PDF/A 101

archives are forever

Presenter

David van Driessche - CTO, Four Pees

Your presenter...

- Studied physics a long-time ago
- Worked with PDF since 1996
  - Enfocus Software
  - Gradual Software
  - Four Pees
- Standardisation around PDF since 2000
  - CGATS / ISO
  - Ghent Workgroup (and lots of European national organisations)

The topic...

Limited to PDF/A

Physical Archives

Archiving video, audio...

Archival Facilities

Hardware requirements for digital archival

Non-PDF file formats

Digital Archival using PDF

Data rot

What to archive?

Retrieval, security, access rights

How much of a problem?

- I wrote my thesis in 1994 - that's 20 years ago
  - "Using MLCFA for decomposition of energy-widening spectra of positron-annihilation events"
- How?
  - Ms-DOS 3.31 and Windows for Workgroups 3.11
  - WordPerfect 5 and Word for Windows 6.0
  - Saved on floppy disks (lots of them)
  - State-of-the-art backup on an Iomega ZIP-disk

It’s only 20 years!

- Early models of the cell phone
- The release of “Mosaic 0.9” from Netscape
- No laptops, tablets, DVD or viagra…
- The invention of the PlayStation (Remember Donkey Kong?)
- No Facebook, Twitter, Tumblr, …
- The birth year of Justin Bieber!

The solution?

- A file format that:
  - Is self-contained (does not need elements outside of the file)
  - Is compatible with different operating systems and their hardware
  - Is global (usable world-wide)
  - Is compact
  - Is not under the control of one particular vendor
  - Will exist for a long time
  - Well-behaved

The solution!

PDF...

- Invented by Adobe in 1993
  - Originally as an electronic documentation and Internet format
  - Rapidly adopted for other purposes starting from 1996
- Does it qualify?
  - Self-contained - compact - compatible - global
  - Vendor-independent?
  - Will last for a long time?
  - Well-behaved?

Standardised!

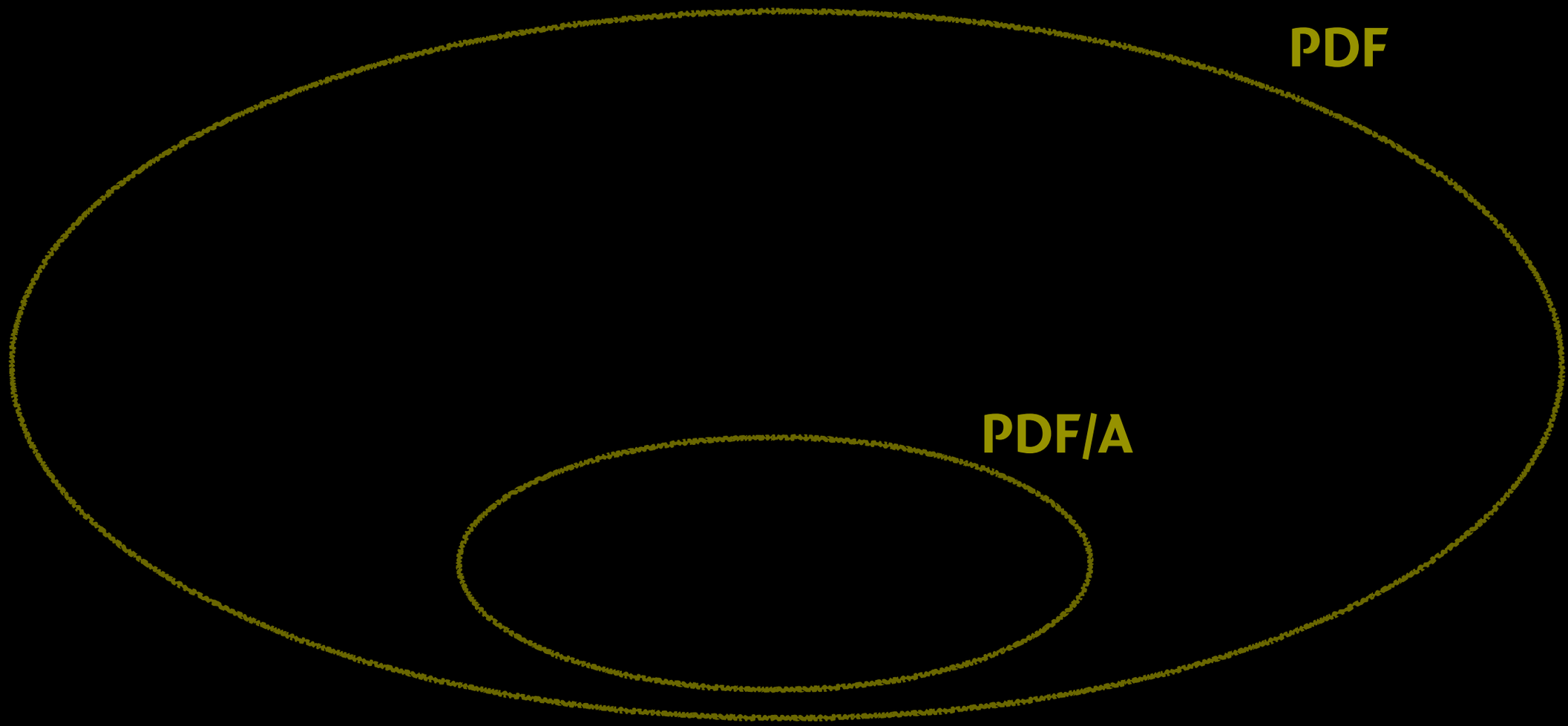
- PDF was adopted by the ISO
  - ISO 32000-1:2008
- New versions are developed by anISO committee, not by any individual vendor

PDF/A

- ISO published the PDF/A standard in 2005
  - ISO 19005-1:2005
- “Electronic document file format for long-term preservation – Part 1, Use of PDF 1.4 (PDF/A-1)”

Well-behaved?

- PDF files can be good
- PDF/A files have to be good



PDF/A-1

- All fonts must be embedded
- All colorspaces must be device-independent
- No Javascript, embedded files, encryption, external content references, transparency, layers, XFA forms data…
- Standards-based metadata must be used

Parts and Flavors

PDF/A-1

PDF/A-2

PDF/A-3

PDF/A-2a

PDF/A-2b

PDF/A-2u

Vanilla or Strawberry?

- Different archives require different things…
- Different flavours of the PDF/A standard cater to that:
  - PDF/A-1b, the basic flavour
    - Guarantees visual reproduction
  - PDF/A-1a, the accessible or advanced flavour
    - Incorporates all requirements of the basic flavour
    - Also focuses on the meaning of the document content

PDF/A-1b

Paris

Visual Reproduction

PDF/A-1a

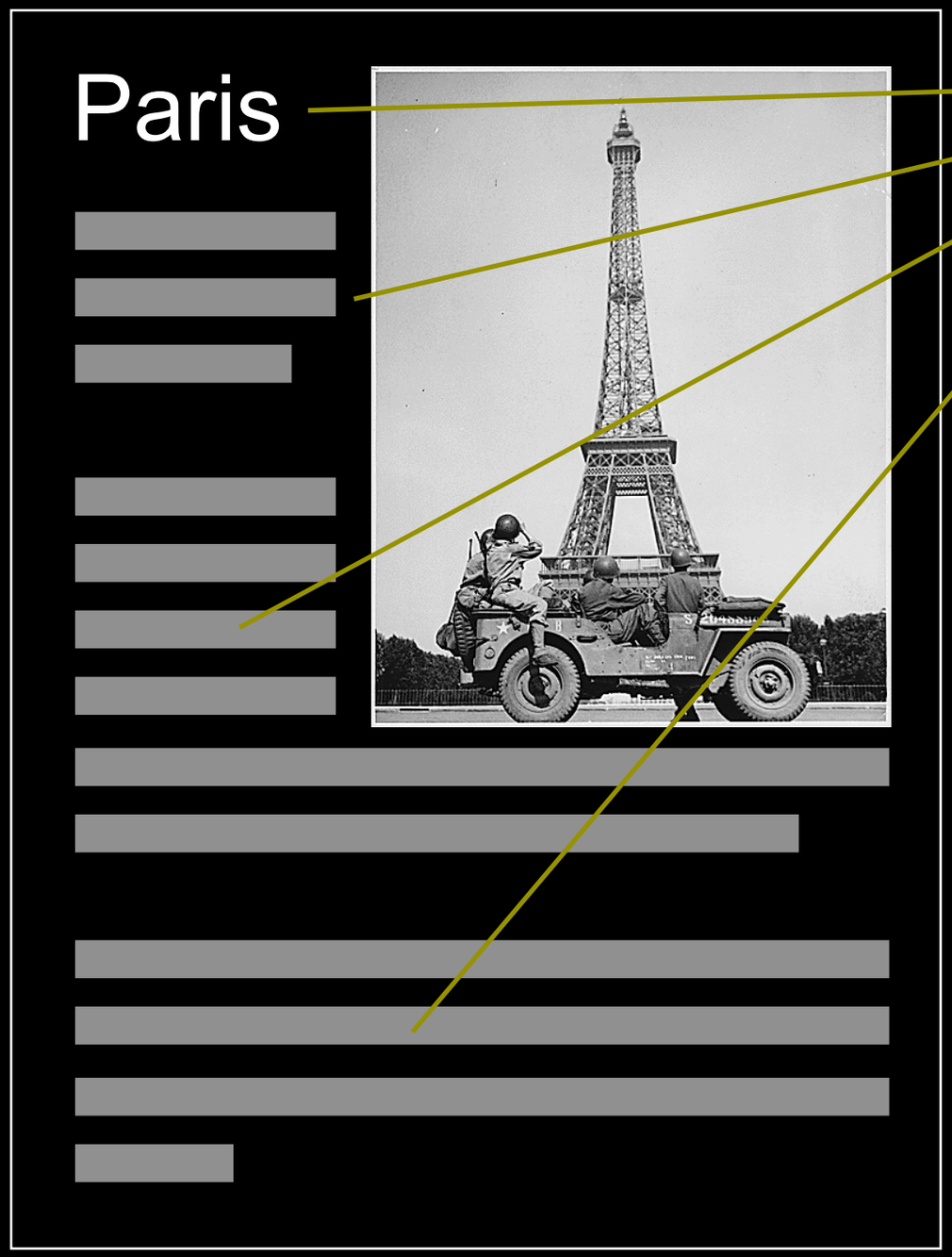
Paris

Visual Reproduction

+

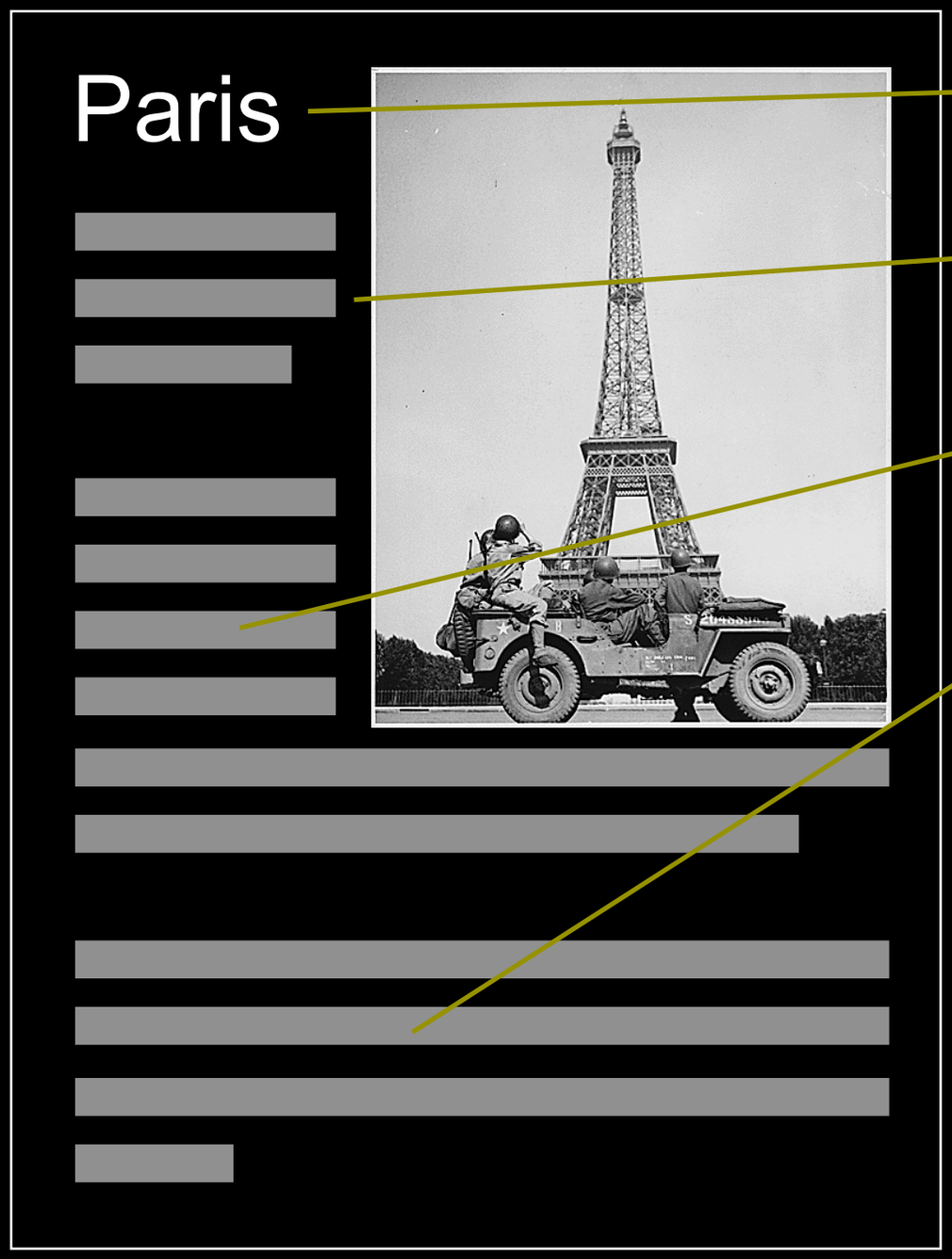
Meaning

PDF/A-1a: text



Text must have mapping to Unicode

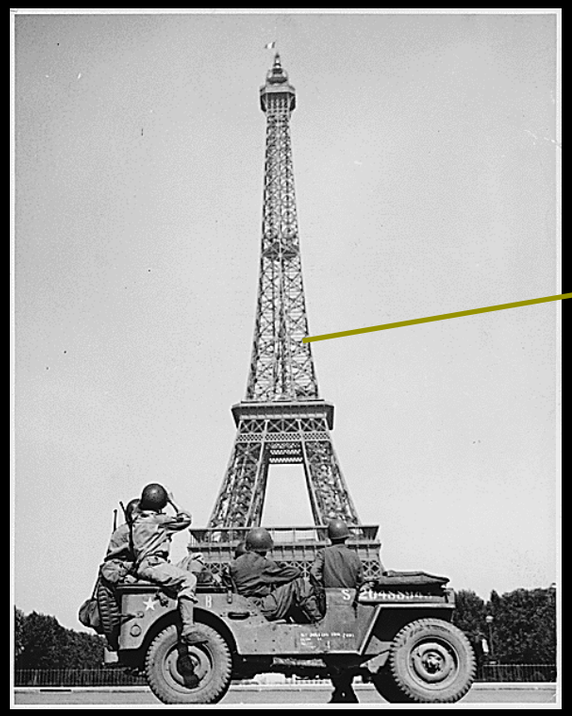
PDF/A-1a: structure



- Document Title
- Paragraph
- Paragraph
- Paragraph

PDF/A-1a: tagging

Paris



Description

"WWII: American soldiers watch as the tricolor flies from the Eiffel Tower again."

So… what to use?

- PDF/A-1b
  - Relatively easy to make and check
  - Easy to make and check without human intervention
- PDF/A-1a
  - Contains more usable content -> great for searchability of archives and additional knowledge about stored documents
  - Very difficult to create automatically *unless* the source document already contains the necessary information
  - Close to impossible to fully check automatically

Parts and Flavors

PDF/A-1

PDF/A-2

PDF/A-3

PDF/A-2a

PDF/A-2b

PDF/A-2u

What's next?

- First of all, PDF/A-1 will always remain a valid ISO standard
  - So you can continue to use PDF/A-1a and PDF/A-1b as long as you want
- But new PDF/A standard parts have been developed

PDF/A-2

- Adds support for a number of additional features that are not allowed in PDF/A-1. Specifically:
  - Transparency
  - Layers (also referred to as “optional content”)
  - JPEG-2000 compression
  - Embedding of OpenType fonts
  - Digital signatures in accordance with the PAdES standard
  - Embedding of other PDF/A-2 documents

Embedding

Re: Archiving Penguins

From: David van Driessche
Date: December 25,2015
To: info@callassoftware.com

Dear sirs,

It has come to my attention that you have software to archive documents using the PDF/A standard. Would this technology be up to the task of archiving penguins?

Looking forward to your answer!
Kind regards,
David.

Penguin Info.pdf

PDF/A-2

Penguin Info.pdf (PDF/A-2)

PDF/A-1 or PDF/A-2?

- None is automatically compatible with the other
  - A PDF/A-1 document is not necessarily compliant with PDF/A-2
  - A PDF/A-2 document is not necessarily compliant with PDF/A-1
- Both will remain in existence
- Look at your source content and the needs of the archive

PDF/A-2: a new flavour

- PDF/A-2b (basic)
- PDF/A-2u (Unicode)
  - basic
  - Unicode mapping for text
- PDF/A-2a (advanced)
  - basic
  - Unicode mapping for text
  - Structure & tagging

PDF/A-3

- Same flavours as in PDF/A-2
- Adds support to embed arbitrary documents in PDF/A-3 files

Embedding

Re: Archiving Penguins

From: David van Driessche
Date: December 25,2015
To: info@callassoftware.com

Dear sirs,

It has come to my attention that you have software to archive documents using the PDF/A standard. Would this technology be up to the task of archiving penguins?

Looking forward to your answer!
Kind regards,
David.

Penguin Info.ppt

PDF/A-3

Penguin Info.ppt

Penguin Info.pdf

Thanks!

Questions?

David van Driessche
Chief Technical Officer, Four Pees
Executive Officer, Ghent Workgroup

david.van.driessche@fourpees.com
www.fourpees.com