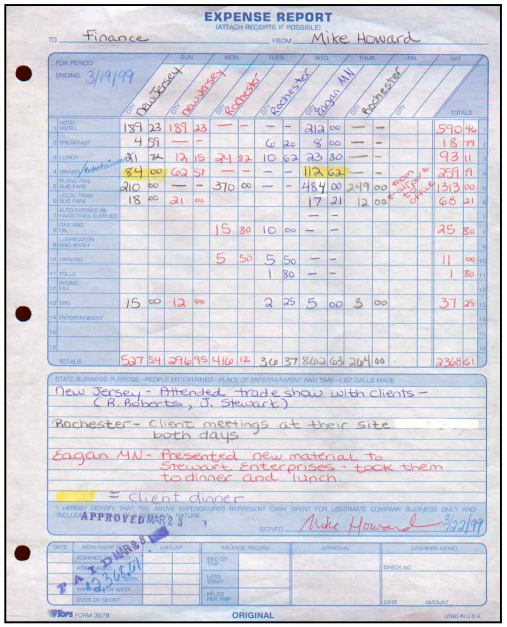
bitonal



grayscale



color

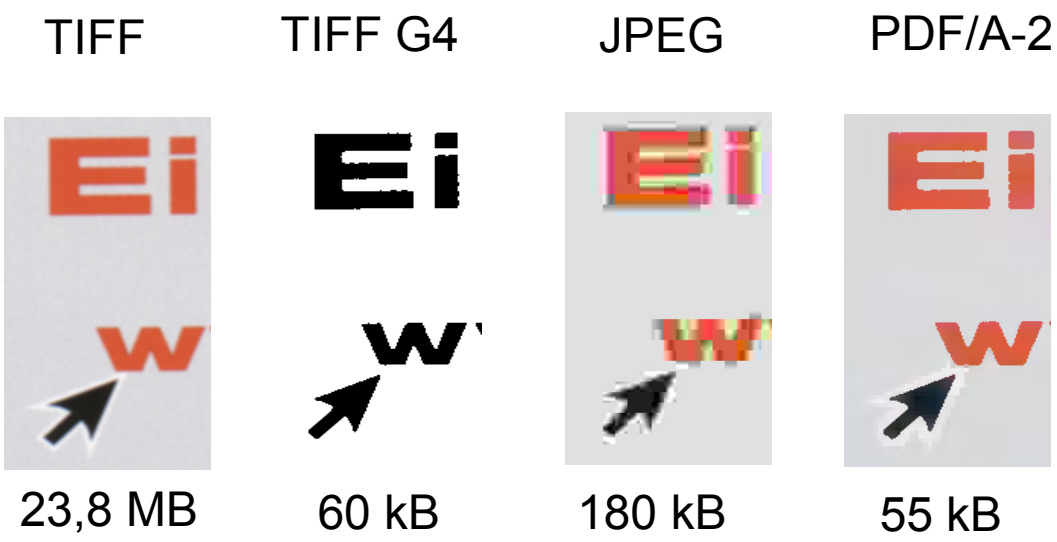


Adobe

Leonard Rosenthol
PDF Architect

PDF – Enhanced Compression

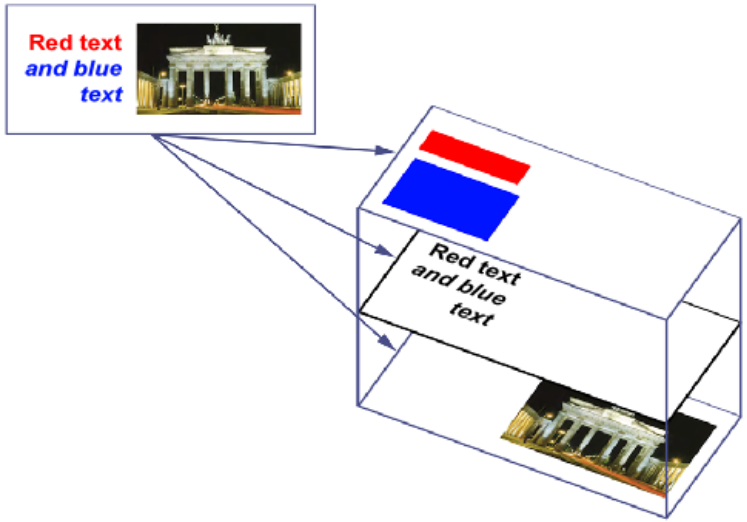
- For Color Documents – lots of options!
  - Flate/ZIP
  - TIFF
  - JPEG
  - JPEG2000



Adobe
Leonard Rosenthol
PDF Architect

PDF – Enhanced Compression

- MRC-compression (Mixed Raster Content)
- Splitting documents in three layers
  - Each compressed independently
- PDF/A-2 adds: JPEG2000, Unicode and layers

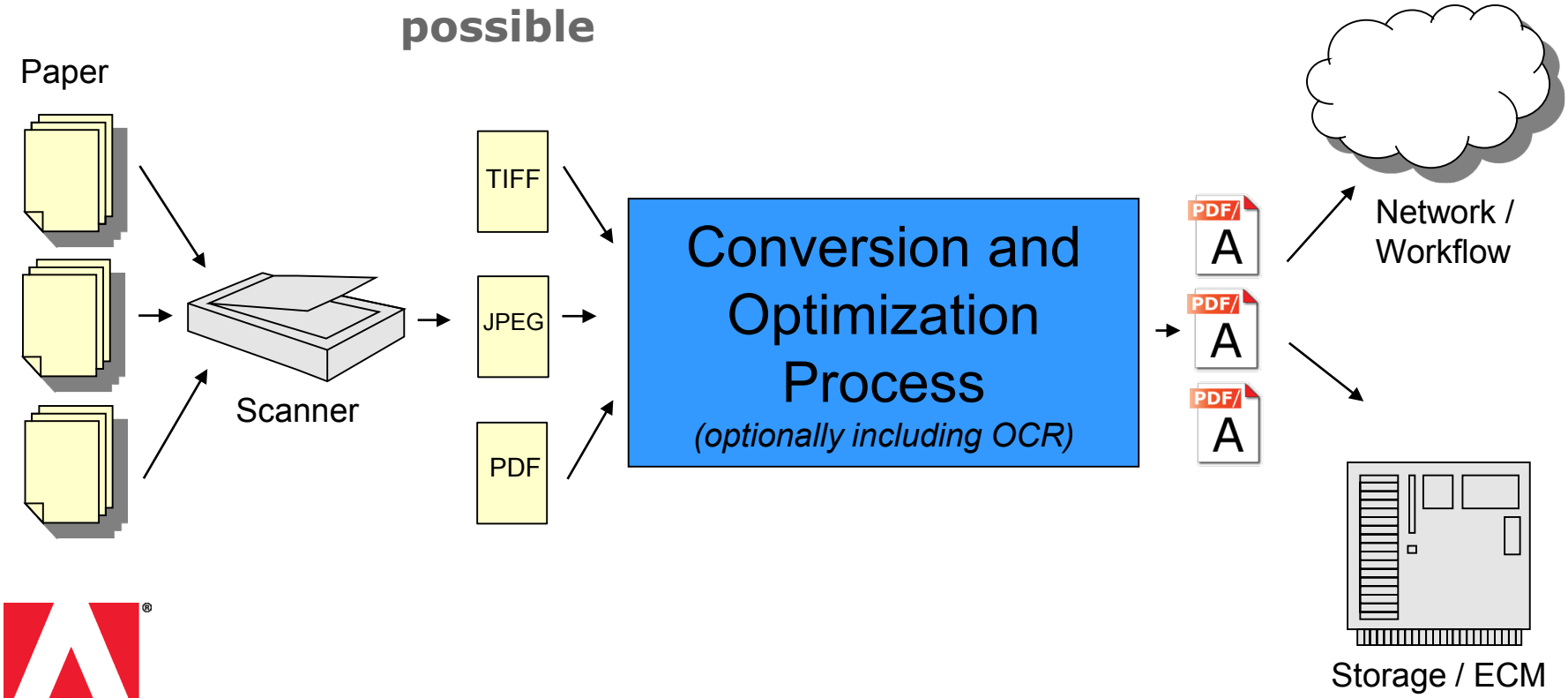


Layer	PDF/A-1	PDF/A-2
Text Color Foreground	JPEG	JPEG JPEG2000
Text b/w Mask	TIFF G4 JBIG2	TIFF G4 JBIG2
Color Background	JPEG	JPEG JPEG2000

Leonard Rosenthol
PDF Architect

How to convert scanned documents into PDF

- **Full automated unattended processing is possible**



Adobe
Leonard Rosenthol
PDF Architect

Example Credit Files

- Mailroom for credit files & international checks
- Example: HeLaBa (German State Bank)
  - 168 Billion Euro balance sheet total
  - 5.700 employees
- Project Outline
  - Convert 20 million page, paper based archive to PDF/A
  - Convert all daily incoming mail to PDF/A
  - Convert incoming E-Mail to PDF/A
  - Create complete electronic credit files
- Results
  - Full color scans in electronic archive
  - High compressed PDF/A files
  - Full text searchable credit files
  - Long term readability of credit files
  - First steps on the way to single archiving format

Adobe

Leonard Rosenthol
PDF Architect

PDF – 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

Adobe
Leonard Rosenthol
PDF Architect

PDF – Summary for scanned documents

- **Scan to PDF (or PDF/A) with optional OCR is straight-forward and doable with off-the-shelf tools (commercial and open source)**
- **Reasons to do it are indisputable**
  - **because it's a well defined open standard**
  - **because of Searchability (embedded OCR+metadata)**
  - **because of Higher Compression**
    - JBIG2 in b/w scanning = 40 – 60% smaller files
    - MRC in color scanning = 10% larger as TIFF G4 of same document

Adobe

Leonard Rosenthol
PDF Architect

Scan to PDF/A with OCR

Thanks a lot for your interest!

Your questions?

Adobe

Leonard Rosenthol
PDF Architect