



PDF Days Europe 2022 | Berlin

What's stopping PDF from ubiquitous acceptance?

How the inclusion of explicit information will empower the PDF standard

Shawn Gaither | Principal Scientist | Adobe, Inc.

A 30-Year Journey



- 1993: Postscript Simulator
- 1995: *TeX* font substitution
- 1997: Capture OCR
- 1999: Palm eBook reader
- 2000: *Make Accessible*
- 2003: Font Recognition
- 2004: Form Field Detection
- 2007: PDF Compare
- 2014: FFD in *Adobe Sign*
- 2015: *Liquid Mode*
- 2018: *Deep Table Model*
- 2020: *Form Fields Model*

Evolution of PDF via ISO Standards

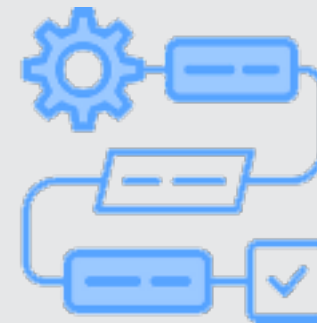


ISO Standard	Index	Applicability
PDF 2.0	32000	Full specification of the PDF language
PDF/X	15930	Prepress for conventional printing of PDFs
PDF/A	19005	Archival of digital PDF documents
PDF/UA	14289	Accessible PDF documents and processors
PDF/E	24517	Engineering PDF documents and exchange
PDF/VT	16612	High volume and personalized printing of PDFs
PDF/R	23504	Raster image and transport

How PDF has Progressed



- *The Camelot Project*
- Interactivity with forms
- Structure and accessibility
- Richer authoring
- Mobile viewing
- Document workflows
- Document analysis



How Authoring has Transgressed



- Original resources `<\001\002\002\003\002> Tj`
- Document structure `missing /StructTreeRoot`
- Style information `missing /ClassMap`
- Document layout `missing << /O /Layout >>`
- Table data and types `missing << /O /Table >>, only /TD, no /AF`
- Formulae components `missing /AF`
- Internal / external links `missing /Link , /GoToE, /GoToDp`

Why is it so Hard to Conform?



- Favoring size, simplicity over content
- Lack of respect for contained data
- Failure to see future possibilities
- Failure to understand PDF specs
- No library support or editability
- *No one else does it so why should I?*
- Lack of accountability [PDF police]



Why Should You Care?



- Explicit Information removes ambiguity/doubt
- Not at the mercy of least common denominator
- Rich information preserves PDF as a valuable format
- No way to know what future workflows may be used
- Learning from past failures helps in the long-term
- Adobe Lessons: *Liquid Mode*, *pdfTools*, *Smart Search*
- We need to ensure the future of PDF as *the* standard



Restore the Original Information

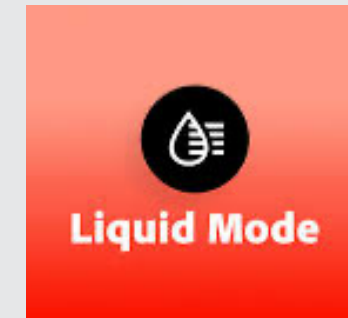


- Use Unicode font encoding `/ToUnicode` stream is a must
- Specify document structure `/StructTreeRoot`
- Use *Namespace* and *ClassMap* `/Namespace, /ClassMap`
- Specify references (*URI*, *Ref*) `/Link, /GoToE, /GoToDp`
- Include original data (*Metadata*, *AF*) `/Metadata, /AF (/Source, /Data)`
- Supplement original data (MathML) `/AF (/Alternative, /Supplement)`
- Augment content for accessibility `/Lang, /ActualText, /Alt, /E`

Progress in PDF Processors and Writers



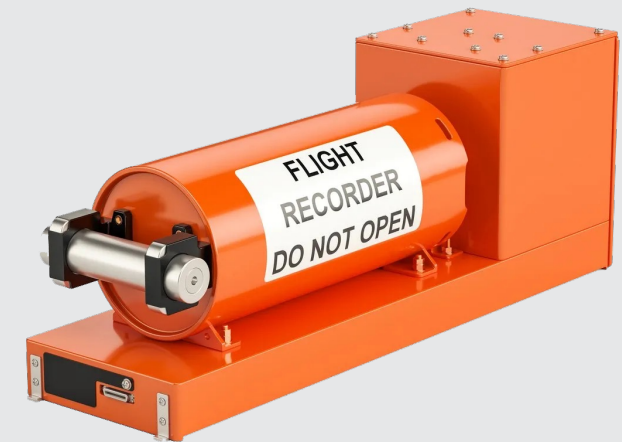
- *Adobe Illustrator* serialization
- *Project Colorado* and *Liquid Mode*
- *LaTeX* tagged and accessible PDFs
- Next generation forms
- *pdfTools* for creation and editing
- Creating Adobe PDFs from *Office*
- *LibreOffice* Hybrid PDFs



Takeaways from this Discussion



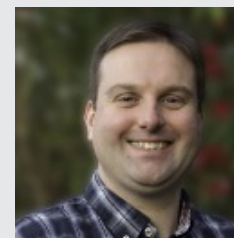
- PDF has grown well beyond rasterization
- Original information is crucial evidence
- PDF workflows are constantly changing
- Do your part to maintain PDF as a standard



Related Talks: *PDF Days Europe 2022*



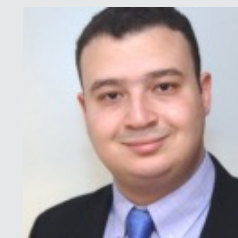
- *Aligning PDF 1.7 and 2.0 through ISO 32005 ...*
- *Learning to Tag*
- *What's holding PDF back?*
- *Tagged and Accessible PDF with LaTeX*
- *PDF Optimization horror-stories*
- *How document understanding can leverage your PDF workflow*
- *Ideas for interoperable self-updating PDF documents*
- *Next Generation Forms for PDF*
- *It's your PDF Association*



Matthew Hardy



Richard Cohn



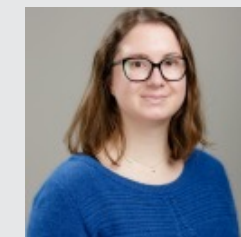
Dr. Tamir Hassan



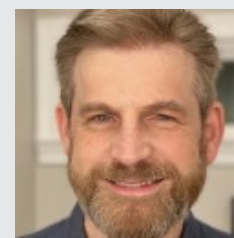
Frank Mittelbach



Patrick Gallot



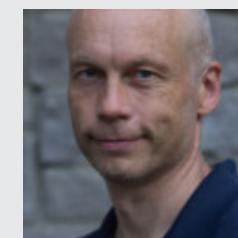
Elodie Tellier



Duff Johnson



Thomas Zellmann



John Brinkman



Mattias Valvekens

Discussion

