Preserving PDF at the coalface

ads
ARCHAEOLOGY
DATA SERVICE

PDF/A at the Archaeology Data Service

Tim Evans 15-07-2015

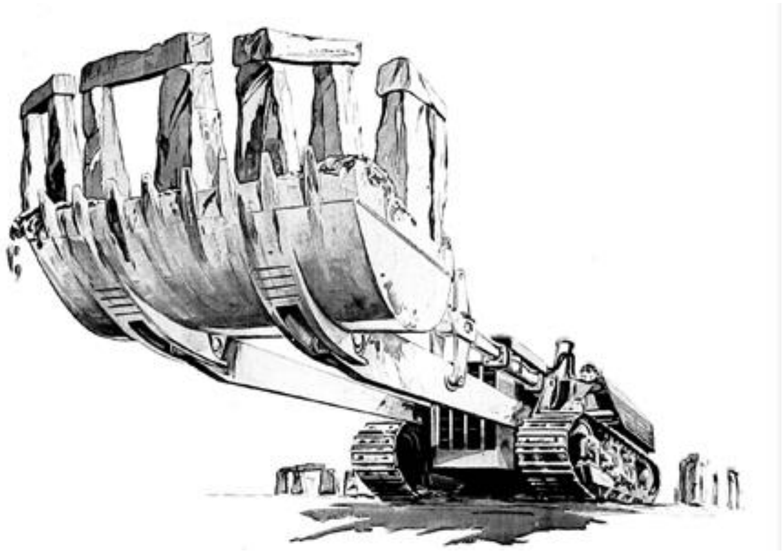
Introduction

The Archaeology Data Service:

- Established in 1996
- Based within the Department of Archaeology, University of York
- Digital archive for UK-based fieldwork and research in archaeology
  - Academic 'research' archives produced by Higher Education
  - The results of development-led fieldwork

What was the impetus in 1996?

- Is destructive!



- Resources are unique
- The preserved record becomes the main resource for future interpretation

Progress to-date

- Collections
  - 1,100,000 metadata records
  - 700+ rich archives
  - **30,000+ reports**

Cotswold Archaeology

Land at Dunkeswell Airfield
Dunkeswell
Devon

Archaeological Strip, Map and Sample
Excavation Report

for
Conergy West Sussex Ltd

CA Project: 4986
CA Report: 14418
RAMM 14/44

May 2015

CONERGY
Conergy West Sussex Ltd

The problem

Traditionally archives deposited as part of a formal process:

- Negotiation and 'enforcement' of accepted formats
- Documents normally deposited as DOC, DOCX, RTF, ODT etc

The problem

- Documents often comprise text and raster/vector elements created in a range of Softwares e.g. Adobe Illustrator
- ADS recommended different elements deposited as separate files for preservation
  - Text as DOC etc
  - Raster at TIF etc
  - Vector as DXF
- A PDF (usually 1.4) created for dissemination

The problem

On occasion a PDF was the only format that could be deposited

- Lack of technical expertise
- Departure of staff member
- Apathy!

The problem

- Preservation Headache!
  - Copying and pasting into Word!
  - Saving in alternative formats – XML
  - Last resort – saving each page as a TIF
- Not ideal:
  - Down-sampling of embedded images
  - Lack of OCR
  - Extra files and management for any future migration

Things get worse

- Only dealing with academic archives
  - Negotiation was controlled (mostly!)
  - Relatively small volume of PDFs deposited
- In the early 2000s the ADS (and partners in national and local government) launched the OASIS system…

What is OASIS?

- Online system for recording investigations
- Aimed at ‘development led’ works undertaken as mitigation in the planning process
- Over 4000+ events every year in England alone
- Each event has its own report, submitted to the local authority as part of ‘preservation by record’.

http://dx.doi.org/10.3789/isqv25no3.2013.04

What is OASIS?

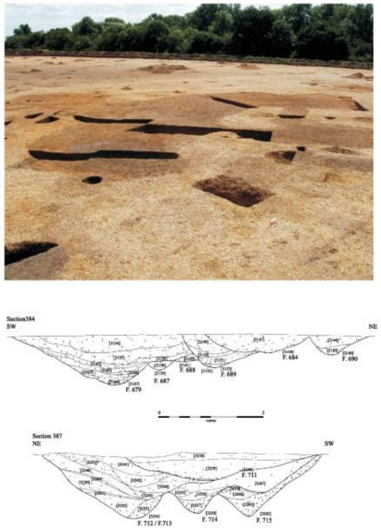
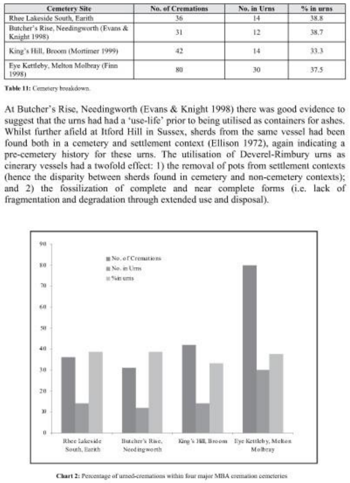


Figure 19. Compound A Photo and Sections



Cemetery Site	No. of Cremations	No. in Urns	% in urns
Rhee Lakeside South, Earith	36	14	38.8
Butcher's Rise, Needingworth (Evans & Knight 1998)	31	12	38.7
King's Hill, Broom (Mortimer 1999)	42	14	33.3
Eye Kettleby, Melton Molbray (Finn 1998)	80	30	37.5

Table 11: Cemetery breakdown.

At Butcher's Rise, Needingworth (Evans & Knight 1998) there was good evidence to suggest that the urns had had a 'use-life' prior to being utilised as containers for ashes. Whilst further afield at Itford Hill in Sussex, sherds from the same vessel had been found both in a cemetery and settlement context (Ellison 1972), again indicating a pre-cemetery history for these urns. The utilisation of Deverel-Rimbury urns as cinerary vessels had a twofold effect: 1) the removal of pots from settlement contexts (hence the disparity between sherds found in cemetery and non-cemetery contexts); and 2) the fossilization of complete and near complete forms (i.e. lack of fragmentation and degradation through extended use and disposal).

90
80
70
60
50
40
30
20
10
0
No. of Cremations
No. in Urns
%in urns
Rhee Lakeside South, Earith
Butcher's Rise, Needingworth
King's Hill, Broom
Eye Kettleby, Melton Molbray

Chart 2: Percentage of urned-cremations within four major MBA cremation cemeteries

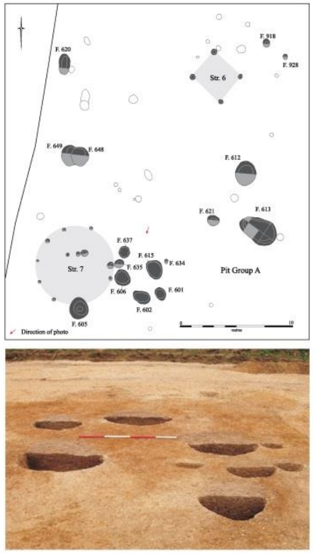


Figure 15. Pit Group A

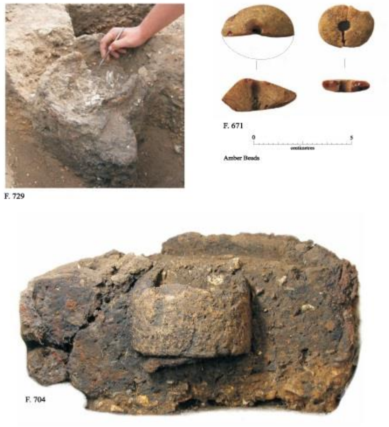
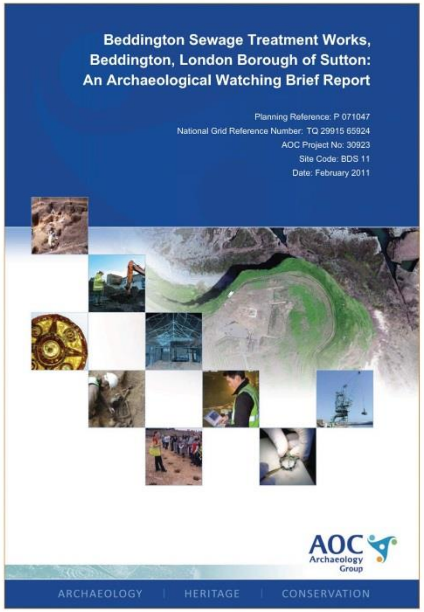


Figure 8. Selected Cremations



Beddington Sewage Treatment Works, Beddington, London Borough of Sutton: An Archaeological Watching Brief Report

Planning Reference: P 071047
National Grid Reference Number: TQ 29915 65924
AOC Project No: 30923
Site Code: BDS 11
Date: February 2011

AOC Archaeology Group

ARCHAEOLOGY | HERITAGE | CONSERVATION

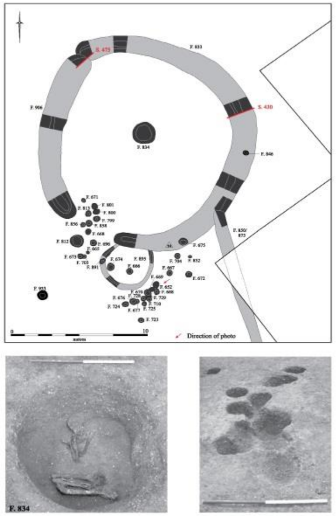
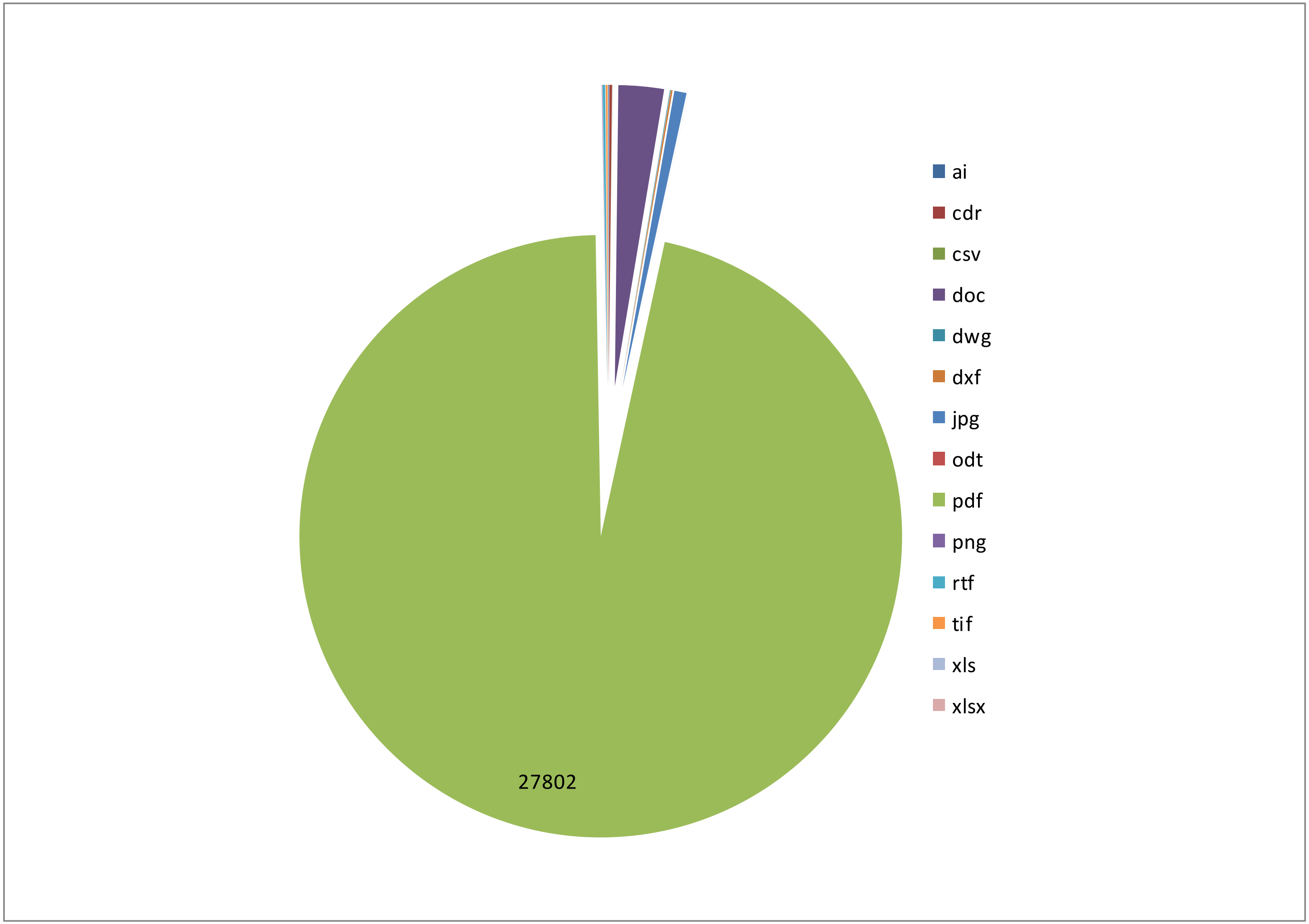


Figure 6. Ring-ditch Monument and Cremation Cemetery

The rise of the PDF

- Offers a compact package
- Free to view
- Perceived as static – content and layout will remain the same
- Traditionally stored on a Local Authority server or even printed out!



Files accessioned through OASIS March 2008 – July 2015

Can’t you ask for non-PDF?

Reports are ‘policed’ by the relevant curator in Local Government, not the ADS

- These curators (understandably) prioritise getting the report and validity of content
- Local Government under severe financial pressures – more important things to worry about!
- Besides, the PDF offers a simple solution for all these parties, why not use it?

Archival solutions (2008)

“PDF/A-1 aims to preserve the static visual appearance of electronic documents over time and also aims to support future access and future migration needs by providing frameworks for embedding metadata about electronic documents, and defining the logical structure and semantic properties of electronic documents’ PDF/A standard“

At the beginning

- All conversions using
- Adobe Acrobat 8 - Pro 9 and 10
- Manually using the Preflight tool
- Almost impossible to migrate an archaeological report to the 1A standard
- PDF/A 1B was thus used as the default option

Initial problems

- PDF to PDF/A 1B conversions were problematic, common issues included:
  - Issues with XMP properties
  - Glyphs
  - Fonts have not been embedded.

However…

- A lot of these can be fixed with manual intervention
- Dreaded ‘Print to PDF/A 1B’ option
    - Uncertainty as to what we’re creating?
    - Are we damaging images, losing text?
- A lot of time spent in manual checks of PDF/A files to ensure significant properties retained

PDFTron

- Since mid-2011 the ADS utilised PDF/A Manager created by PDFTron
  - batch-level processing;
  - automation of previously manual fix-ups (such as the editing/stripping of XMP properties)
- Success rate – i.e. the PDFTron conversion script returning a successful post-conversion validation – is on average around 80% of all files processed (2013)

http://archaeologydataservice.ac.uk

However…

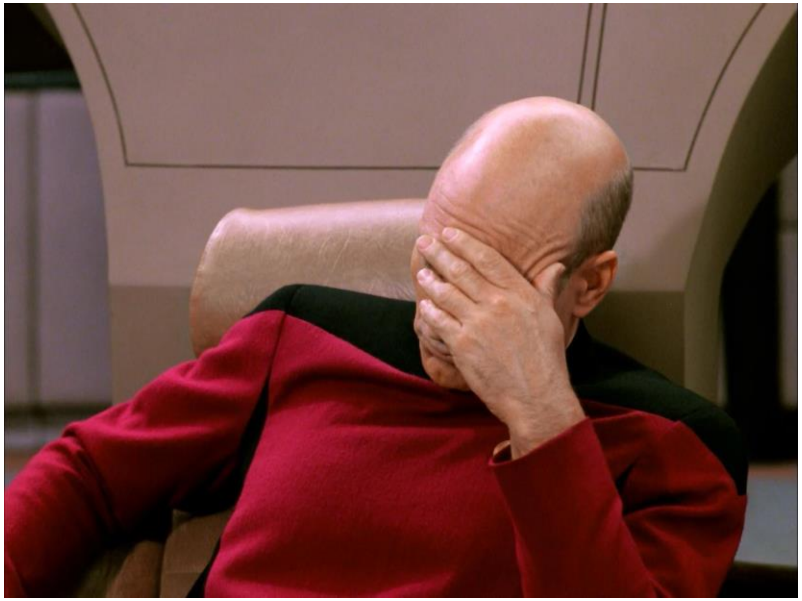
- PDFTron conversion rate was notably less successful when dealing with the growing numbers of PDF 1.5-1.7 (2015)
- Although a useful tool when dealing with a large corpus of mixed PDF formats, has not been the overarching solution
- Still using a mixture of PDFTron and Preflight

http://archaeologydataservice.ac.uk

However…

- PDFTron PDF/A 1B files checked in Preflight (Pro X)

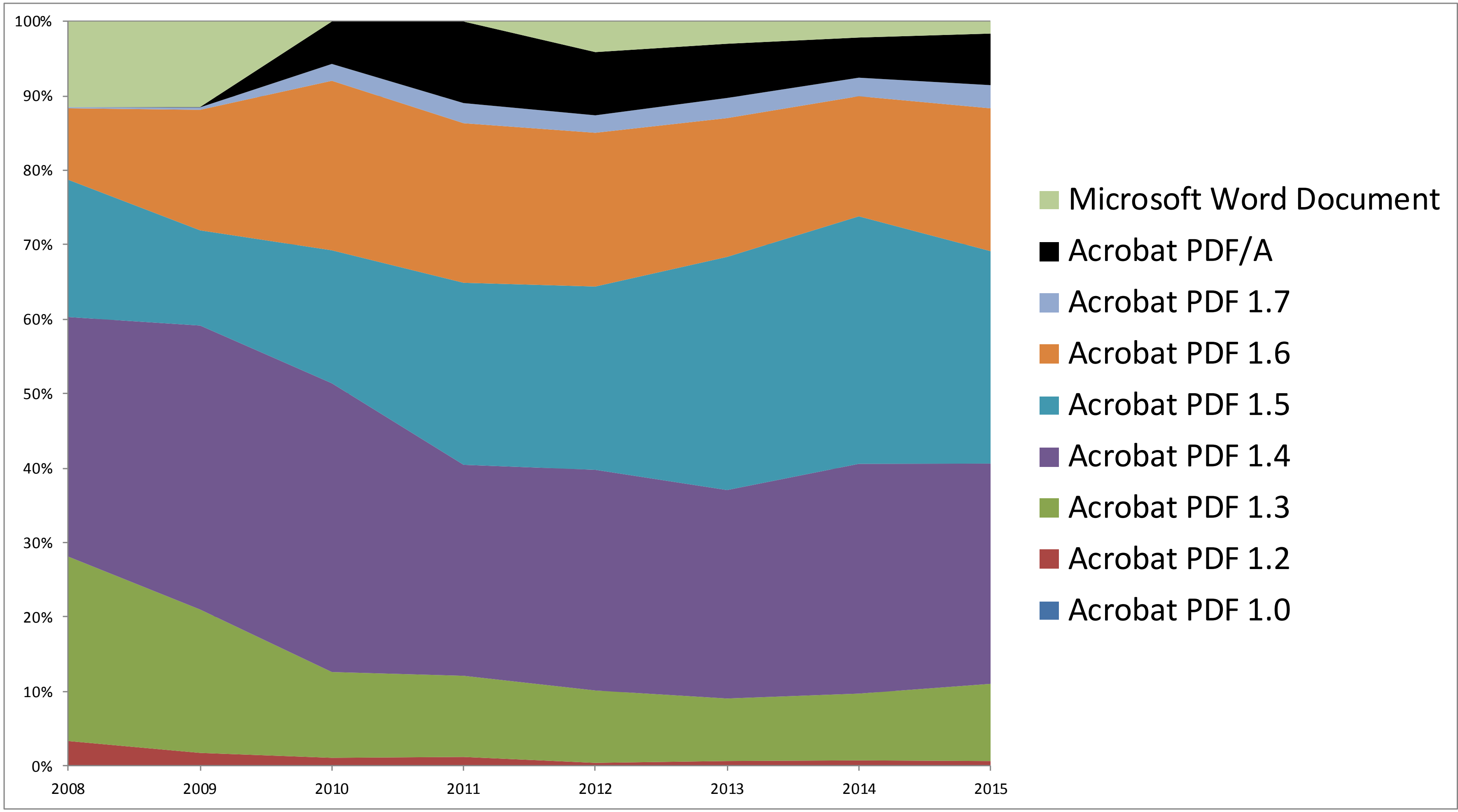
On average between 5-10% have failed to conform to the profile



Uncertainty over the PDF/A profile

- We returned to files created in Acrobat 8:
  - significant numbers (c. 25%) of these were not validating as such in the Preflight tool for either Pro 9 or 10
  - a very small number created in Acrobat 9.5.5 that were not validating in Pro 10
- On closer investigation it seems these issues have not been confined to the ADS experience. The Bavaria report identifies the same key problems, and attribute them to the fine-tuning of the PDF/A-1 standard throughout versions 7-9 of Adobe Software, as well as third party tools.

http://archaeologydataservice.ac.uk


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2008
2009
2010
2011
2012
2013
2014
2015
Microsoft Word Document
Acrobat PDF/A
Acrobat PDF 1.7
Acrobat PDF 1.6
Acrobat PDF 1.5
Acrobat PDF 1.4
Acrobat PDF 1.3
Acrobat PDF 1.2
Acrobat PDF 1.0

Identification of incoming PDF/A

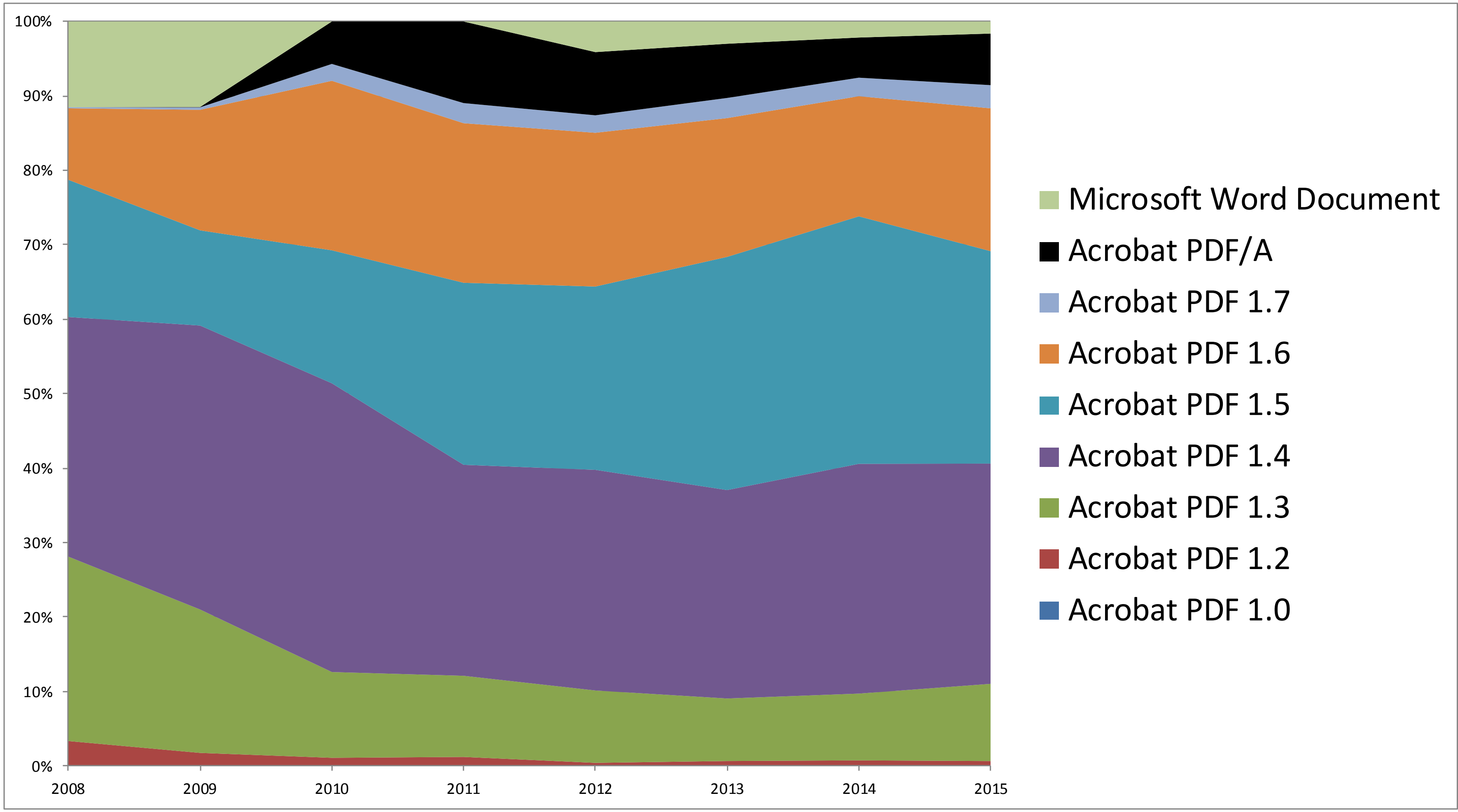
- DROID (The National Archives) run on ingest
- Detects file type <pdfaid:part> namespace declaration in the PDF/A Identification
- ON subsequent checking (Preflight) most of these failing: ‘false positives’
- Much alarm!

Third party tools

- CutePDF
- OmniPage Capture SDK
- Nitro PDF
- PDFCreator
- Acrobat PDFMaker for Word (v.8)
- = the majority of PDF/A files we were receiving from depositors were not validating

Moving forward

- Open Planets Foundation (OPF) - PDF Identify, validate, repair. Hamburg 1-2 Sept 2014
  - Some ‘errors’ in validation would not cause significant preservation risks in the future
- PDF/A 2: (ISO 19005-2). PDF/A 1 is too restrictive and doesn’t deal well with aspects of PDF which have become available since PDF1.4


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2008
2009
2010
2011
2012
2013
2014
2015
Microsoft Word Document
Acrobat PDF/A
Acrobat PDF 1.7
Acrobat PDF 1.6
Acrobat PDF 1.5
Acrobat PDF 1.4
Acrobat PDF 1.3
Acrobat PDF 1.2
Acrobat PDF 1.0

Current Practice

- Use of PDF/A 1 and PDF/A 2
  - Notable that some files will only convert to A1 not A2 (in Adobe)
  - Adopting a 'best-fit' policy using the two standards (yet to get to grips with A3)

Callas

- ADS use of Callas PDF Toolbox
  - Same validation tool as Adobe Acrobat X Pro and thus consistent validation
  - Significant advantages in reporting
  - Batch conversion

Use of PDF/A 2

- From Adobe Preflight Syntax problem: Real value out of range (too high) (also commonly get the other extreme - too low)

  Solution to this issue was to use PDFA2 as the original PDFA1b spec is unable to cope with large numbers either negative or positive. The error in this case was caused by a bezier curve with a high value.

http://archaeologydataservice.ac.uk

Use of PDF/A 2

- Preflight turned the pages into raster images and putting the text as an invisible layer so it was still searchable.
- This was not an ideal solution as it lost the vector aspects of some of the illustrations (archaeological site plans).
- As the vector aspects of these illustrations is often seen as a significant property of the file it is not adequate solution for this use case

http://archaeologydataservice.ac.uk

Common problems

- Still an issue with **Fonts**: propriety font or unusual font size e.g. 11.00343pt
  - Workaround for all softwares (inc. Callas) is to rasterize and image
  - A solution, but not always best quality

http://archaeologydataservice.ac.uk

Thoughts

- Do we (ADS) need to be less fussy about the ‘significant properties’ for reports
  - Does vector content *always* need to be retained?
  - Should we worry about rasterizing pages (as long as OCR is present)?
- Now tied to a mixed economy of softwares and tools (some free, some commercial) to ensure consistent and accurate creation and validation.

Thanks for listening!

tim.evans@york.ac.uk