



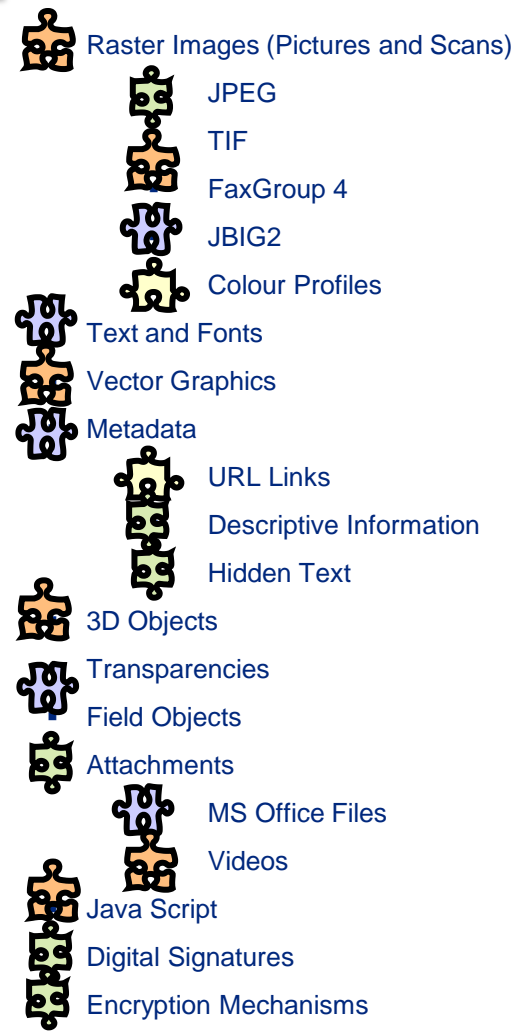
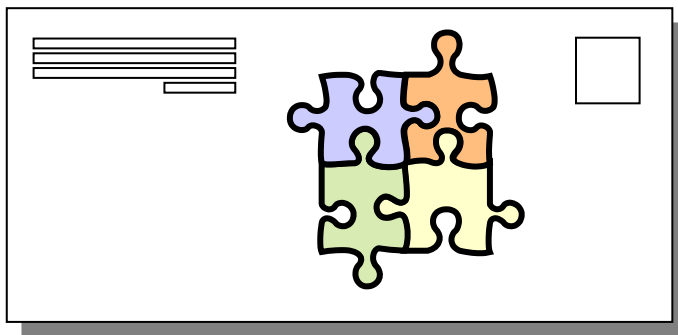
Introducing the PDF/A-3 Standard

Gary Hodkinson

gary.hodkinson@pdfa.org



- What is a **PDF** file?
 - A file format created by Adobe
 - It is a *wrapper*
 - NOT just “electronic paper”
 - It can store many different objects



PDF/A: The Basics

- **What is a PDF/A file?**
 - **The ISO standard for Document Management**
 - **A document file format for long-term preservation (see later)**
 - Secures future ability to read the file
 - Preserves original formatting
- **Why do we need PDF/A?**
 - **Valuable information is being created in PDF form**
 - **There are issues with non-standardised PDF files**
 - Adherence to the PDF syntax varies
 - The Adobe Reader is NOT a PDF Validator
 - Dynamic, interactive PDFs can be created



PDF/A: The Basics

- **PDF/A is:**
 - **Device/ Software/ Versions-independent, i.e. the content is displayed consistently**
 - **Self Contained: A PDF/A-compatible file contains all the components needed to display it**
 - **An evolving standard: additional PDF features are continually being added to the PDF/A standard definition**
- **PDF/A is NOT:**
 - **“Read Only”: this is a common misunderstanding**
 - **A static definition: as new PDF options evolve, they are incorporated into the standard**



PDF/A: The Basics

The ISO Standard family:

- **“PDF/A-1” – released in 2005, ISO 19005-1**
 - Based on PDF 1.4
- **“PDF/A-2” – released in 2011, ISO 19005-2**
 - Based on PDF ISO 32000
 - Allows compression, transparency, Open Type font embedding, digital signatures, Unicode
- **“PDF/A-3” – released 2012, ISO 19005-3**
 - Based on PDF ISO 32000
 - Allows embedded files



PDF/A is currently the best option for long-term archiving

Definition of “long term”:

"The period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository, which may extend into the indefinite future."

Goals of PDF/A

- **To maintain a static visual representation of documents**
- **To provide consistent handing of Metadata**
- **To guarantee future access by remaining transparent**
- **To limit number of restrictions imposed in return**



PDF/A: The Basics

Prohibited content for a PDF/A file:

- Active/Variable content (e.g. JavaScript)
- ~~External content and embedded files~~
- Encryption

In Summary:

- No component of a PDF/A file may jeopardise the consistent format of the document
- The presentation of the document must not rely on external dependencies



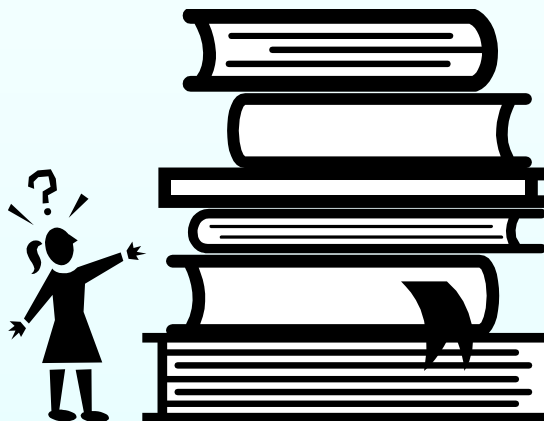
Limitations of Alternative File Formats

- **TIFF, TIFF G4, JPEG (typically scanned images)**
 - Cannot include hidden text and standardised metadata
 - May be of bad image quality (e.g. highly compressed JPEG)
 - Sub-optimal compression algorithms
 - No unified file format possible
- **Digital Documents (MS Office, CAD, XML, PDF)**
 - No unified file format, and multiple versions (Office 95, 97, 2003, 2007, 2010)
 - “Ordinary” PDFs carry risks of inconsistency
 - Formats can easy change (e.g. switch to OpenOffice)
 - Proprietary software vendors may make irreversible changes



Is PDF/A-3 a contradiction?

Native formats should be converted to PDF/A



Now native formats are embedded in PDF/A-3

Distinguish between archive and non-archive



Goals achieved by embedding files in PDF/A-3

- ✓ Save archive ready rendition with source format
- ✓ Save archive ready rendition with machine readable data
- ✓ Save meta data that is inseparable from archive material

Avoid saving archive material as an attachment

Attachments in PDF/A-3

- Embedded files have a relationship to PDF objects
 - Attachment is *Data, Source, Alternative* to/of an object (will be further defined in ISO 32000-2)
- Embedded files are “Associated Files” for:
 - A page, pages or the whole PDF/A document
 - Specific content, e.g. a diagram or text paragraph

For example:

- data.xls is *Data* for a specific diagram
- metadata.csv is *Data* for a number of pages
- text.doc is the *Source* of the PDF/A-3 file
- audio.m4a is an *Alternative* for text content



PDF/A-3 Practical Limitations

- **PDF/A-3 allows for embedding of files of any type**
- **In practice, there are some limitations**
 - **Adobe Reader prevents viewing and/or saving of embedded dangerous files**
 - e.g. .exe, .bat, .cmd, .zip, .gz, .pst, .url
 - Only a registry hack can circumvent that, thus not a practical solution
 - **User: black/white listing**
 - Additional formats to be configured by user
 - Acrobat only opens .pdf by default “without comment”



PDF/A-3 Application: Scanned Documents

- **Metadata is often the most important component**
 - **Metadata is usable without images**
 - **Images without metadata are often not usable**
- **PDF/A-3 offers a complete solution**
 - **Documents are readable for the long term**
 - **Compresses to very small (colour) documents**
 - **OCR enables searchable documents**
 - **Metadata attachments mean documents are in context**
- **Always check and decide whether to use XMP, XMP Extension Schemas or native format for meta data**



PDF/A-3 Application: Born Digital Files

- **Hybrid Archiving**
 - For a document still in its life cycle, further versions might be created
 - Answers the question of when to create an archive ready rendition (PDF/A)
- **Using PDF/A-3**
 - Still open to change native format as attachment embedded into archive ready PDF/A rendition
- **Scope of use**
 - **PDF/A is a long term archiving format**
 - Works for many years and decades
 - **Attachments may have a limited lifetime**
 - Maybe a couple of years
 - Dependent on external software application



PDF/A-3 Application: Document Communication

- **Electronic invoice exchange format**
 - **Goal: no capture needed on receiving side**
 - Also for users, without EDI systems
 - **German initiative of Ministry of Economics, associations, stakeholders**
 - **German DIN standard under construction, thereafter:**
 - PDF/A-3 is the electronic invoice, XML based invoice data embedded
 - XML: Core Invoice Data Model MUG (CEN EU specification)
 - Subset of UN/CEFACT Cross Industry Invoice (CII)
 - CWA 16356-1 to -3 describes the data model (very technical)
 - **Outlook: get it to the EU level and international standards**



PDF/A-3 Application: Document Communication

**PDF/A file is the invoice.
XML data is embedded files
with **Data** relationship**

- Invoice can be CI-invoice
- Alternatively rendering XML using CEN Core Invoice Standard View

Item	Description
1	<p>Widgets Seller's ID: SE54321 Standard ID: ST12345 Commodity Classification: A Commodity Sub-Classification: 1 Order Line: 1 Note: A brief note about these widgets. Length: 200MLT Width: 100MLT Thickness: 90MLT Gross Unit Price: 10.00 Discount: 1.00 Country of Origin: GB</p>
2	Nuts and Bolts

Part 3 Annex C_CEN_Core_Invoice_Standard_View_v1.0.pdf - Adobe Reader

Datei Bearbeiten Anzeige Fenster Hilfe

1 / 2 73,3% Werkzeuge Signieren Kommentar

CEN Core Invoice - Standard View v1.0

Seller Seller name Seller address line1 Seller address line2 Seller address city Seller address country sub entity Seller address post code GB VAT No: GB123456789 Contact: Seller contact person name Telephone: Seller contact telephone number Fax: Seller contact fax number Email: Seller contact email address Seller's Published ID: 123456789 Seller's Legal Registration ID: Seller legal registration identifier Seller's Tax Registration Status: Seller tax registration status	Invoice date and No Invoice No: Invoice identifier Invoice Date: 09/05/2011 Invoice Period Start: 01/05/2011 Invoice Period End: 31/05/2011 Other References Buyer's Accounting Code: BUY0001 Buyer's Reference ID: Buyer reference identifier Order Reference ID: Order reference identifier Contract Reference ID: Contract reference identifier Contract Type: Contract type, text Contract Type Code: AXA
Buyer Buyer name Buyer address line1 Buyer address line2 Buyer address city Buyer address country sub entity Buyer address post code FR VAT No: Buyer VAT identifier Contact: Buyer contact person name Telephone: Buyer contact telephone number Fax: Buyer contact fax number Email: Buyer contact email address Buyer's Published ID: 3210987654321 Buyer's Legal Registration ID: Buyer legal registration identifier	Terms of delivery and payment Delivery Date: 31/03/2011 Delivery Location ID: Delivered to location identifier Delivered to address line 1 Delivered to address line 2 Delivered to address city Delivered to address country sub division Delivered to address post code Delivery Country: FR Payment Due Date: 09/05/2011 Invoice Currency Code: EUR Payment Means Type: 31 Payment Terms: Payment terms Seller's Bank Account Details Please quote Payment Reference: Seller payment identifier Bank Account Number: Seller account identifier Bank Branch Number: Seller financial institution branch ID Bank BIC: Seller financial institution identifier Bank IBAN: Seller account identifier

Item	Description	Quantity	Unit	Unit Price	Amount	VAT Code
1	<p>Widgets Seller's ID: SE54321 Standard ID: ST12345 Commodity Classification: A Commodity Sub-Classification: 1 Order Line: 1 Note: A brief note about these widgets. Length: 200MLT Width: 100MLT Thickness: 90MLT Gross Unit Price: 10.00 Discount: 1.00 Country of Origin: GB</p>	5	no	9.00	45.00	T01
2	<p>Nuts and Bolts Seller's ID: SE56789 Standard ID: ST98765 Order Line: 2 Note: A brief note about these nuts and bolts. Gross Unit Price: 2.00 Discount: 0.50 Country of Origin: GB</p>	100	net	1.50	150.00	T01
Sum of Line Amounts					195.00	
Total Charges "See buyer's letter"					8.90	E
Total Charges reason code: 78					10.00	T01
Total Allowances reason code: 94					25.00	T01
Total Allowances					25.00	

1 of 2





Introducing the PDF/A-3 Standard

Gary Hodkinson

gary.hodkinson@pdfa.org