electronic
document
CONFERENCE

I Didn't Know PDF Could Do That!

Exploring the Capabilities of PDF Documents
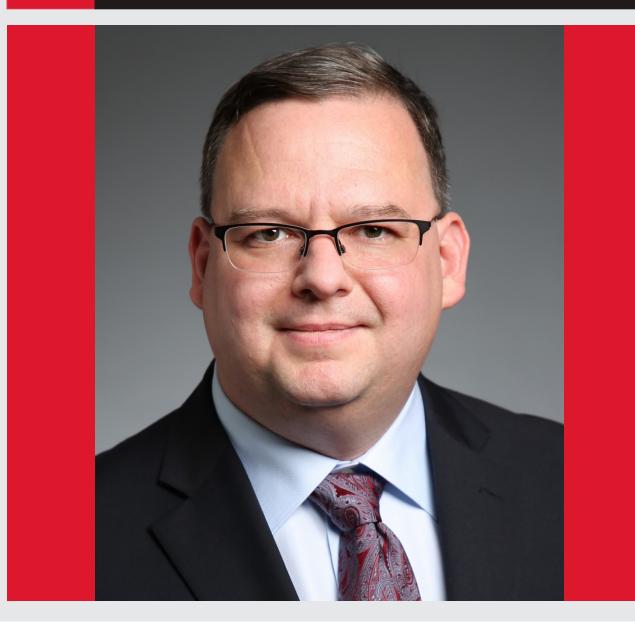
PDF contains multitudes

Let's explore some lesser-known PDF capabilities in the areas of

- Security
- Collaboration
- Interoperability
- Information exchange

Just because PDF can do it… doesn't mean your software can!

About the presenter

Matt Kuznicki

- Chairman of the PDF Association

- Involved in PDF developer community since 2000

- Involved in PDF standards development since 2012

- Interested in promoting widespread adoption of PDF

PDF: "Millions of trees saved"

With more than 10 trillion PDFs estimated to have been produced

If PDF has saved only 10% of these from being printed – and each is only a one page file

That's still a trillion sheets of paper!

PDF: "Millions of trees saved"

What's a trillion sheets of paper?

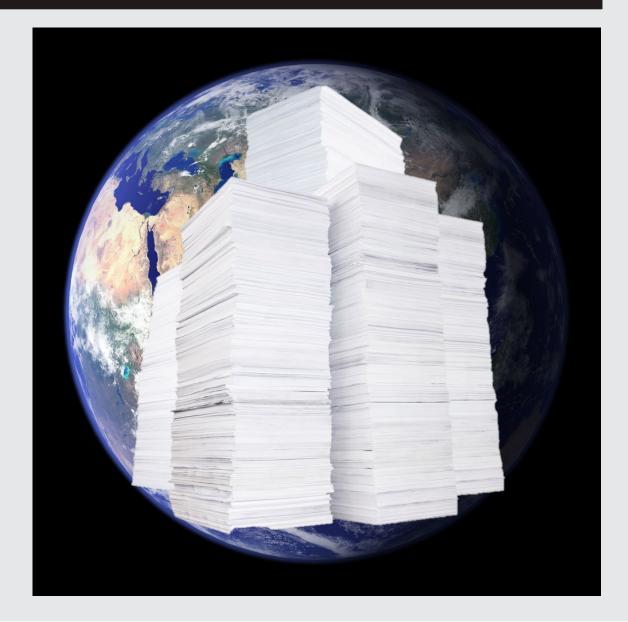That's 50 to 100 million trees worth of paper that have been saved

Enough to cover the area of Seattle and then some!

PDF goes around the world

One trillion pages would circle the Earth more than 2 and a half times!

Combining it all together

We can make one PDF to hold all these unprinted pages

- No longer a 10GB file size limit – we can make a trillion pages or more!
- No longer a 200" x 200" page size limit – we could impose all trillion pages on 1 PDF page
  - On less than 1% of the maximum sheet size of one page
  - Maybe we could impose all 10 trillion PDFs that have been made on one page?

Securing the World

Or At Least the World's Documents

Keeping information safe

PDF document security is standardized and in broad deployment

- Standardized mechanism for password protection allows for implementation interoperability

- Advisory mechanism for specifying user-permitted actions

- PDF 2.0 uses AES 256 security: generally regarded as safe even against quantum computers

But security in PDF includes much more than just password protection

For your eyes only

PDF public key security handlers (section 7.6.5) allow encrypting files so that only select people can open them

- A file can specify any number of specific individuals to grant access

- Access permissions – print, comment, etc. – can be specified in similar manner to standard password security

- X.509 certificate based workflow – broad industry compatibility

Embedded and associated files may also be encrypted – or not – as desired via Crypt filters

Let's keep this between us

Custom security handlers can be created for specialized workflows via Crypt filters (sections 7.4.10, 7.6.6)

- Online verification of access permission
- Digital rights management (DRM) implementations
- Higher-security encryption beyond what's in the PDF specification

PDF 2.0 files may include an unencrypted wrapper document (section 7.6.7) for use with software unable to decrypt the document

Trust but verify

Digital signatures are commonly used to signify agreements – they can also be used to verify document contents have not been altered

- Invisible signatures allows tampering detection without implying a legal agreement

- Document security store (section 12.8.4) and document timestamp dictionary (12.8.5) capture information needed for long-term validation

- Digital signatures include capabilities to specify permissible modifications that do not constitute alteration (such as commenting)

Collaboration

People Working Together

Collaborative commenting

PDF's standardized annotations allow for collaboration on documents

- Defined annotation types and behavior allow for document passing

- FDF (section 12.7.8) and XFDF allow for storing annotations in addition to PDF forms

  - Collaboration without passing documents back and forth – only the FDF with the annotation data needs to be transferred

Additional pages may be carried in the FDF also (12.7.8.3.3) to support adding content from a common document baseline or template

Collaborative versioning

PDF's incremental update capability (section 7.5.6) is well known for allowing small changes to files without rewrites. But did you know…

- Incremental updates can be used to facilitate PDF file versioning?

- Updates do not need to be a linear chain – branched updating is allowed via different /Prev entries in incremental update cross-reference sections

- Incremental updates allow a PDF to target multiple, different PDF specification versions
  - Example: a PDF 1.7 document with an incremental update to PDF 2.0 compatibility

Collaborative versioning

PDF incremental updates allow users to

- Store and access working drafts and revisions of documents in one file

- Pass drafts and revisions around to others for collaboration

- Maximize compatibility for older PDF software that does not support updated versions of PDF

Interoperability

Programs Working Together

PDF as an open data platform

Associated files (section 14.13) and file specification dictionaries (section 7.11.3) allow data files to be embedded in a PDF along with their meaning

- Store source pages or documents for PDF renditions in their native format for import and export from editing workflows

- Associate schemas and data definitions to metadata and other streams

- Provide alternate formats for tagged sections of PDF content

- Connect logical content and tagged PDF elements to data packages

These features combine to allow using PDF as an open data container

Interoperability and collaboration

Standardization of annotation and form data allows for collaboration across different tools from different vendors

- Propose redactions in one user tool with Redaction annotations (section 12.5.6.23), import and apply these in a server workflow or different tool

- Pre-populate form data from user information and preferences; user can open, updates and save in their preferred environment

- Allow multiple users to markup and comment on a common PDF file from different tools – concurrently – allowing for real-time updates

Standards adherence and compatibility

A PDF file can adhere to multiple different PDF versions and standards

- A PDF/A-3 file is a valid PDF 1.7 file and can be viewed by all appropriate software

- A PDF 1.5 file with an incremental update for PDF 2.0 conformance can be processed by both PDF 1.x and PDF 2.x – aware programs

- A PDF file may be a valid PDF/UA and PDF/A file at the same time

Document requirements (section 12.11) allow documents to state specific functionalities processors must support for proper file processing

Information Exchange

Computers Working Together

Combining documents together

Document parts (section 14.12) allows grouping collections of PDF pages together into single units

- Users can combine documents that are different pieces of a finished work together

- Allows for a different page view ordering to be presented for different processors or use cases

- Processors can extract specific collections of PDF pages targeted for specific handling conditions

Semantic content – it's just logical

Logical content (section 14.7) allows a workflow to define its own semantics and apply its structure to PDF content

- Tagged PDF is one implementation of semantic structure – the PDF standard allows for implementers to make their own

- Logical content may use or reference the PDF standard structure types – but is not required to do so. A workflow may structure PDF logical content however is best for its use case

- Structure elements can have attributes from multiple different processors or sets (14.7.6.1) for maximum interoperability

Bringing meaning to data

Tagged PDF is well-known for accessibility enablement – but did you know

- Content and semantics allow for transforming visual pages into alternate forms and formats

- Tagged PDF allows for information repurposing – exchange with other applications

- Rolemapping and namespaces allow for mapping 3rd party tagging definitions into tagged PDF for maximum reusability

Logical Content and Visual Content

Logical content is separated from a PDF's visual appearance

- Logical content is stored separately from visual page appearance
- Logical content can be ordered and segmented differently than the visual page appearance ordering

PDF page appearances designed for exchange with other humans

Logical content is designed for information exchange with other programs

Wrapping Up

Whew! What a Tour!

Wrapping Up

PDF is well known as a reliable platform for document presentation

Less known are PDF capabilities as a secure platform for collaboration, open data exchange and content reuse

Powerful features of PDF support different people, tools from different vendors, and processes with different needs all working together

Let's Work Together
To Work Together!

Thank you!

Reach out!

matt.kuznicki@pdfa.org

Notes and Figures

- 10 trillion PDFs? https://www.pdfa.org/pdf-in-2016-broader-deeper-richer/

- Between 10,000 and 20,000 sheets of paper per tree: https://www.sierraclub.org/sierra/2014-4-july-august/ask-mr-green/how-much-paper-does-one-tree-produce

- 1,000 trees per acre for planting seems a commonplace recommendation, such as in http://www.davey.com/hoa/elements/pdf/2-13%20reforestation.pdf