

Best Practices For Converting Paper to Digital Documents

Carsten Heiermann

**CEO Foxit Europe &
Board Member PDF Association**



Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



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, huge file size, e.g. with JPEG**
 - **No standardized metadata spread over all formats**
 - **No full text searchability (OCR) inside files**
 - **Thus no “big data” nor simple search**
 - **0% accessible documents**

 foxit

Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



PDF - the better way

- **ISO standards**
 - **ISO 32000**
 - **PDF 2.0**
 - **Applicable special formats:**
 - **PDF/A – ISO 19005**
 - **“Archiving”**
 - **PDF/UA – ISO 14289**
 - **“Universal Access”**
- **Features for scanned documents**
 - **“One format fits all”**
 - **Advanced compression**
 - **Searchability**
 - **508 compliance**




Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



PDF - Advanced Compression


- **For black/white documents**
 - **JBIG2 - ISO/IEC 14492**
 - Created as alternative/successor to TIFF G4
 - Full and visual lossless mode
 - Embedded in PDF, available in readers

FAX G4



60 kB

JBIG2/lossless



46 kB

JBIG2/lossy



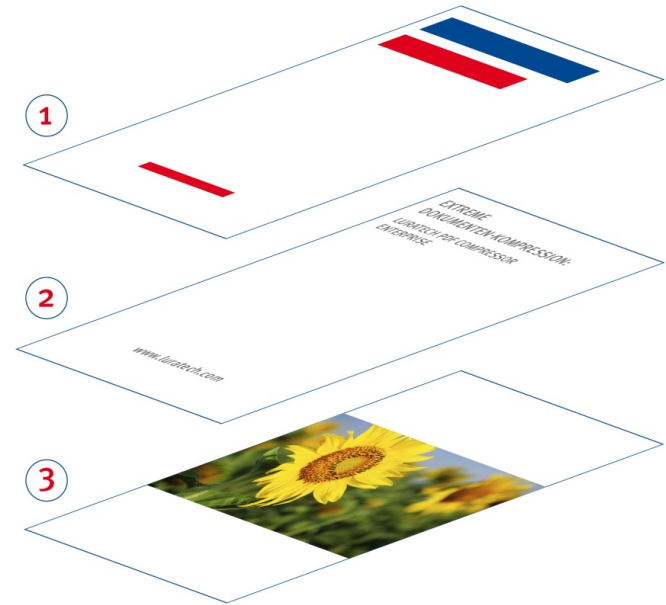
29 kB



Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



- **Color: segmentation**



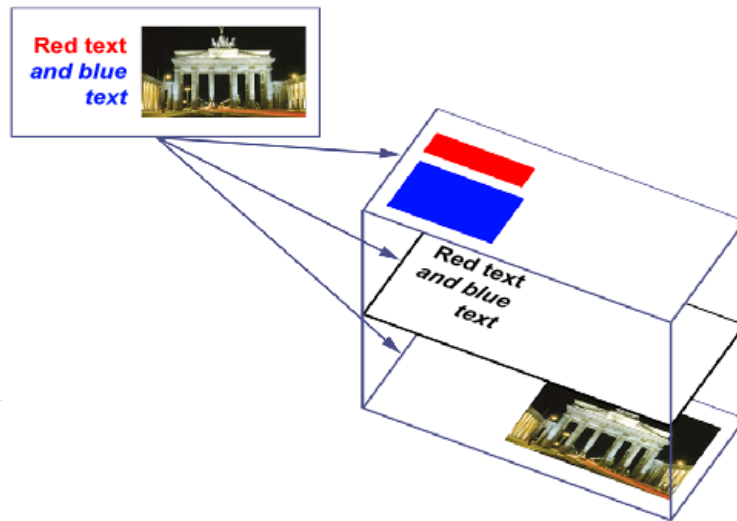
- 1 **Foreground image**
Color information for text and images
- 2 **Mask**
Text and image portions
- 3 **Background image**
Background and images



Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



- **Color: encoding**
 - **MRC-compression (Mixed Raster Content)**
 - **Splitting documents in three layers, to be compressed independently and stored in PDF**
 - **PDF/A-2 adds: JPEG2000, 2u 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



Carsten Heiermann
 CEO Foxit Europe &
 Board Member PDF Association



PDF - Advanced Compression

- **Color: File size**
 - Extreme compression, fully legible
 - Saves the color and the visual quality

TIFF



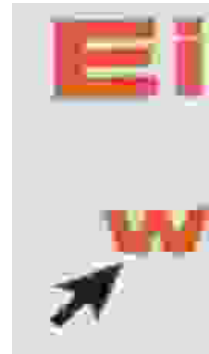
23,8 MB

TIFF G4



60 kB

JPEG



180 kB

PDF/A-1



65 kB

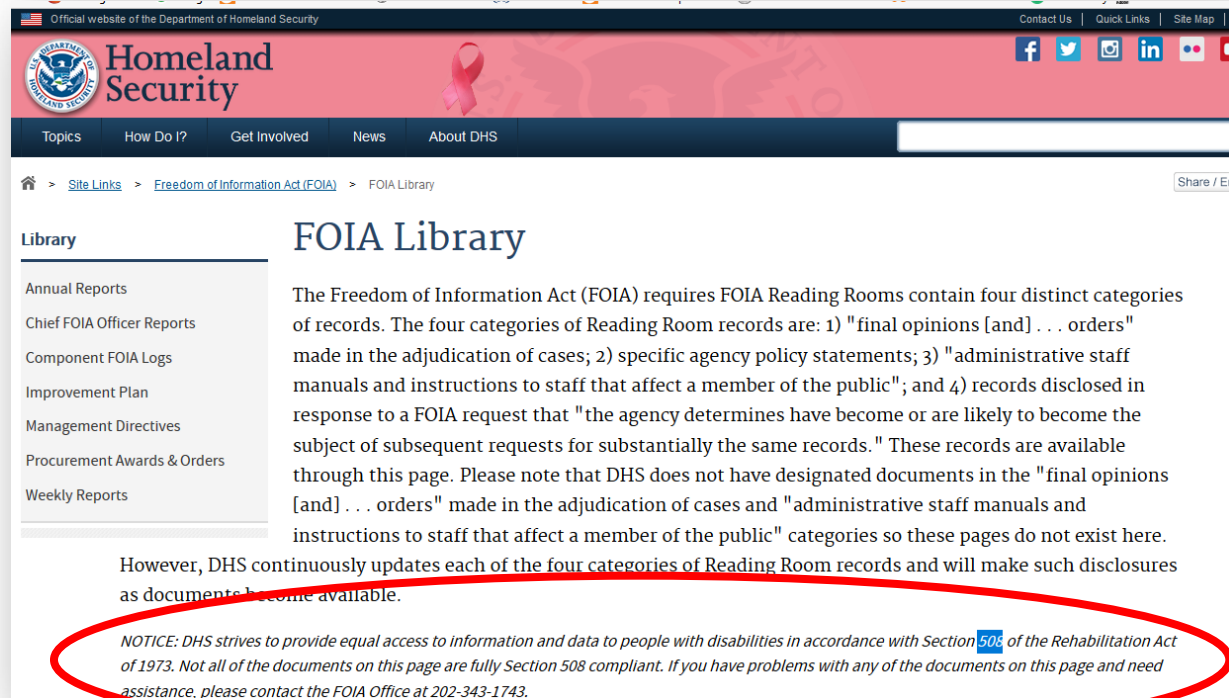
PDF(/A-2)



55 kB



- Agency responsibility
- Non-compliance with Section 508 of the Rehabilitation Act should not be used as a justification by agencies to remove documents from their FOIA websites or refuse to post information proactively



The screenshot shows the official website of the Department of Homeland Security (DHS) for the Freedom of Information Act (FOIA) Library. The page title is "FOIA Library". The main content area contains the following text:

The Freedom of Information Act (FOIA) requires FOIA Reading Rooms contain four distinct categories of records. The four categories of Reading Room records are: 1) "final opinions [and] . . . orders" made in the adjudication of cases; 2) specific agency policy statements; 3) "administrative staff manuals and instructions to staff that affect a member of the public"; and 4) records disclosed in response to a FOIA request that "the agency determines have become or are likely to become the subject of subsequent requests for substantially the same records." These records are available through this page. Please note that DHS does not have designated documents in the "final opinions [and] . . . orders" made in the adjudication of cases and "administrative staff manuals and instructions to staff that affect a member of the public" categories so these pages do not exist here.

However, DHS continuously updates each of the four categories of Reading Room records and will make such disclosures as documents become available.

NOTICE: DHS strives to provide equal access to information and data to people with disabilities in accordance with Section 508 of the Rehabilitation Act of 1973. Not all of the documents on this page are fully Section 508 compliant. If you have problems with any of the documents on this page and need assistance, please contact the FOIA Office at 202-343-1743.



PDF - accessibility

- **Measures around content need to be taken**
 - **Introduce alternative (descriptive) text to images**
 - **508 has more samples on “content hurdles to overcome”**
- **Measures around structure need to be taken (most part of the work)**
 - **PDF pages are not coded linear, but object oriented**
 - **Objects like text, tables, images -> building blocks**
 - **Dictionaries, which object where and how to form a page**
 - **Non human readable structure!**
 - **But a structure to “organize a view”**
 - There is no difference in drawing the lower right corner first in full page view
 - In a screen reader it makes a difference between all or nothing!

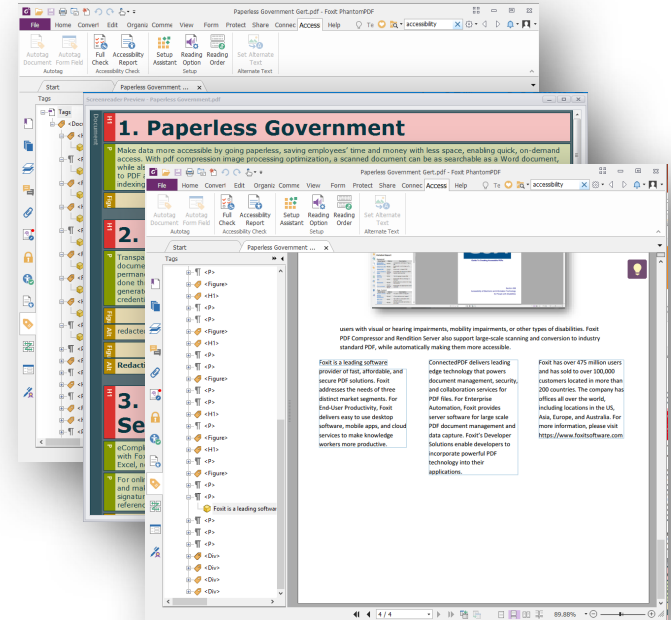


foxit

Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



- **Big thing to do: add the "reading/consuming structure"**
- **Extra "logical layer" in a PDF file**
 - **Called "tagging"**
 - **Equal to give a "meaning" rather than just an "appearance"**
 - **Samples:**
 - `<H1>` tags a line of large bold font as level one headline
 - **Same way:**



- Running title on book page: "artifact" (screen readers skip)
- Paragraph of text. Easy?! How about set in three columns
 - Screen Readers: Don't read the first line, then second, but use tags to follow the reading order column by column!

- **Need to put our experience in code (tag)**



Carsten Heiermann
 CEO Foxit Europe &
 Board Member PDF Association



- **If no structure available**
 - **Re-engineering the authors intent**
 - Never as accurate or easy as getting it right in the first place
 - Think about a newspaper front page (to be continued on page 7...)



Continued on page 3...

- **Layout Recognition Engine**
 - Technically, complex algorithms in software
 - Adds structure to unstructured pages as good as can
 - High accuracy achieved, but not 100%
 - 80/20 rule for automation/manual or achievement/effort
 - Last mile: manual work, if that level is needed
 - Useful toolset (LRE, checker, good tag editor)



Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



- **Scanned documents**
 - **Extra layer of complexity**
 - It is all just an image of a full page with zero structure
 - **First step:**
 - OCR (using dictionaries and confidence level – character, word, sentence)
 - Segmentation (image areas, text areas)
 - Not as 100% accurate as a “Word-text”, but close
 - **Second step:**
 - Layout analysis, again: Re-engineering the authors intent
 - Header, footer, headlines, chapter, paragraph:
 - can be done well
 - Images, detected, tagged as “image”:
 - but can’t “read” the describing text
 - Reading order:
 - documents, books – good
 - “Newspaper front page” – challenge



- **Scanned documents:**
 - **Derive a meaningful(!) tag structure**
 - Don't fulfill checklist criteria using shortcut (tagged image "scanned page")
 - Make the content really accessible
 - 80/20 rule for automation/manual or achievement/effort
 - Last mile: manual work, if that level is needed
 - **Feedback from visually challenged customers:**
 - Resulting PDFs of the automated processor:
 - Very useable!
 - Tested on different screen readers, braille devices – first time that it works well!
 - **Use the automated step, create accessible scanned PDFs!**
 - Low cost per document, as no labor is involved
 - Bumping it up to 100%
 - Normally not affordable for large batches of scanned documents (old Archive,...)



Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



PDF - best practices for scanned documents

- ✓ **Color scanning**
- ✓ **MRC based high compression**
- ✓ **Searchability (use OCR)**
- ✓ **PDF/A**
- ✓ **Auto tag: accessibility for mass conversions**
- ✓ **Check accessibility and fix up for important documents**



-> Look for tools with the right feature sets!

Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association



Thank you! **Questions?**

Get in touch: c_heiermann@foxitsoftware.com
Web site: www.foxitsoftware.com



Carsten Heiermann
CEO Foxit Europe &
Board Member PDF Association

