An introduction to PDF

Sarah Higgins

Aberystwyth University

Department of Information Studies

sjh@aber.ac.uk

What is PDF?

(Portable Document Format)

- Developed to enable **document** sharing across platforms while retaining “look and feel”
- Originally a proprietary format - Adobe Systems
- Specification available free of charge from 1993
- Became an open standard in 2008
- ISO 32000-1:2008 (PDF 1.7)

PRIFYSGOL ABERYSTWYTH UNIVERSITY

Digital Preservation Awards 2014

Features of PDF

- Displays as intended across platforms and devices due to embedded:
  - Instructions for layout, dimensions and graphics (PostScript based)
  - Fonts
  - Colours and transparency
  - Internal and external links
  - Layers e.g. OCR layer
  - Structural storage system
- Can deal with vector, raster, text, media, interactive forms
- Embedded metadata – header or streams
- Compression - relatively small file sizes
- Security: can be encrypted (owner and user)
- Authenticity: can add digital signatures
- Backwards compatible
- Tagging for content extraction for re-use

Flavours of PDF

PDF	PDF/X	PDF/A	PDF/E	PDF/VT	PDF/UA
V 1.0 – 1.6 proprietary V1.7 = ISO 32000-1: 2008 5 extensions Levels 1, 3, 5, 6 & 8) ISO13200-2 forthcoming	**PDF for Exchange** Based on PDF 1.3, 1.4 & 1.6 ISO 15929 ISO 15930:1 ISO15930:3 ISO15930:4 ISO15930:6 ISO15930:7	**PDF for Archive** Based on PDF 1.4 &1.7 ISO 19005-1 (A-1) ISO19005-2 (A-2) ISO19005-3 (A-3)	**PDF for Engineering** Based on PDF 1.6 ISO 24517	**PDF for Exchange of Variable Data and Transactional (VT) Printing** Based on PDF 1.4 or 1.6 (as restricted by PDF/X-4 and PDF/X-5 ISO 16612-1 ISO16612-2	**PDF for Universal Accessibility** Based on PDF 1.7 (ISO32000-1) ISO 14289-1

Information from Wikipedia https://en.wikipedia.org/wiki/Portable_Document_Format and International Organization for Standardization (ISO) http://www.iso.org/iso/home.htm

Document v. Record

- ISO 15489 defines a the characteristics of a record as
  - Authentic
  - Reliable
  - Has integrity
  - Useable

What is PDF/A?

- A sub-set of PDF for the “Long Term Electronic Preservation of multi-media page documents that may contain a mixture of text, raster images and vector graphics”*.
- Developed by ISO/TC 171/SC2/WG5
- Includes requirements for developing a viewer – how the file should be treated for rendering e.g. font handling, colour management, transparency etc.
- Encapsulates all information needed for display e.g. text, fonts, graphics etc.

*Leonard Rosenthal (2013) What PDF/A is and what PDF/A is not

What is PDF/A?

- Self contained documents
- Predictable and accurate rendering
- Robust – inconsistencies in PDF removed
- Audit trail maintained
- No encryption
- No interactivity (JavaScript, video, 3D)
- Limited colour-space

What PDF/A is not

- A hardware / migration solution
- A complete archiving solution
- An authentication or security system

Flavours of PDF/A

PDF/A-1 (ISO 19005-1:2005)

Supports

- Embedded fonts
- XMP metadata
- Device independent colour-spaces

Forbids

- Encryption
- LZW compression
- External content references
- Transparency
- Multimedia
- Javascript

PDF/A-2 (ISO 19005-2:2011)

Adds support for

- JPEG2000 compression
- Compressed object streams
- Compressed cross reference streams
- Transparency
- Digital signatures
- Embedded PDF/A files (archive quality)
- XML Forms Architecture (XFA) forms
- Content layers e.g. OCR layer
- Forwards compatibility

PDF/A-3 (ISO 19005: 2012)

Adds support for:

Embedded file formats (not archive quality - no guarantee that they can be viewed)

Conformance

4 levels of conformance:

- A – All / Accessible (born-digital material)
- B – Basic Conformance – basic requirements (digitised material)
- U – Unicode conformance

  (PDF A/2 & PDF A/3 - for searching and indexing digitised material)
- U/A – Accessible

Issues for PDF/A

- 3D material (PDF/E not necessarily suitable)
- Authenticity – Migration to PDF/A
  - Is a modification
  - Breaks the digital signature
  - Document v record