ARCHIVING INFORMATION WITH PDF

A Brief Introduction

PRESENTED BY

Matt Kuznicki

Chief Product Officer | Datalogics

Chairman | PDF Association

Agenda

Archiving Information With PDF

- **PDF and Information Interchange**
  - What archiving means in a PDF context
- **PDF/A: PDF for Long-Term Archiving**
  - Versions of PDF/A
  - Benefits of using PDF/A
- **User and Usage Examples**
- **Creating PDF/A Files**
- **Verifying PDF/A Files**

PDF and Information Interchange

What is PDF?

PDF

Portable Document Format

What is PDF?

A series of open standards describing digital documents

- 25 years of continual improvements and expanding software ecosystem
- Typically known for graphical descriptions of page appearances

A portable container of visual information allowing

- Embedding resources required to view and process documents
- Interactive elements: page markup, digital signatures, audio, video, 3D mode
- Conventions for reliable structured content extraction

PDF allows different trade-offs between conciseness and expressiveness

Archiving in a PDF Context

Standardized, open standards based content representation – PDF:

- Is royalty-free and open for any and all implementers without IP or other restrictions
- Enjoys a large community of implementers, creators and processors
- Allows the construction of digital documents in a format that will exist and be readable many years / centuries into the future

PDF does not cover physical storage of the file or other (very important) matters for archiving outside this scope

PDF: Reliable Document Interchange

- PDF is the leading format for reliable document interchange
- Designed from the start to facilitate reliability across different readers and through time
- But PDF is a vast set of capabilities, different workflows and producers use different capabilities and write PDFs differently
- How to condense everything one can do in PDF into the most useful set for archiving?

PDF/A: PDF for Long-Term Archiving

PDF/A – PDF Standards for Archiving

For the most reliable information archiving, PDF/A was created

ISO 19005: Electronic document file format for long-term preservation

A set of ISO standards that provide for PDF creators and producers

- Requirements to ensure presence of resources needed for document processing
- Restrictions on PDF syntax and features for greater compatibility across implementers
- Requirements for PDF creators and processors for more standardized behaviors

A collection of best practices, technical notes, and community support

Versions of PDF/A

As PDF has evolved, so has PDF/A:

- PDF/A-1 is based on PDF 1.4
- PDF/A-2 is based on ISO 32000-1 and allows including other PDF/A-2 files
- PDF/A-3 is based on ISO 32000-1 and allows including other files of any type
- PDF/A-4 upcoming will bring maximally reliable archiving to PDF 2.0 (ISO 32000-2)

PDF/A flavors to best suit different workflows:

- a – structured content extractability requirements
- b – focused on reliable visual archiving and rendering
- u – includes information to improve text retrieval by machine processes

Benefits of PDF/A

Key benefits over unrestricted PDF:

- More reliable interchange than PDF alone
- More reliable text, metadata and content extraction than PDF alone
- Greater portability across different viewers and processors
- More reliable rendering across different viewers than PDF alone
- Most reliable interpretation of included contents across different PDF processors

User and Usage Examples

Some Users of PDF/A

PDF/A use in government and industry is prolific.

Examples:

- US Courts: PDF/A for case management and for electronic case files submissions [1]
- Library of Congress: PDF/A is a preferred electronic format for page-oriented digital documents [2]
- ZUGFerd: German PDF/A based standard for human readable and machine readable invoices [3]

[1] https://www.pacer.gov/announcements/general/pdfa.html

[2] https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml

[3] http://www.ferd-net.de/zugferd/faq/index.html?changelang=4

Some Users of PDF/A

More Examples:

- US DoD Defense Technical Information Center: PDF/A recommended for electronic document submissions [4]
- US FDA: PDF/A recommended for electronic document submissions [5]
- EU drug information submissions: PDF/A recommended for electronic document submissions [6]

[4] http://www.dtic.mil/dtic/submit/formats.html

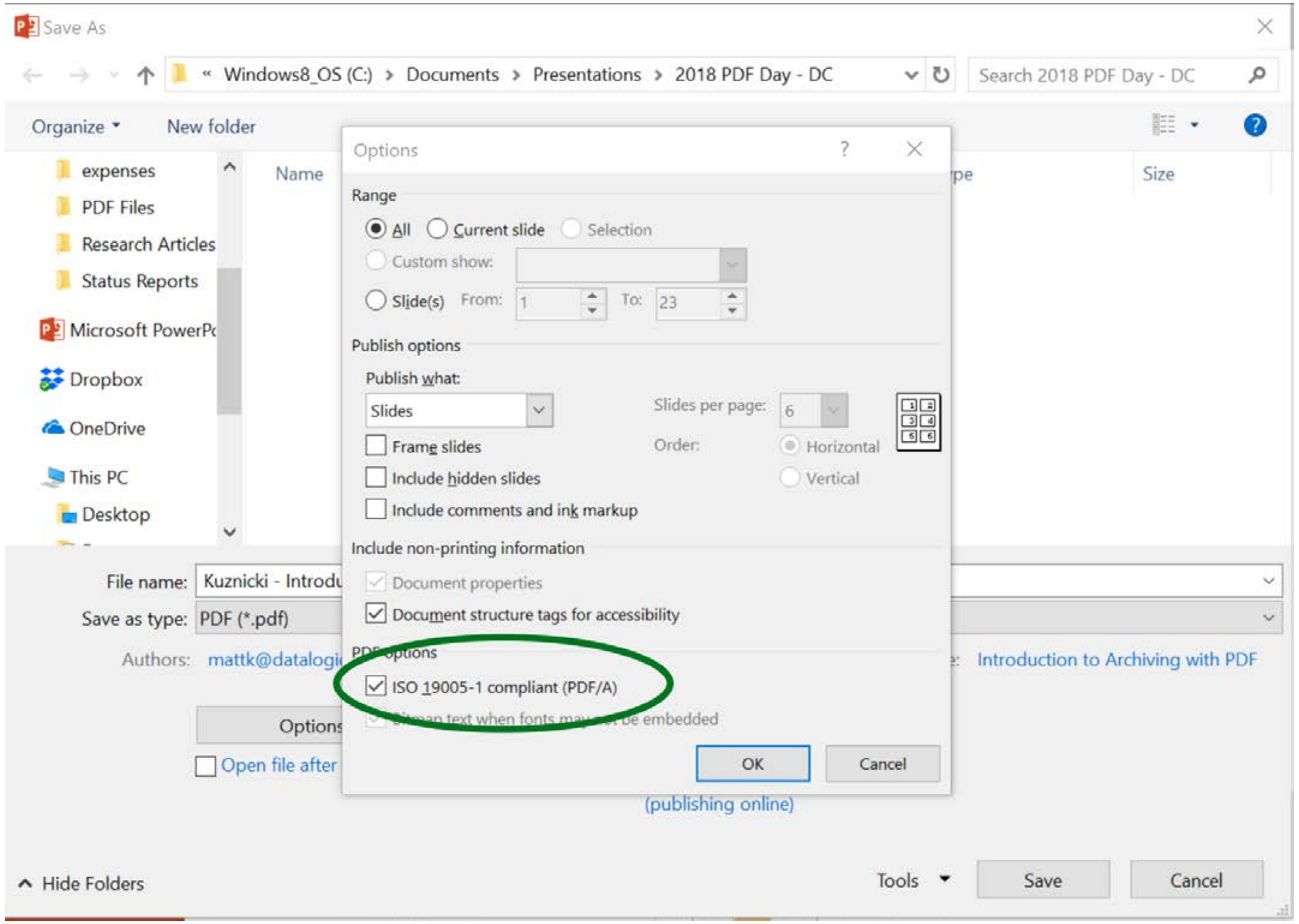
[5] https://www.fda.gov/downloads/ForIndustry/FDAeSubmitter/UCM332980.pdf

[6] http://esubmission.ema.europa.eu/tiges/docs/eCTD%20Guidance%20v3.0%20final%20Aug13.pdf

Creating PDF/A Files

Creating PDF/A Files

Creation support built into Microsoft Office Suite



Save As

Windows8_OS (C:) › Documents › Presentations › 2018 PDF Day - DC

Options

Range

● All ○ Current slide ○ Selection

○ Custom show:

○ Slide(s) From: 1 To: 23

Publish options

Publish what: Slides

Slides per page: 6

Order: ● Horizontal ○ Vertical

□ Frame slides

□ Include hidden slides

□ Include comments and ink markup

Include non-printing information

☑ Document properties

☑ Document structure tags for accessibility

PDF options

☑ ISO 19005-1 compliant (PDF/A)

☑ Bitmap text when fonts may not be embedded

File name: Kuznicki - Introdu

Save as type: PDF (*.pdf)

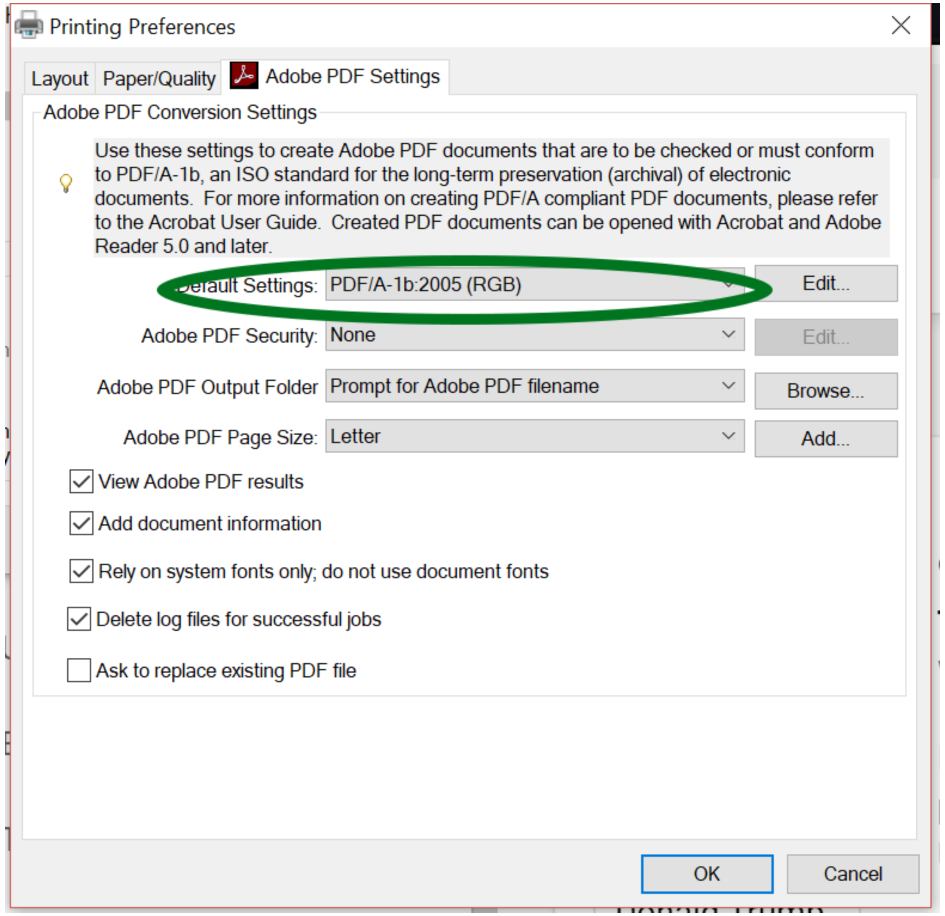
Authors: mattk@datalogi

Title: Introduction to Archiving with PDF

□ Open file after

Creating PDF/A Files

Adobe Acrobat print to PDF/A printer



Printing Preferences

Adobe PDF Conversion Settings

Use these settings to create Adobe PDF documents that are to be checked or must conform to PDF/A-1b, an ISO standard for the long-term preservation (archival) of electronic documents. For more information on creating PDF/A compliant PDF documents, please refer to the Acrobat User Guide. Created PDF documents can be opened with Acrobat and Adobe Reader 5.0 and later.

Default Settings: PDF/A-1b:2005 (RGB)

Adobe PDF Security: None

Adobe PDF Output Folder: Prompt for Adobe PDF filename

Adobe PDF Page Size: Letter

☑ View Adobe PDF results

☑ Add document information

☑ Rely on system fonts only; do not use document fonts

☑ Delete log files for successful jobs

☐ Ask to replace existing PDF file

Creating PDF/A Files

Other examples of support for creation from end-user tools:

- Adobe Creative Cloud & Document Cloud
- Apache OpenOffice
- DocsCorp pdfDocs
- Foxit PhantomPDF
- Nitro Pro
- Nuance Power PDF

PDF/A creation tools for users are prolific and widespread in use

Creating PDF/A Files

Broad ecosystem of tools for programmers and software developers.

Examples

- API/SDK driven programmatic creation:
  - Apache FOP, iText SDK, PDFLib SDK
- API/SDK driven conversion of existing PDF files:
  - Adobe PDF Library, callas pdfToolbox SDK, Datalogics PDF Java Toolkit, PDFTron
- Cloud and Web API creation:
  - ABBYY Cloud SDK, ASPOSE, Qoppa, and more

Verifying PDF/A Files

PDF/A Validation

Free and commercial tools for verifying PDF/A conformance:

- veraPDF - PREFORMA project
- Adobe Acrobat
- callas pdfToolbox SDK
- PDF Tools AG
- SOLID Documents software development kit

Wrapping Up

Summary

Archiving Information with PDF

- Is a key capability baked into the Portable Document Format
- Enables reliable exchange and reading of documents
- Is standardized in the PDF/A series of open standards
- Used in a variety of government and industry workflows
- Is supported by a large ecosystem of PDF viewers, creators and processors

PDF is the accepted choice for reliable archiving of visual documents and information

More Information About PDF/A Software

Find more information about PDF and PDF/A software at the PDF Association website:

https://www.pdfa.org/

Thank You!

Matt Kuznicki

Chief Product Officer, Datalogics
Chairman, PDF Association

datalogics.com | mattk@datalogics.com | +1.312.853.8200