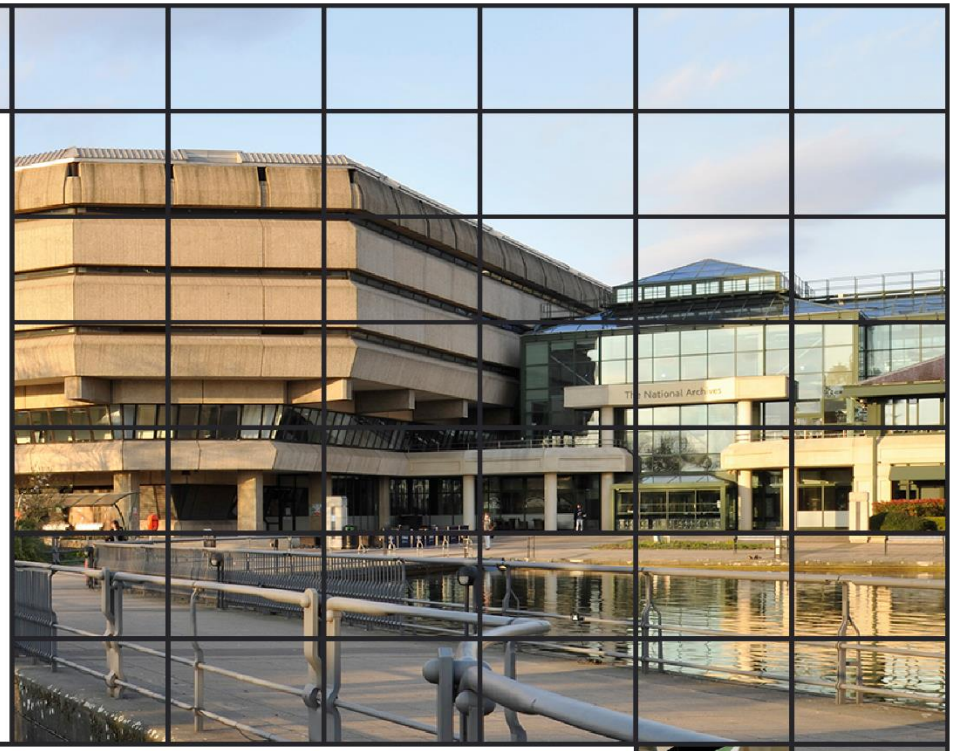
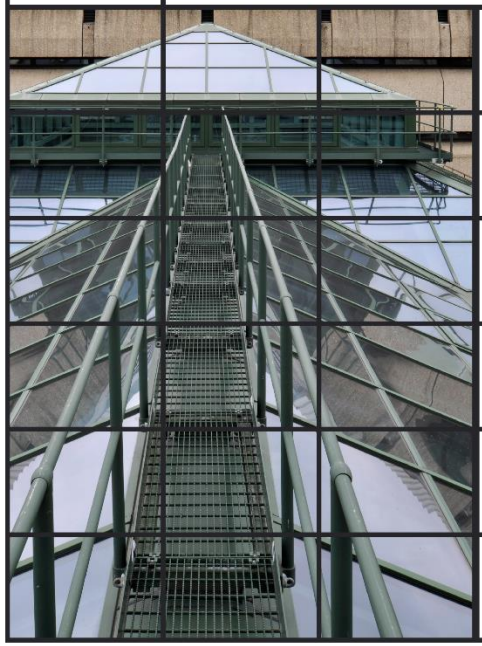
Digital Preservation 101

Session 2:

Knowing what you have part 2





Paul Young
Digital Preservation Specialist/Researcher

THE NATIONAL ARCHIVES

Photo by Belinda Fewings on Unsplash

Schedule

- Background of PRONOM
- **Break**
- How does PRONOM work?
- **Lunch**
- What do you do when PRONOM doesn’t work? File format process at TNA
- Homework (what are we looking for in a DROID report?)
- **Break**
- File format surgery
- **Finish**

THE NATIONAL ARCHIVES

Overview of Today's Session

Today is all about file format identification!

Aim of today

In continuation of the previous session learning about what you have, we will be looking at file format identification, how DROID and PRONOM work and what to do when you come across new and unknown formats.

THE NATIONAL ARCHIVES

Photo by Gia Oris on Unsplash

Sources of help

PRONOM mailbox - PRONOM@nationalarchives.gov.uk

PRONOM request form - www.nationalarchives.gov.uk/contact-us/submit-information-for-pronom

PRONOM Google group - https://groups.google.com/forum/#!forum/pronom

DROID Google group – https://groups.google.com/forum/#!forum/droid-list

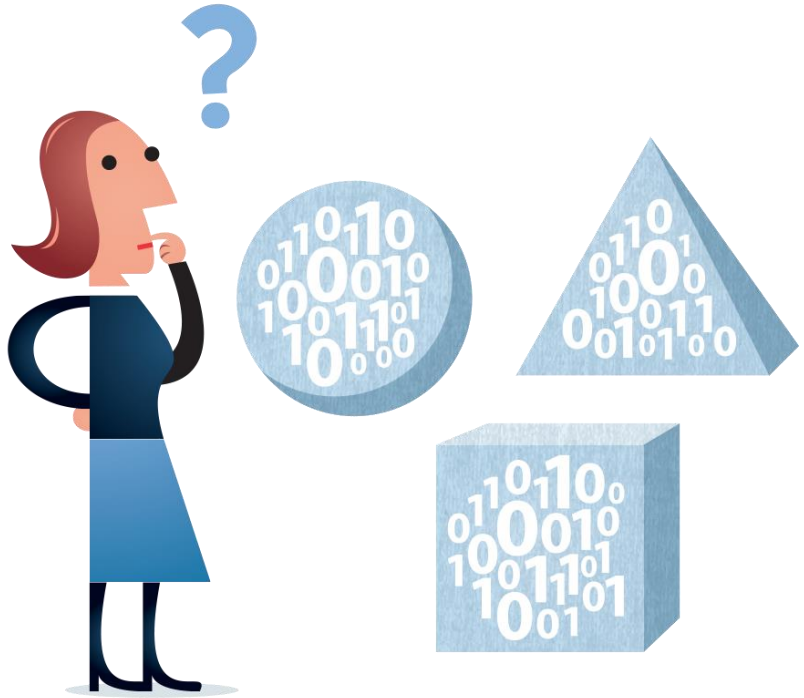
Twitter - #PRONOM #DROID

THE NATIONAL ARCHIVES

Recap from Previous Session – Why is File Format Identification Important?

One of the key aspects around knowing what you have is understanding the formats in your collection, it allows you to:

- Take preservation actions such as migration
- Ensure you have access methods e.g. for legislative purposes (FOI) and for researchers
- Metadata extraction, format validation depends on identification of format

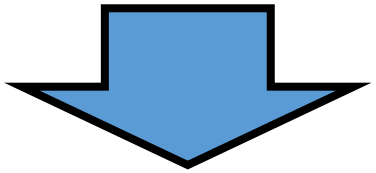


Digitalbevaring.dk

THE NATIONAL ARCHIVES

The technical registry
PRONOM

- File format registry which provides DROID with information
- File format identification signatures (patterns in byte sequences) Often referred to as ‘Magic’
- PRONOM data embedded in all major digital preservation systems



- DROID is a tool which uses the PRONOM file format registry to scan byte sequence and identify PUID by comparing known signatures in signature file

THE NATIONAL ARCHIVES

The technical registry

PRONOM

DROID
Digital Record Object Identification

fido

Preservica
Active digital preservation

DAITSS
Dark Archive in the Sunshine State

ExLibris
Rosetta

archivematica

RODA

libsafe

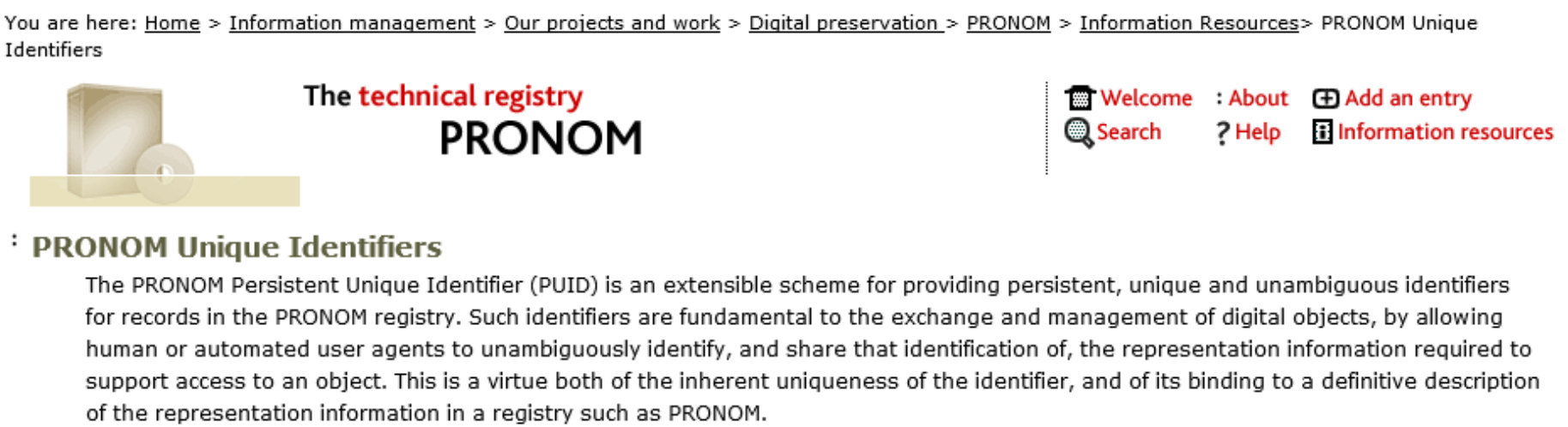
Fedora™

PROGRAMME
vitam
archivage numérique

THE
NATIONAL
ARCHIVES

Why PRONOM?

- PRONOM has been the most consistent file format registry over time
- PRONOM aims to provide ‘certain’ file format identification by matching consistent patterns to file format specifications



https://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm

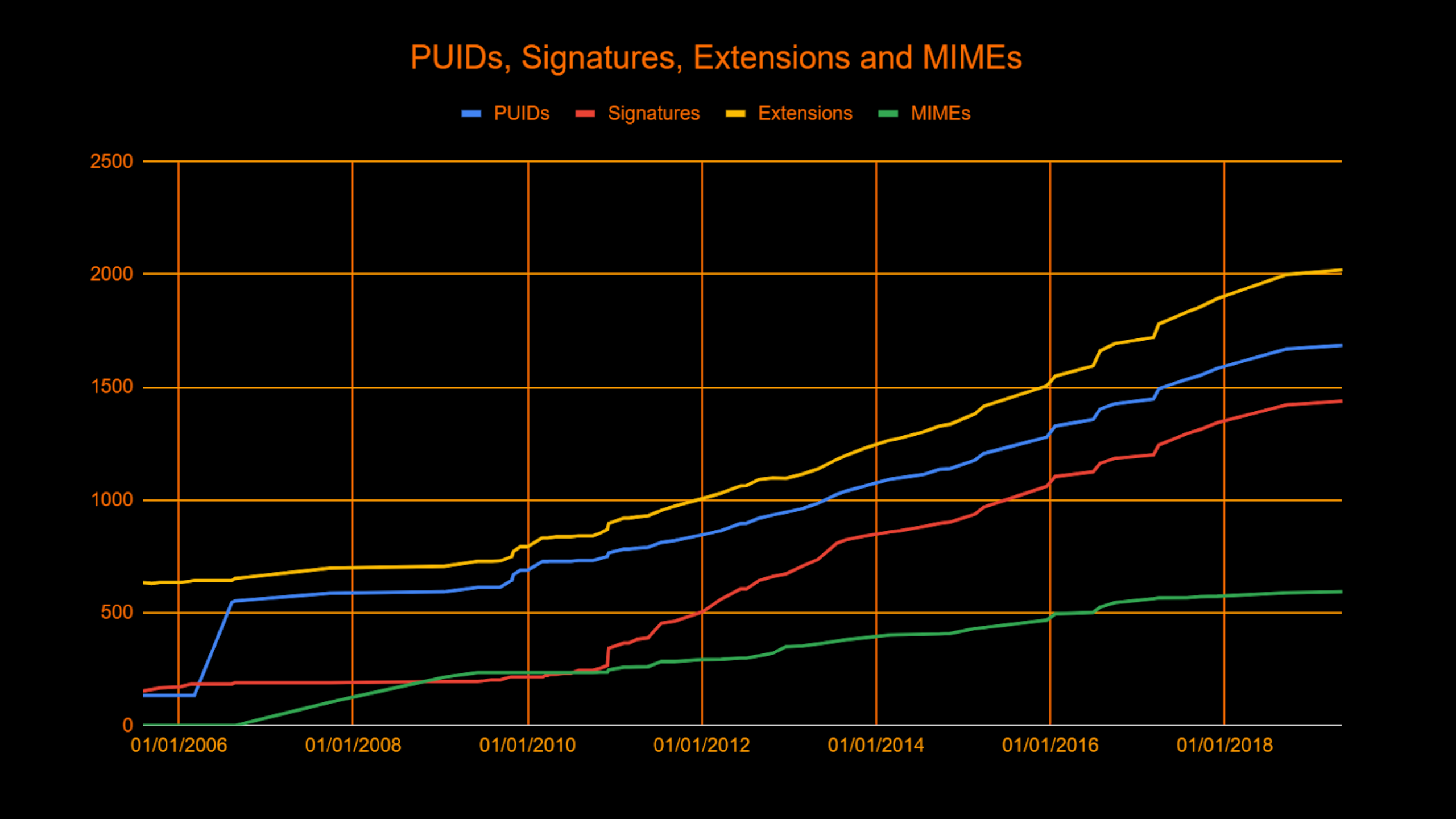
THE NATIONAL ARCHIVES

DROID & PRONOM timeline

- 2001 – 1st edition of PRONOM (Internal)
- 2004 – Opened up as externally browsable resource
- 2005 – 1st edition of DROID – developed to utilise the PRONOM data
- 2005 – PRONOM made publicly available
- Ongoing – TNA commitment to PRONOM research ongoing to meet UK Gov and community needs

Target of 100 new PUIDS a year for PRONOM research team

THE NATIONAL ARCHIVES


PUIDs, Signatures, Extensions and MIMEs
PUIDs
Signatures
Extensions
MIMEs
2500
2000
1500
1000
500
0
01/01/2006
01/01/2008
01/01/2010
01/01/2012
01/01/2014
01/01/2016
01/01/2018

PRONOM CONTRIBUTORS WORLDWIDE

https://www.google.com/maps/d/u/0/viewer?mid=1zWzV6G-CZDzq_kvIFGFYTgYxATI&ll=35.27438564888553%2C42.991917126214O5&z=2

Georgia Tech Research Institute
The National Archives and Records Administration
Lars Clausen
Museum of London
Digital Library Research Group
Swiss Federal Archives
Datable
Historic Environment Scotland
National Library of New Zealand
National Library of Singapore
Bibliothèque nationale de France
University of Southampton
Archaeology Data Service
Adobe Systems Incorporated
Hagley Museum and Library
Rhizome
British Library
The Borthwick Institute for Archives
Australian National University
l'Institut national de l' Audiovisuel
Asta Development
Axiell
Nokia
Tessella
National Library of Wales
Central Digital Archive
Artefactual Systems Inc
Yale University Library
KULeuven - LIBIS
Keep Solutions
Provincie West-Vlaanderen
Public Record Office of Northern Ireland
Swiss Federal Office of Topography
Walker Art Center
Catenda
Ricoh
Sebastian Lasse
Michelle Lindlar
State Records New South Wales
Archives New Zealand
University of Edinburgh
The Canadian Centre for Architecture
Hull History Centre
Bibliothèque cantonale et universitaire - Lausanne
The Church of Jesus Christ of Latter-Day Saints
Neil Gaiman
National Library of Scotland
Bodleian Library, University of Oxford
Landesarchiv Nordrhein-Westfalen
Harry Ransom Centre
New York Public Library
University of British Columbia
Technische Informationsbibliothek (TIB) Hannover
Victoria University of Wellington
National Library of Australia
SLUB Dresden
The National Archives of Estonia
Esri

THE NATIONAL ARCHIVES

MIME type vs PUID

- MIME (or media type) is a identifier for format, consisting of two parts *type* and *subtype*

  e.g. - image/jpeg

  - application/pdf

  - text/plain

- Internet Assigned Numbers Authority (IANA) is official authority for standardisation and publication - https://www.iana.org/
- Originally used to denote attachments in emails, often used by browser to determine how to render file
- Often determined by extension

THE NATIONAL ARCHIVES

MIME type vs PUID

Untitled-1 × Untitled-2 ×

Resource	Extension	Format	PUID	Mime type	Size
H:\File format tr...					
10.pdf	pdf	Acrobat PDF 1.4 - Portable Document Format	fmt/18	application/pdf	
cat.JPG	jpg	Exchangeable Image File Format (Compressed)	fmt/645	image/jpeg	
Closed cover i...	pdf	Acrobat PDF 1.3 - Portable Document Format	fmt/17	application/pdf	
images (4).jpg	jpg	JPEG File Interchange Format	fmt/43	image/jpeg	
images (5).jpg	jpg	JPEG File Interchange Format	fmt/43	image/jpeg	
images (6).jpg	jpg	JPEG File Interchange Format	fmt/43	image/jpeg	
images.jpg	jpg	JPEG File Interchange Format	fmt/43	image/jpeg	
New Microsoft...	pdf	Acrobat PDF 1.5 - Portable Document Format	fmt/19	application/pdf	
New Microsoft...	pdf	Acrobat PDF 1.5 - Portable Document Format	fmt/19	application/pdf	
New Microsoft...	pdf	Acrobat PDF 1.5 - Portable Document Format	fmt/19	application/pdf	
New Microsoft...	pdf	Acrobat PDF 1.5 - Portable Document Format	fmt/19	application/pdf	
penguins.jpg	jpg	JPEG File Interchange Format	fmt/43	image/jpeg	
Template for v...	pdf	Acrobat PDF 1.5 - Portable Document Format	fmt/19	application/pdf	

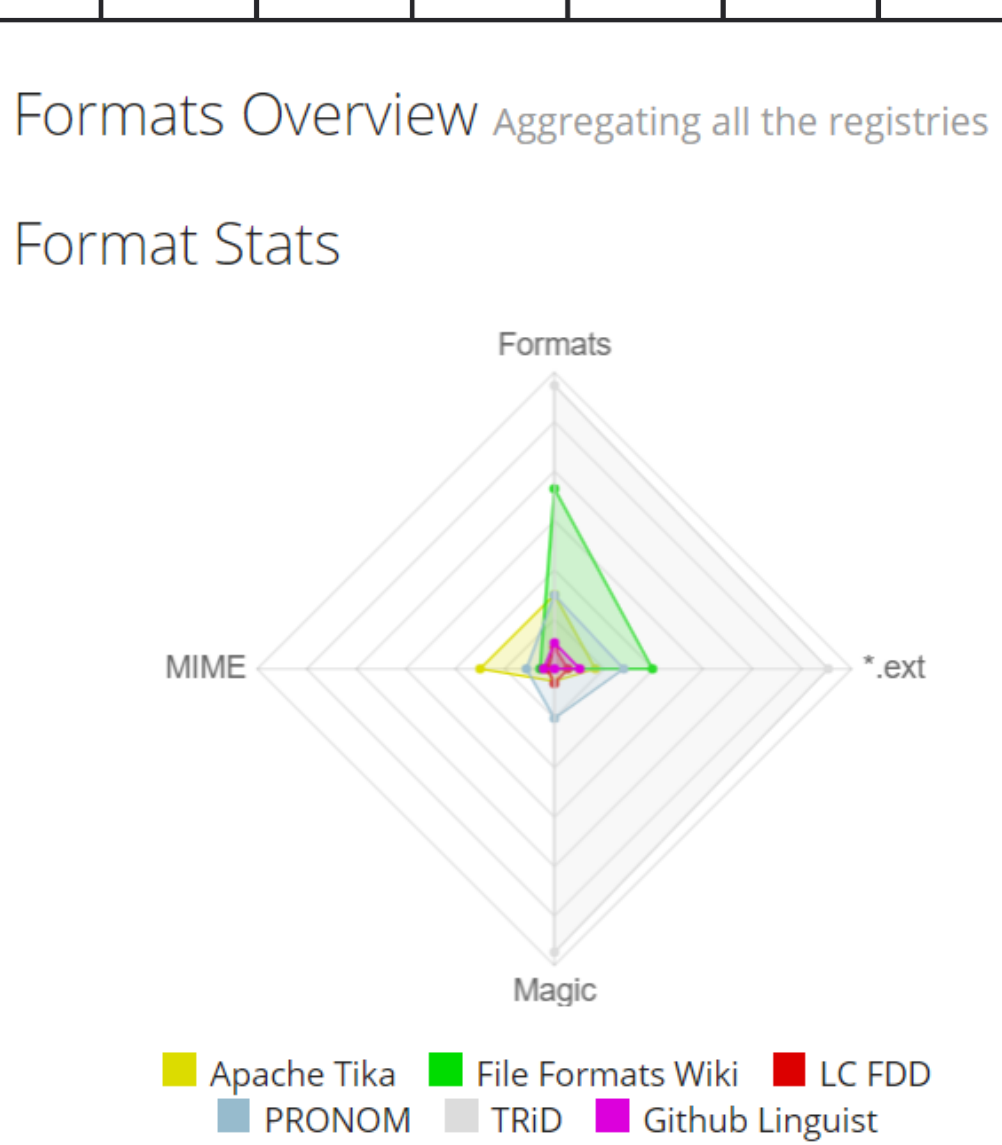
THE NATIONAL ARCHIVES

Exercise 1 – Have a browse around PRONOM

- Browse around PRONOM and have a look - https://www.nationalarchives.gov.uk/PRONOM/

Photo by Bruno Nascimento on Unsplash

THE NATIONAL ARCHIVES



Other sources of file format information

LOC – detailed file format information
http://www.loc.gov/preservation/digital/formats/index.html

File Formats Wiki – general file format information
http://fileformats.archiveteam.org/wiki/Main_Page

TRID – programmatic file identifiers
http://mark0.net/soft-trid-e.html

Source:
http://www.digipres.org/formats/

THE NATIONAL ARCHIVES

Photo by Ramiz Dedaković on Unsplash

HOW DOES IT WORK?

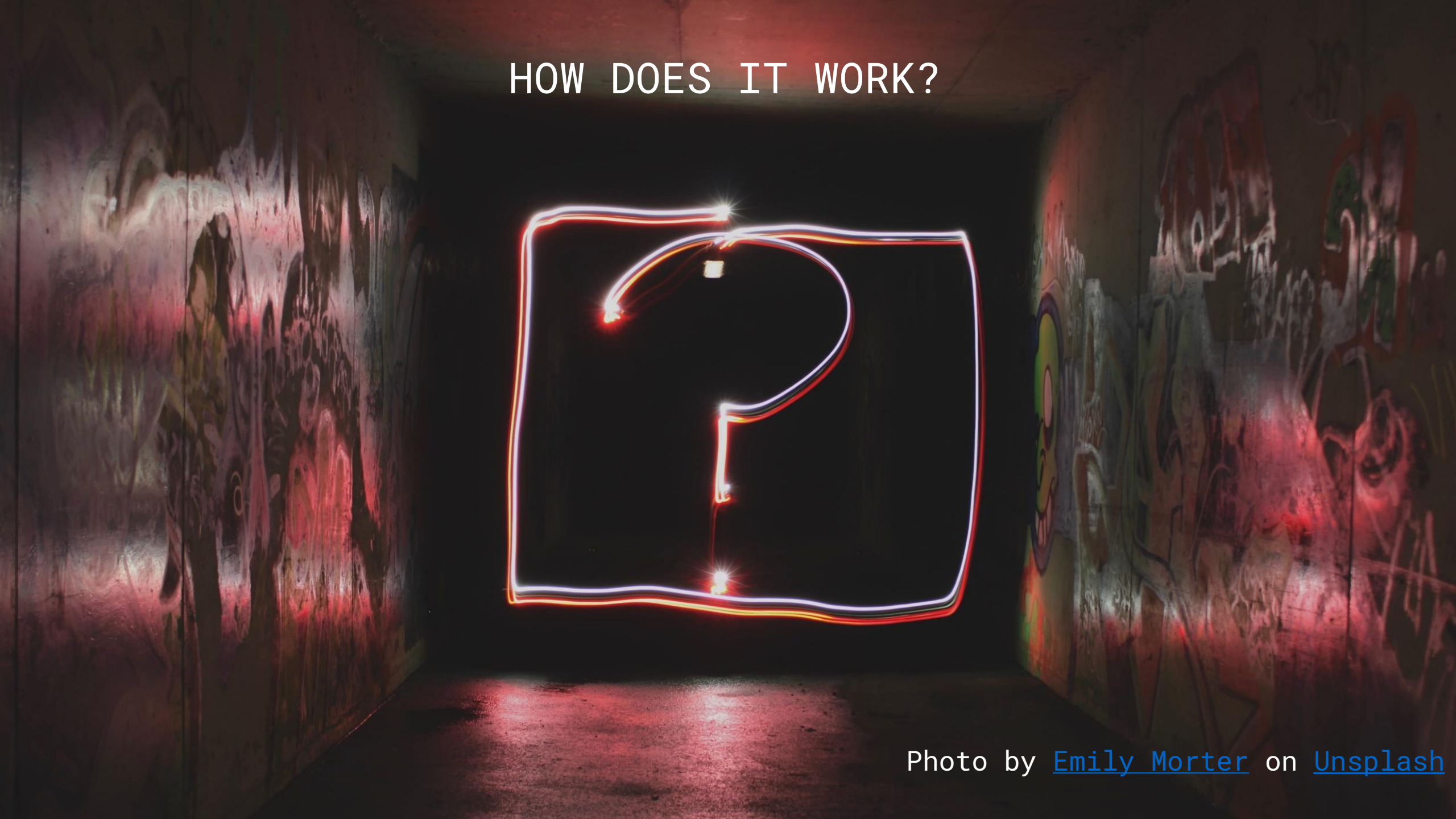


Photo by Emily Morter on Unsplash

How are PUID's assigned?

- Extension – Purely by extension (e.g. txt, .doc, jpg) cannot be as confident or specific about format identification
- Signature – Binary pattern matching, looking for unique and consistent identifying signature
- Container – formats made up of small files contained within a 'ZIP' or 'OLE2' wrapper (.doc, .xlsx, .odt)

New Microsoft Word ...	Container	fmt/412		docx
New Microsoft Word ...	Signature	fmt/19		pdf
New Microsoft Word ...	Container	fmt/412		docx
New Microsoft Word ...	Signature	fmt/19		pdf
New Microsoft Word ...	Container	fmt/412		docx
New Microsoft Word ...	Signature	fmt/19		pdf
New Text Document ...	Extension	x-fmt/111		txt

THE NATIONAL ARCHIVES

What is a PRONOM signature?

A signature attempts to define a series of ‘magic bytes’.

Patterns can show up in a specific part of a file, that can be used as a fingerprint to identify that a specific file conforms to a particular format.

Often these patterns occur at the beginning or end of the file. Often these patterns are detailed in technical specifications by software creators.

Digitalbevaring.dk

THE NATIONAL ARCHIVES

What is a PRONOM signature?

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

Details for: Acrobat PDF 1.5 - Portable Document Format 1.5 Save as... XML | CSV Print

Go to: Summary › | Documentation › | Signatures | Compression › | Character encoding › | Rights › | Reference files › Properties ›

Signatures

External signatures	File extension: pdf		
Internal signatures	**Name**	PDF 1.5	
	Description	BOF: %PDF-1.5 EOF (offset up to 1024 bytes): %%EOF	
	Byte sequences	**Position type**	Absolute from BOF
		Offset	0
		Byte order	
		Value	255044462D312E35
		Position type	Absolute from EOF
		Offset	0
		Maximum Offset	1024
		Byte order	
		Value	2525454F46

https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=618&strPageToDisplay=signatures

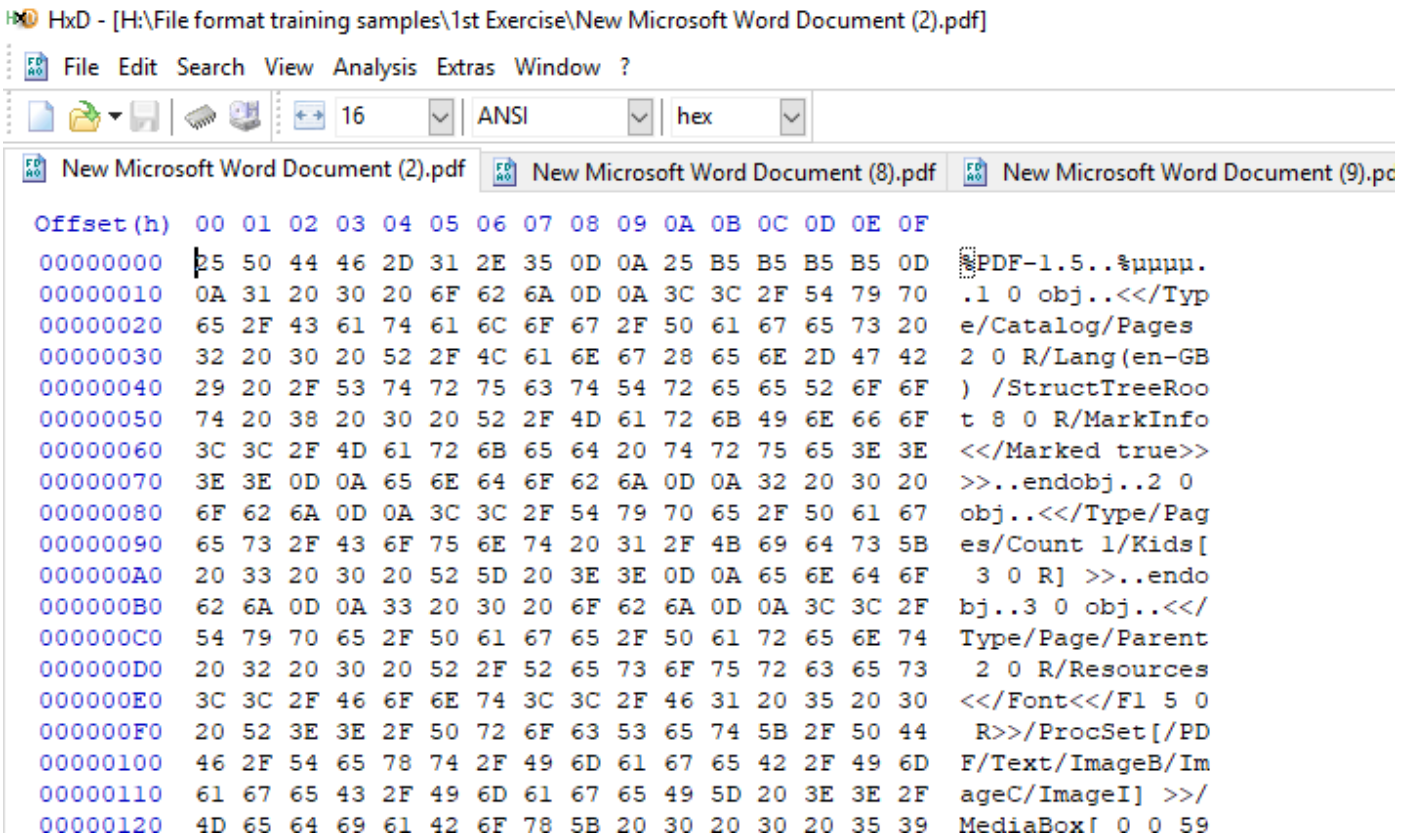
THE NATIONAL ARCHIVES

Reading byte streams

Hexadecimal numbers allow us to read binary data in a more human friendly manner.

Hex editors allow us to view the hexadecimal (16 characters, 0-9 & A-F) notation of a file.

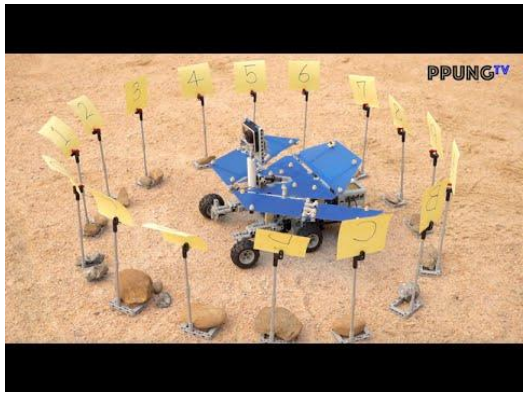
Each hex character represents half a byte or a ‘nibble’ – 4 bits



THE NATIONAL ARCHIVES

ASCII and Hexadecimal (via Matt Damon)

PUBLIC DOMAIN




```
48 4F 57 41 4C 49 56 45
H  O  W  A  L  I  V  E
```

Hex	Binary	Value
41	0100 0001	A
42	0100 0010	B
43	0100 0011	C
44	0100 0100	D
45	0100 0101	E
46	0100 0110	F
47	0100 0111	G
48	0100 1000	H
49	0100 1001	I
4A	0100 1010	J
4B	0100 1011	K
4C	0100 1100	L
4D	0100 1101	M
4E	0100 1110	N
4F	0100 1111	O
50	0101 0000	P
51	0101 0001	Q
52	0101 0010	R
53	0101 0011	S
54	0101 0100	T
55	0101 0101	U
56	0101 0110	V
57	0101 0111	W
58	0101 1000	X
59	0101 1001	Y
5A	0101 1010	Z

THE NATIONAL ARCHIVES

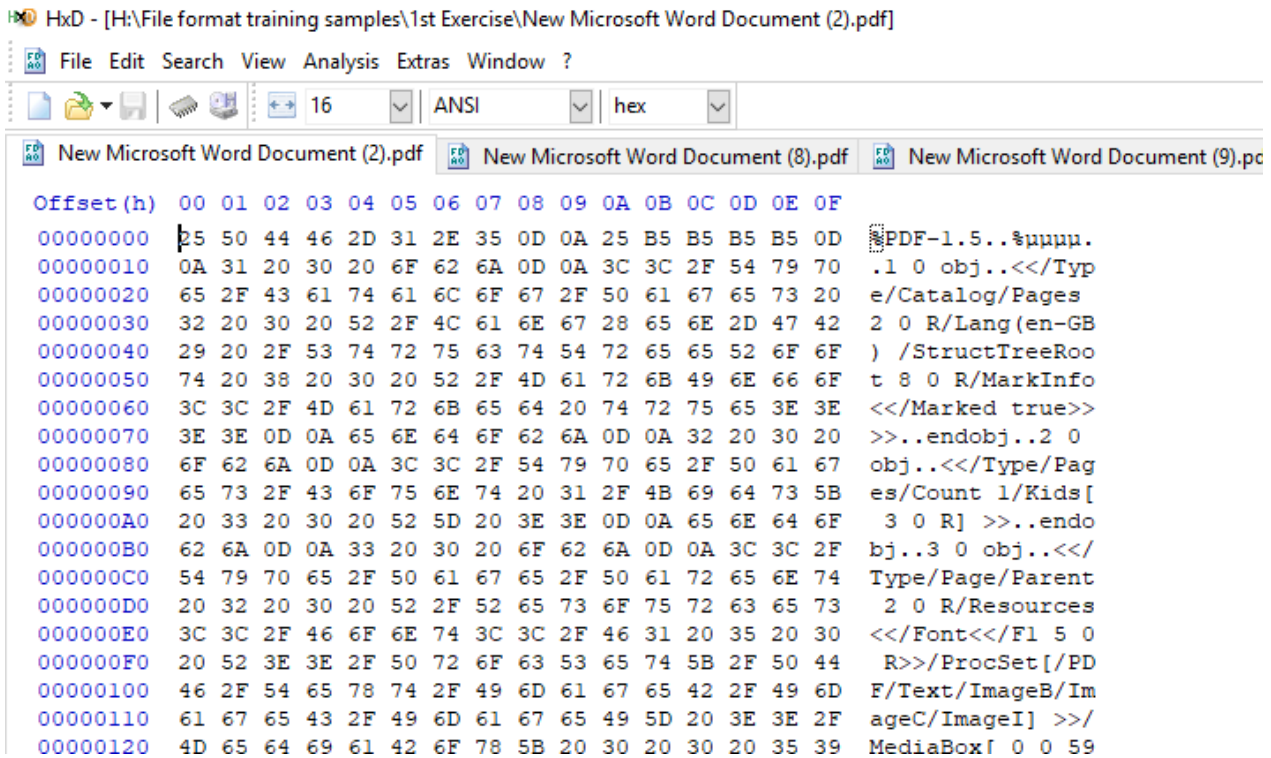
Hex Editors

Hex editors

- Windows - HxD
https://mh-nexus.de/en/hxd/

- OS X - HexFiend
http://ridiculousfish.com/hexfiend/

- Linux - Bless
https://apps.ubuntu.com/cat/applications/precise/bless/20



THE NATIONAL ARCHIVES

Exercise 2 – comparing signatures

- Open DROID (click on droid.bat on desktop)
- Run DROID over '2nd Exercise folder'
- Click on any PUID to bring up PRONOM page, click on signature tab
- Open HXD (on desktop)
- Drag file (same one you clicked on PUID for) into HXD
- Compare!

Photo by Robert Baker on Unsplash

THE NATIONAL ARCHIVES

PRONOM terms, basic syntax

BOF = Beginning of File.

EOF = End of File.

Var = Variable (anywhere in the file)

Offset/Max Offset = Exact or positional range in which a signature starts

Ranges –

{n} = specific number of wildcard bytes, e.g. A2{5}F3

{n-n} = range of wildcard bytes, e.g. 4D{0-12}E4

Byte range: [hh:hh] = single byte value between range, e.g [00:FA]

Either/or:(hhhh|hhhh|hh) = either/any or these byte values, e.g. (0D|0A|0D0A)

THE NATIONAL ARCHIVES

Container signatures

Some file formats use other formats as containers. In these cases, the format signature first needs to identify the container, and then identify the format using that container.

The two most common container formats are

- OLE2 - Used primarily by Microsoft in the 90s and 00s
- ZIP - Used by Microsoft and many others in the 00s to the present day

THE NATIONAL ARCHIVES

Exercise 3 - Microsoft Office XML

Open Office XML formats use ZIP as a container for XML

THE NATIONAL ARCHIVES

Microsoft Office XML

```
<ContainerSignature Id="1030" ContainerType="ZIP">
  <Description>Microsoft Word OOXML</Description>
  <Files>
    <File>
      <Path>[Content_Types].xml</Path>
      <BinarySignatures>
        <InternalSignatureCollection>
          <InternalSignature ID="302">
            <ByteSequence Reference="BOFoffset">
              <SubSequence Position="1" SubSeqMinOffset="0"
                SubSeqMaxOffset="32768">
                <Sequence>'ContentType="application/vnd.openxmlformats-officedocument.wordprocessingml.document.main+xml"'</Sequence>
              </SubSequence>
            </ByteSequence>
          </InternalSignature>
        </InternalSignatureCollection>
      </BinarySignatures>
    </File>
  </Files>
</ContainerSignature>
```

THE NATIONAL ARCHIVES

Photo by Dan Gold on Unsplash

Photo by Delaney Dawson on Unsplash

Text based file formats

(when identification by extension is the best PRONOM can do)

- Non-binary formats which don't have a consistent pattern which can be used for a PRONOM signature
- Most can only be identified by 'extension'
- Common formats you may recognise are 'txt' and 'csv'

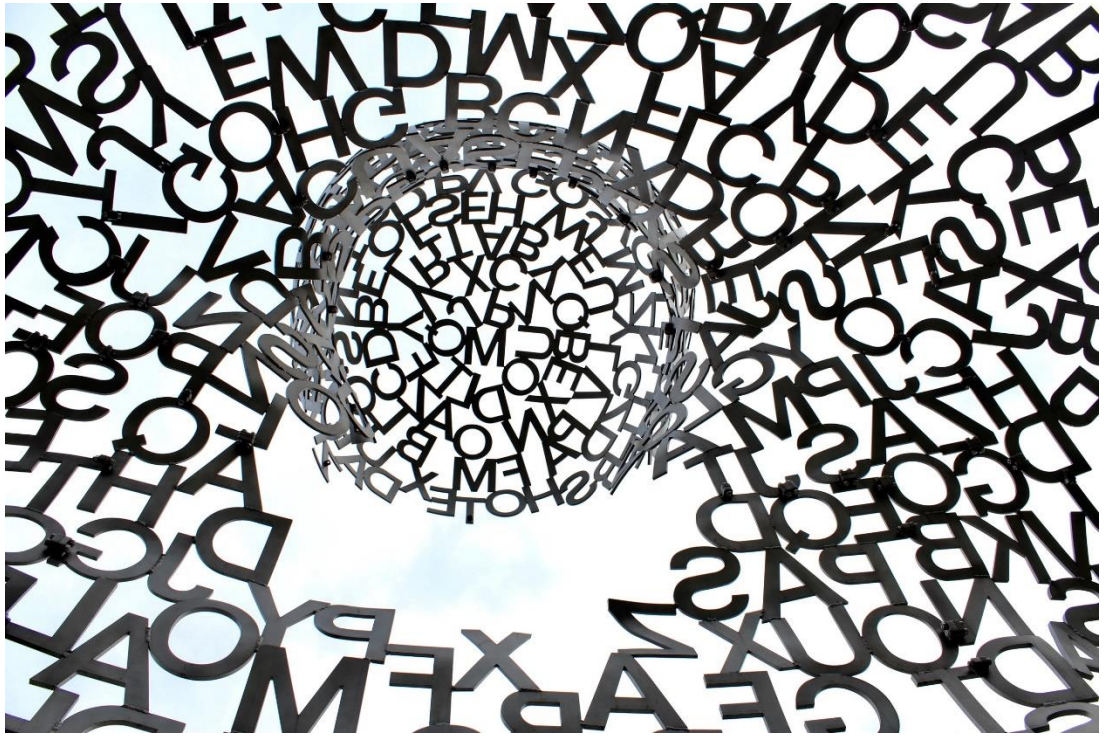


Photo by Nathaniel Shuman on Unsplash

THE NATIONAL ARCHIVES

Text files (txt)

Running DROID.txt - Notepad

File Edit Format View Help

```
====
    Copyright (c) 2016, The National Archives <pronom@nationalarchives.gsi.gov.uk>
    All rights reserved.

    Redistribution and use in source and binary forms, with or without
    modification, are permitted provided that the following
    conditions are met:

     * Redistributions of source code must retain the above copyright
       notice, this list of conditions and the following disclaimer.

     * Redistributions in binary form must reproduce the above copyright
       notice, this list of conditions and the following disclaimer in the
       documentation and/or other materials provided with the distribution.

     * Neither the name of the The National Archives nor the
       names of its contributors may be used to endorse or promote products
       derived from this software without specific prior written permission.

    THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS"
    AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
    IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
    PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR
    CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL,
    EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO,
    PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR
    PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF
    LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING
    NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS
    SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
====
```

HxD - [T:\pyoung\droid-binary-6.5-SNAPSHOT-bin-unix\Running DROID.txt]

File Edit Search View Analysis Extras Window ?

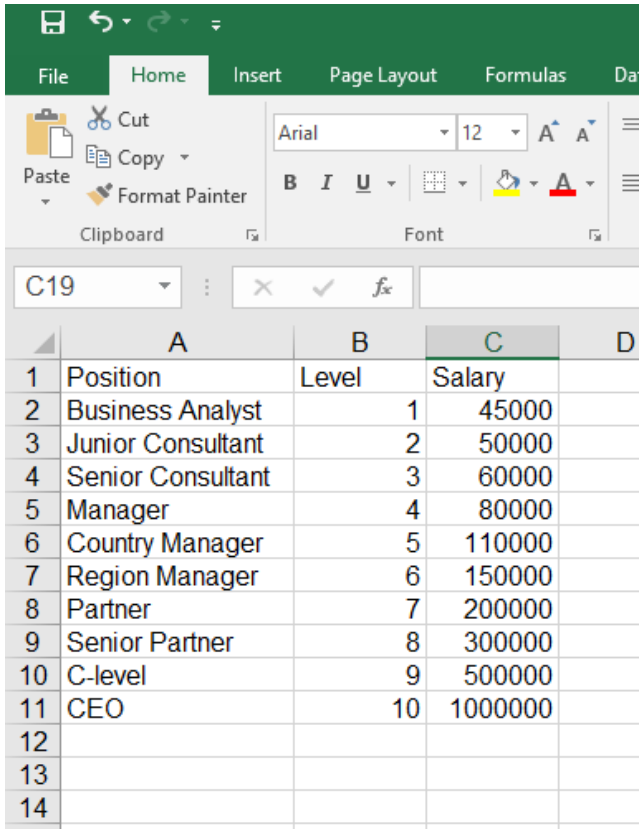
16 ANSI hex

Running DROID.txt

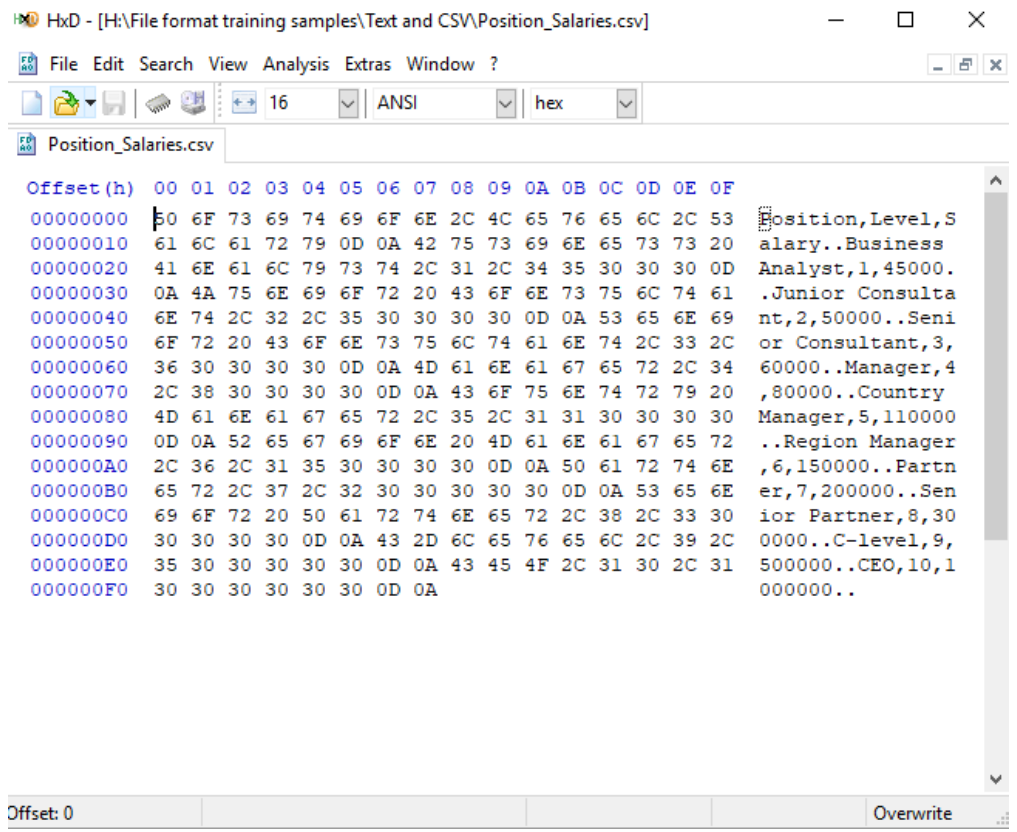
```
Offset(h)  00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F

00000000  3D 3D 3D 3D 0D 0A 20 20 20 20 43 6F 70 79 72 69  ====..    Copyri
00000010  67 68 74 20 28 63 29 20 32 30 31 36 2C 20 54 68  ght (c) 2016, Th
00000020  65 20 4E 61 74 69 6F 6E 61 6C 20 41 72 63 68 69  e National Archi
00000030  76 65 73 20 3C 70 72 6F 6E 6F 6D 40 6E 61 74 69  ves <pronom@nati
00000040  6F 6E 61 6C 61 72 63 68 69 76 65 73 2E 67 73 69  onalarchives.gsi
00000050  2E 67 6F 76 2E 75 6B 3E 0D 0A 20 20 20 20 41 6C  .gov.uk>..    Al
00000060  6C 20 72 69 67 68 74 73 20 72 65 73 65 72 76 65  l rights reserve
00000070  64 2E 0D 0A 0D 0A 20 20 20 20 52 65 64 69 73 74  d.....    Redist
00000080  72 69 62 75 74 69 6F 6E 20 61 6E 64 20 75 73 65  ribution and use
00000090  20 69 6E 20 73 6F 75 72 63 65 20 61 6E 64 20 62   in source and b
000000A0  69 6E 61 72 79 20 66 6F 72 6D 73 2C 20 77 69 74  inary forms, wit
000000B0  68 20 6F 72 20 77 69 74 68 6F 75 74 0D 0A 20 20  h or without..
000000C0  20 20 6D 6F 64 69 66 69 63 61 74 69 6F 6E 2C 20    modification,
000000D0  61 72 65 20 70 65 72 6D 69 74 74 65 64 20 70 72  are permitted pr
000000E0  6F 76 69 64 65 64 20 74 68 61 74 20 74 68 65 20  ovided that the
000000F0  66 6F 6C 6C 6F 77 69 6E 67 0D 0A 20 20 20 20 63  following..    c
00000100  6F 6E 64 69 74 69 6F 6E 73 20 61 72 65 20 6D 65  onditions are me
00000110  74 3A 0D 0A 0D 0A 20 20 20 20 20 2A 20 52 65 64  t:....     * Red
00000120  69 73 74 72 69 62 75 74 69 6F 6E 73 20 6F 66 20  istributions of
00000130  73 6F 75 72 63 65 20 63 6F 64 65 20 6D 75 73 74  source code must
00000140  20 72 65 74 61 69 6E 20 74 68 65 20 61 62 6F 76   retain the abov
00000150  65 20 63 6F 70 79 72 69 67 68 74 0D 0A 20 20 20  e copyright..
00000160  20 20 20 20 6E 6F 74 69 63 65 2C 20 74 68 69 73      notice, this
00000170  20 6C 69 73 74 20 6F 66 20 63 6F 6E 64 69 74 69   list of conditi
00000180  6F 6E 73 20 61 6E 64 20 74 68 65 20 66 6F 6C 6C  ons and the foll
00000190  6F 77 69 6E 67 20 64 69 73 63 6C 61 69 6D 65 72  owing disclaimer
000001A0  2E 0D 0A 0D 0A 20 20 20 20 2A 20 52 65 64 69  .....     * Redi
000001B0  73 74 72 69 62 75 74 69 6F 6E 73 20 69 6E 20 62  stributions in b
000001C0  69 6E 61 72 79 20 66 6F 72 6D 20 6D 75 73 74 20  inary form must
000001D0  72 65 70 72 6F 64 75 63 65 20 74 68 65 20 61 62  reproduce the ab
000001E0  6F 76 65 20 63 6F 70 79 72 69 67 68 74 0D 0A 20  ove copyright..
000001F0  20 20 20 20 20 6E 6F 74 69 63 65 2C 20 74 68        notice, th
00000200  69 73 20 6C 69 73 74 20 6F 66 20 63 6F 6E 64 69  is list of condi
00000210  74 69 6F 6E 73 20 61 6E 64 20 74 68 65 20 66 6F  tions and the fo
00000220  6C 6C 6F 77 69 6E 67 20 64 69 73 63 6C 61 69 6D  llowing disclaim
00000230  65 72 20 69 6E 20 74 68 65 0D 0A 20 20 20 20 20  er in the..
00000240  20 20 64 6F 63 75 6D 65 6E 74 61 74 69 6F 6E 20    documentation
00000250  61 6E 64 2F 6F 72 20 6F 74 68 65 72 20 6D 61 74  and/or other mat
```

THE NATIONAL ARCHIVES

Comma Separated Values (CSV)



	A	B	C	D
1	Position	Level	Salary	
2	Business Analyst	1	45000	
3	Junior Consultant	2	50000	
4	Senior Consultant	3	60000	
5	Manager	4	80000	
6	Country Manager	5	110000	
7	Region Manager	6	150000	
8	Partner	7	200000	
9	Senior Partner	8	300000	
10	C-level	9	500000	
11	CEO	10	1000000	
12				
13				
14				



THE NATIONAL ARCHIVES

Exercise 4

- Drag text files into hex editor to see difficulties in identifying patterns for these formats

Photo by roya ann miller on Unsplash

THE NATIONAL ARCHIVES

Photo by Raoul Droog on Unsplash

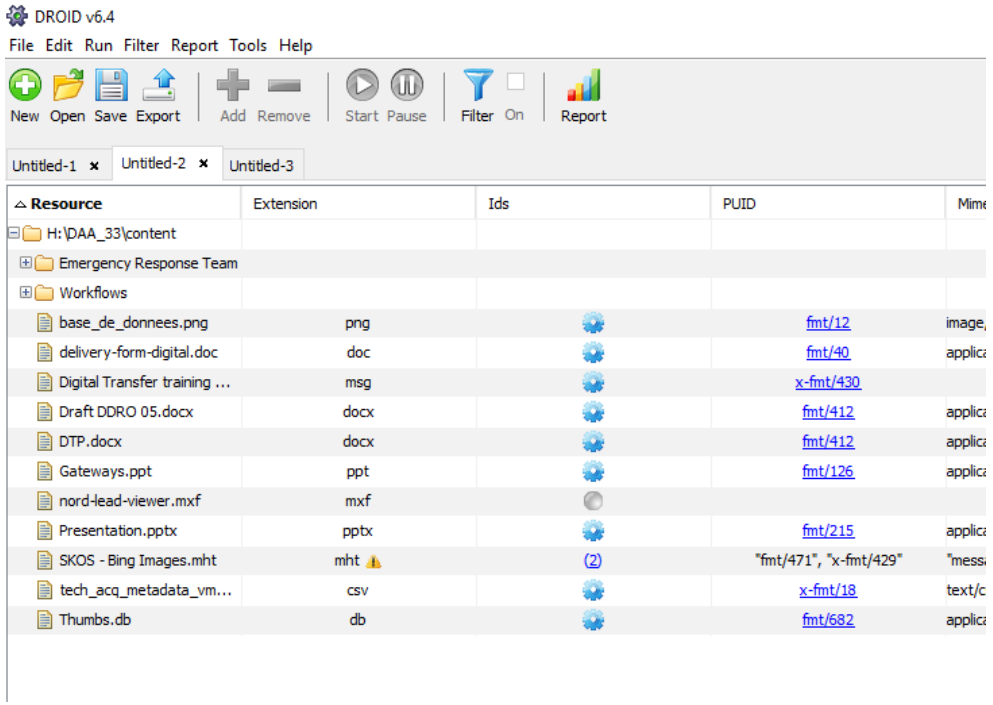
THE NATIONAL ARCHIVES

Unidentified format?

- Grey circle?
- Or identifying as multiple formats?

PRONOM covers lots of formats but there are always more, and sometimes need updating!

Community has a big part to play in reporting new formats or spotting errors in PRONOM



THE NATIONAL ARCHIVES

Sources of help

PRONOM mailbox - PRONOM@nationalarchives.gov.uk

PRONOM request form - www.nationalarchives.gov.uk/contact-us/submit-information-for-pronom

PRONOM Google group - https://groups.google.com/forum/#!forum/pronom

DROID Google group – https://groups.google.com/forum/#!forum/droid-list

Twitter - #PRONOM #DROID

THE NATIONAL ARCHIVES

Why is it not identifying?

- New format never seen before
- New version of an existing format
- Existing format which is not identifying correctly (due to PRONOM error)
- Corrupt format (signature not working because identifying sections are missing or corrupted)

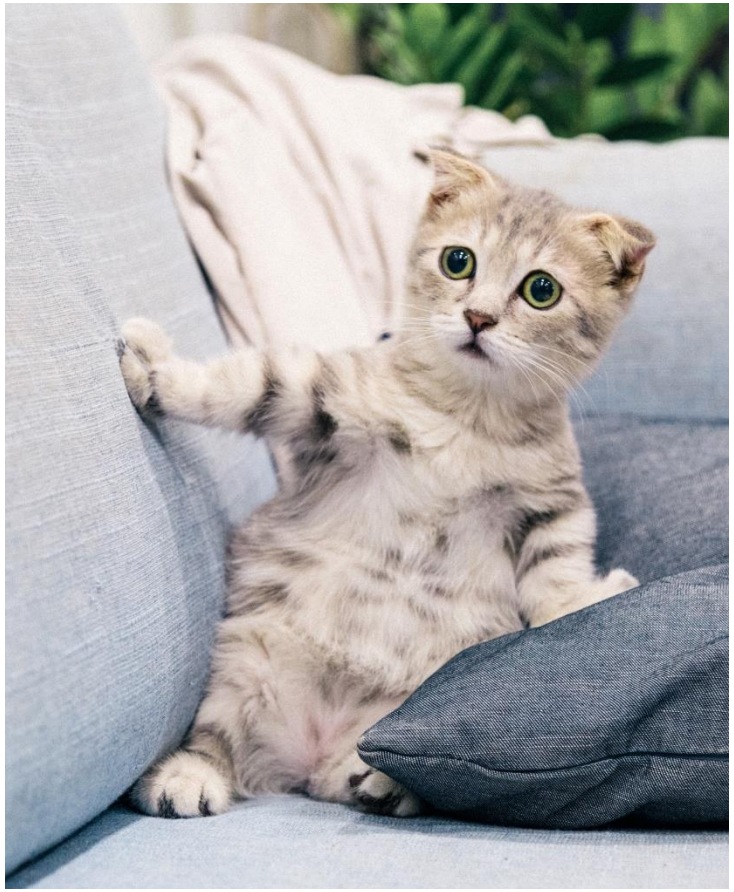


Photo by Tran Mau Tri Tam on Unsplash

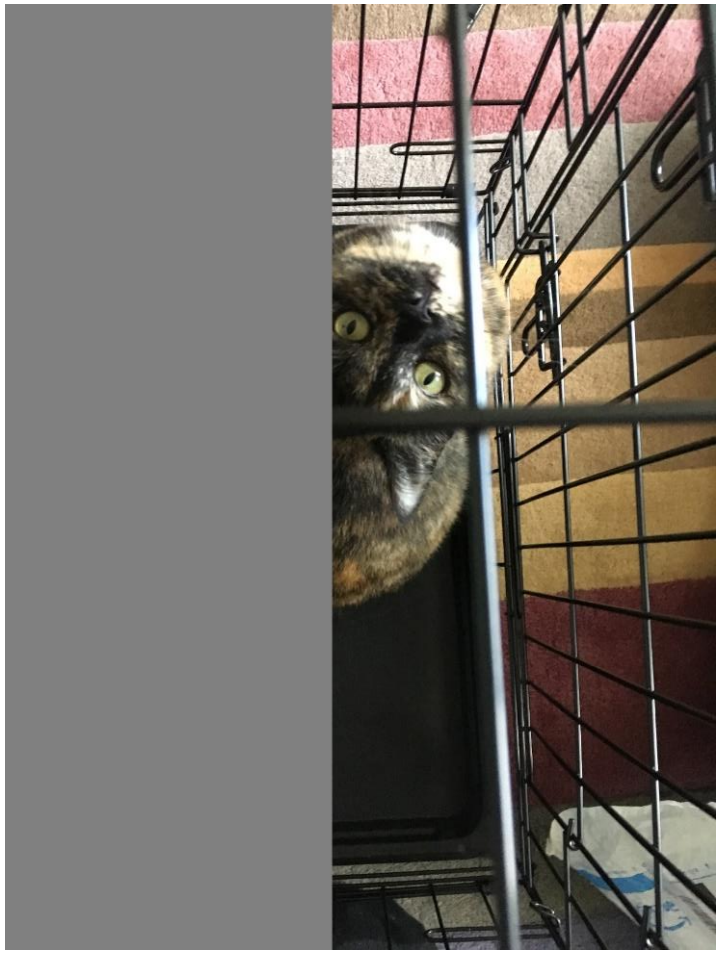
THE NATIONAL ARCHIVES

Corrupt?

- DROID and PRONOM do not validate formats, so will not always pick up corruption but if identifying sections are missing the format will not identify

Untitled-1 ×

Resource	Ids	PUID
H:\File format training samples\2nd Exercise		
cat.JPG		fmt/645
corrupt penguins.jpg		
corrupt_cat.jpg		
penguins.jpg		fmt/43



corrupt_cat.jpg

THE NATIONAL ARCHIVES

Exercise 5

- Run DROID over '5th Exercise' folder
- Click on PUID for 'penguins.jpg'
- Compare signature in PRONOM with 'corrupt_penguins.jpg' file, to see the difference

THE NATIONAL ARCHIVES

Not corrupt but a new format?
Contacting PRONOM

- We receive submissions which range from ‘this file won’t identify and I don’t know what it is,’ through to ‘here is everything there is to know about this format, including signature and samples!’
- Ideally we’d like a format name & samples as a minimum (but don’t let that put you off!)
- We are happy to look into any format but the more you can offer us the quicker it will appear on PRONOM

THE NATIONAL ARCHIVES

What do we want to know?

- Format name
- Format description
- Format extension(s)
- Specification links (if known)
- Links to any sites with relevant information
- MIME/Media type (check IANA)
- Vendor (if known)
- Format signature
- How you wish to be attributed

Photo by Agence Olloweb on Unsplash

THE NATIONAL ARCHIVES

Sending samples to PRONOM

- Samples can be sent via email PRONOM@nationalarchives.gov.uk
- We are happy to sign non-disclosure agreements that samples will be used for PRONOM research purposes only

THE NATIONAL ARCHIVES

File format research process overview

1. Check common file format sites
2. Locate sample files, and scan with latest DROID sig file
3. Observe common byte sequences
4. Locate technical specification (where available) – use to confirm findings

THE NATIONAL ARCHIVES

1.File format sites

Gary Kessler - https://www.garykessler.net/library/file_sigs.html

FileExt - https://filext.com/

As well as

LOC – http://www.loc.gov/preservation/digital/formats/index.html

File Formats Wiki –http://fileformats.archiveteam.org/wiki/Main_Page

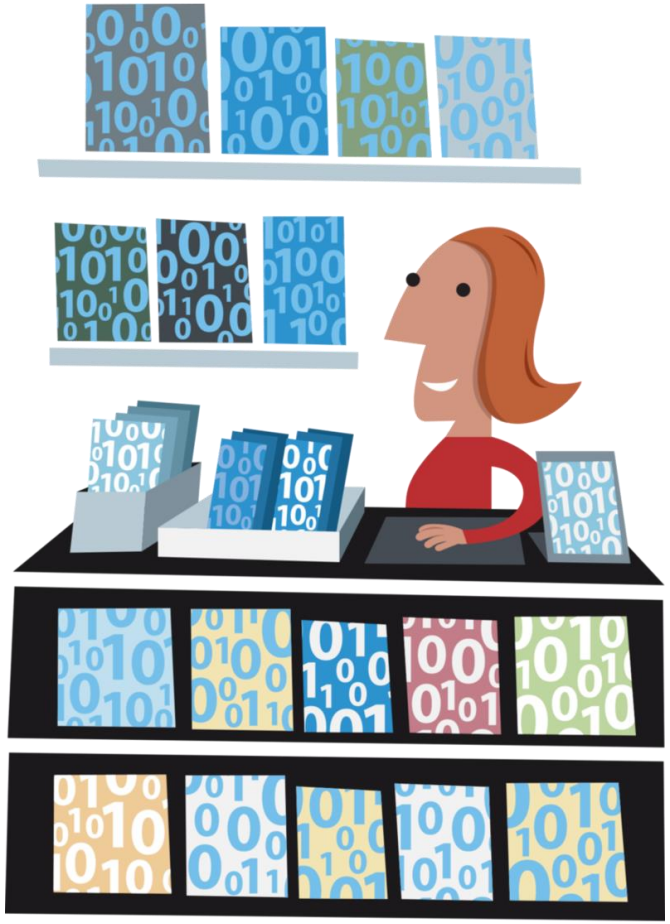
TRID – http://mark0.net/soft-trid-e.html

AND - http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

THE NATIONAL ARCHIVES

2.Samples

- Want a range of samples if possible from different sources
- Trial versions of software are a good source of creating samples yourself
- Ask community for samples
- Be careful of formats with the same extensions but totally different formats – fileext.com is a good source for information on this



Digitalbevaring.dk

THE NATIONAL ARCHIVES

Handy sample searching techniques

- Searching in Google -

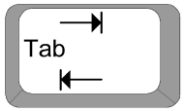
“parent directory” .jpg = good google search for websites holding lots of files

filetype:jpg = another good approach for locating samples

Be careful with Googling! And downloading, check what you are clicking on

THE NATIONAL ARCHIVES

3. Observing patterns
Exercise 6(TZX format)

- Open folder '6th Exercise' with tzx samples, gathered from https://www.worldofspectrum.org/
- Drag into Hex editor
- Cycle files holding Ctrl + 

- Observe any patterns you see

THE NATIONAL ARCHIVES

```
Offset(h)  00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F

00000000   5A 58 54 61 70 65 21 1A 01 01 10 E8 03 13 00 00  ZXTape!....è....
00000010   00 42 6C 6F 63 6B 44 69 7A 7A 79 A8 00 01 00 A8  .BlockDizzy¨...¨
00000020   00 1C 10 E8 03 AA 00 FF 00 01 32 00 E7 30 0E 00  ...è.ª.ÿ..2.ç0..
00000030   00 00 00 00 3A DA 30 0E 00 00 00 00 00 3A FD 32  ....:Ú0......:ý2
00000040   35 33 32 37 0E 00 00 EF 62 00 3A EF 22 22 AF 3A  5327...ïb.:ï""¯:
00000050   F9 C0 36 35 31 34 36 0E 00 00 7A FE 00 0D 00 14  ùÀ65146...zþ....
00000060   6E 00 FD 32 35 33 32 37 0E 00 00 EF 62 00 3A EF  n.ý25327...ïb.:ï
00000070   22 62 64 6D 63 22 AF 3A EF 22 62 64 73 76 6C 64  "bdmc"¯:ï"bdsvld
00000080   22 AF 3A F8 22 74 3A 22 3A F8 22 42 6C 6F 63 6B  "¯:ø"t:":ø"Block
00000090   44 69 7A 7A 79 22 CA 31 0E 00 00 01 00 00 3A F8  Dizzy"Ê1......:ø
000000A0   22 78 22 AF 36 35 31 34 36 0E 00 00 7A FE 00 2C  "x"¯65146...zþ.,
000000B0   31 39 30 0E 00 00 BE 00 00 3A EF 22 62 64 24 22  190...¾..:ï"bd$"
000000C0   AA 3A F9 C0 36 35 30 30 30 0E 00 00 E8 FD 00 0D  ª:ùÀ65000...èý..
000000D0   91 10 E8 03 13 00 00 03 78 20 20 20 20 20 20 20  `.è.....x
000000E0   20 20 BE 00 7A FE A8 80 49 10 E8 03 C0 00 FF F3    ¾.zþ¨€I.è.À.ÿó

Offset(h)  00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F

00000000   5A 58 54 61 70 65 21 1A 01 0A 30 1C 43 72 65 61  ZXTape!...0.Crea
00000010   74 65 64 20 77 69 74 68 20 52 61 6D 73 6F 66 74  ted with Ramsoft
00000020   20 4D 61 6B 65 54 5A 58 10 34 03 14 00 00 00 4C   MakeTZX.4.....L
00000030   4F 41 44 45 52 20 20 20 20 58 02 01 00 58 02 10  OADER    X...X..
00000040   00 10 EA 03 5A 02 FF 00 01 09 00 F9 30 0E 00 00  ..ê.Z.ÿ....ù0...
00000050   00 00 00 0D 00 0A 25 01 E7 30 0E 00 00 00 00 00  ......%.ç0......
00000060   3A DA 30 0E 00 00 00 00 00 3A D9 30 0E 00 00 00  :Ú0......:Ù0....
00000070   00 00 3A FD 32 34 39 39 39 0E 00 00 A7 61 00 3A  ..:ý24999...§a.:
00000080   F5 DB 31 0E 00 00 01 00 00 3B D9 37 0E 00 00 07  õÛ1......;Ù7....
00000090   00 00 27 27 27 22 53 54 4F 50 20 54 48 45 20 54  ..'''"STOP THE T
000000A0   41 50 45 20 42 55 54 20 44 4F 20 4E 4F 54 20 52  APE BUT DO NOT R
000000B0   45 57 49 4E 44 21 22 3A D7 31 0E 00 00 01 00 00  EWIND!":×1......
000000C0   2C 31 30 0E 00 00 0A 00 00 3A F5 D9 37 0E 00 00  ,10......:õÙ7...
000000D0   07 00 00 3B 22 49 4E 53 45 52 54 20 54 48 45 20  ...;"INSERT THE
000000E0   42 4C 4F 43 4B 42 55 53 54 45 52 53 20 54 41 50  BLOCKBUSTERS TAP
```

4.Technical specifications

- If written by vendor then can confirm patterns seen in sample files
- Can be found via Google, on official websites or for older software sometimes can be searching on the Internet Archive
- Do not always exist, for some proprietary software companies do not want to expose information for commercial reasons
- Sometimes ‘unofficial’ specifications may exist, created by software developers working on applications using the format. ‘GitHub’ is a good source for these, cannot be trusted in the same way but can be helpful in signature development
- Technical specifications vary a lot in how easy they are to read, don’t need to read the whole thing! Just scan for any information on format headers and signatures
- You do not need to find a technical specification to create a signature, but it helps

THE NATIONAL ARCHIVES

TZX Header			
length: 10 bytes			
Offset	**Value**	**Type**	**Description**
0x00	"ZXTape!"	ASCII[7]	TZX signature
0x07	0x1A	BYTE	End of text file marker
0x08	1	BYTE	TZX major revision number
0x09	20	BYTE	TZX minor revision number

http://www.worldofspectrum.org/TZXformat.html

```
Offset(h)  00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F

00000000   5A 58 54 61 70 65 21 1A 01 01 10 E8 03 13 00 00  ZXTape!....è....
00000010   00 42 6C 6F 63 6B 44 69 7A 7A 79 A8 00 01 00 A8  .BlockDizzy¨...¨
00000020   00 1C 10 E8 03 AA 00 FF 00 01 32 00 E7 30 0E 00  ...è.ª.ÿ..2.ç0..
00000030   00 00 00 00 3A DA 30 0E 00 00 00 00 00 3A FD 32  ....:Ú0......:ý2
00000040   35 33 32 37 0E 00 00 EF 62 00 3A EF 22 22 AF 3A  5327...ïb.:ï""¯:
00000050   F9 C0 36 35 31 34 36 0E 00 00 7A FE 00 0D 00 14  ùÀ65146...zþ....
00000060   6E 00 FD 32 35 33 32 37 0E 00 00 EF 62 00 3A EF  n.ý25327...ïb.:ï
00000070   22 62 64 6D 63 22 AF 3A EF 22 62 64 73 76 6C 64  "bdmc"¯:ï"bdsvld
00000080   22 AF 3A F8 22 74 3A 22 3A F8 22 42 6C 6F 63 6B  "¯:ø"t:":ø"Block
00000090   44 69 7A 7A 79 22 CA 31 0E 00 00 01 00 00 3A F8  Dizzy"Ê1......:ø
000000A0   22 78 22 AF 36 35 31 34 36 0E 00 00 7A FE 00 2C  "x"¯65146...zþ.,
000000B0   31 39 30 0E 00 00 BE 00 00 3A EF 22 62 64 24 22  190...¾..:ï"bd$"
000000C0   AA 3A F9 C0 36 35 30 30 30 0E 00 00 E8 FD 00 0D  ª:ùÀ65000...èý..
000000D0   91 10 E8 03 13 00 00 03 78 20 20 20 20 20 20 20  `.è.....x
000000E0   20 20 BE 00 7A FE A8 80 49 10 E8 03 C0 00 FF F3    ¾.zþ¨€I.è.À.ÿó
```

Some less perfect examples – Jeff's image format

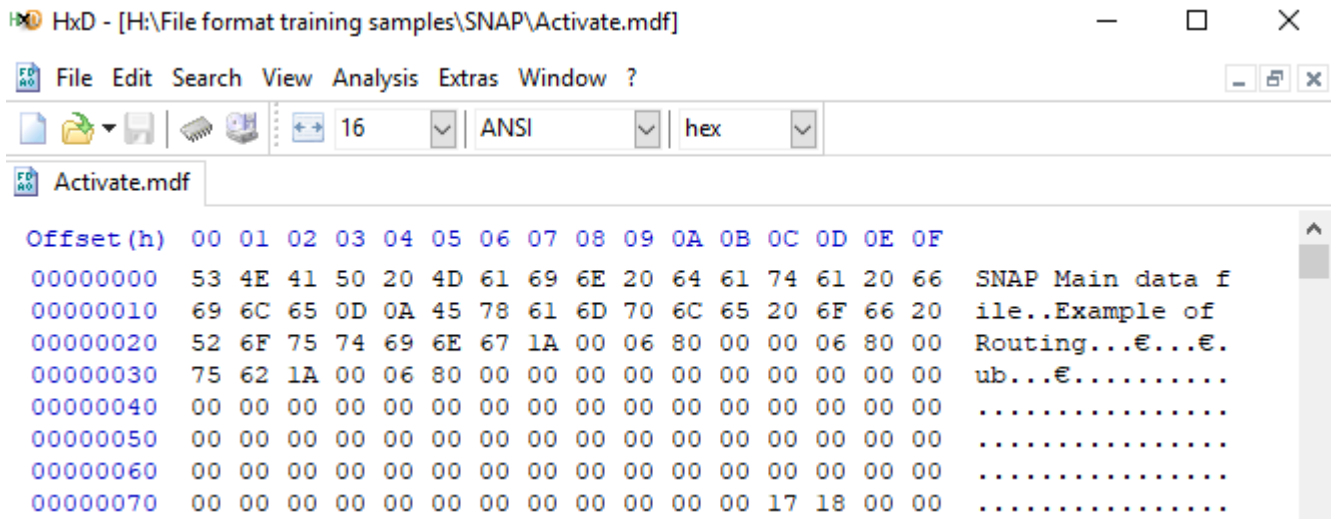
- Jeff created his format to be a rival for GIF – as he was worried about patent for compression used in GIF
- Samples very hard to find
- Creator had made technical specification (available via web archive)
- As it is from creator of the format was deemed enough for a signature

	GIF	JIF
extension	gif	jif
type (for Macintosh files)	GIFf	JIFf
signature (first six bytes of file)	GIF87a or GIF89a	JIF99a
compression	LZW	LZ77 derivative

THE NATIONAL ARCHIVES

Some less perfect examples – Snap Survey

- Snap Survey, doesn't publish technical specifications
- Had a small sample set, trial software available to create more files
- Clear header can be seen in samples
- Contacted software provider to confirm findings



THE NATIONAL ARCHIVES

Photo by [Kate Stone Matheson](#) on [Unsplash](#)

What's next? PRONOM release process

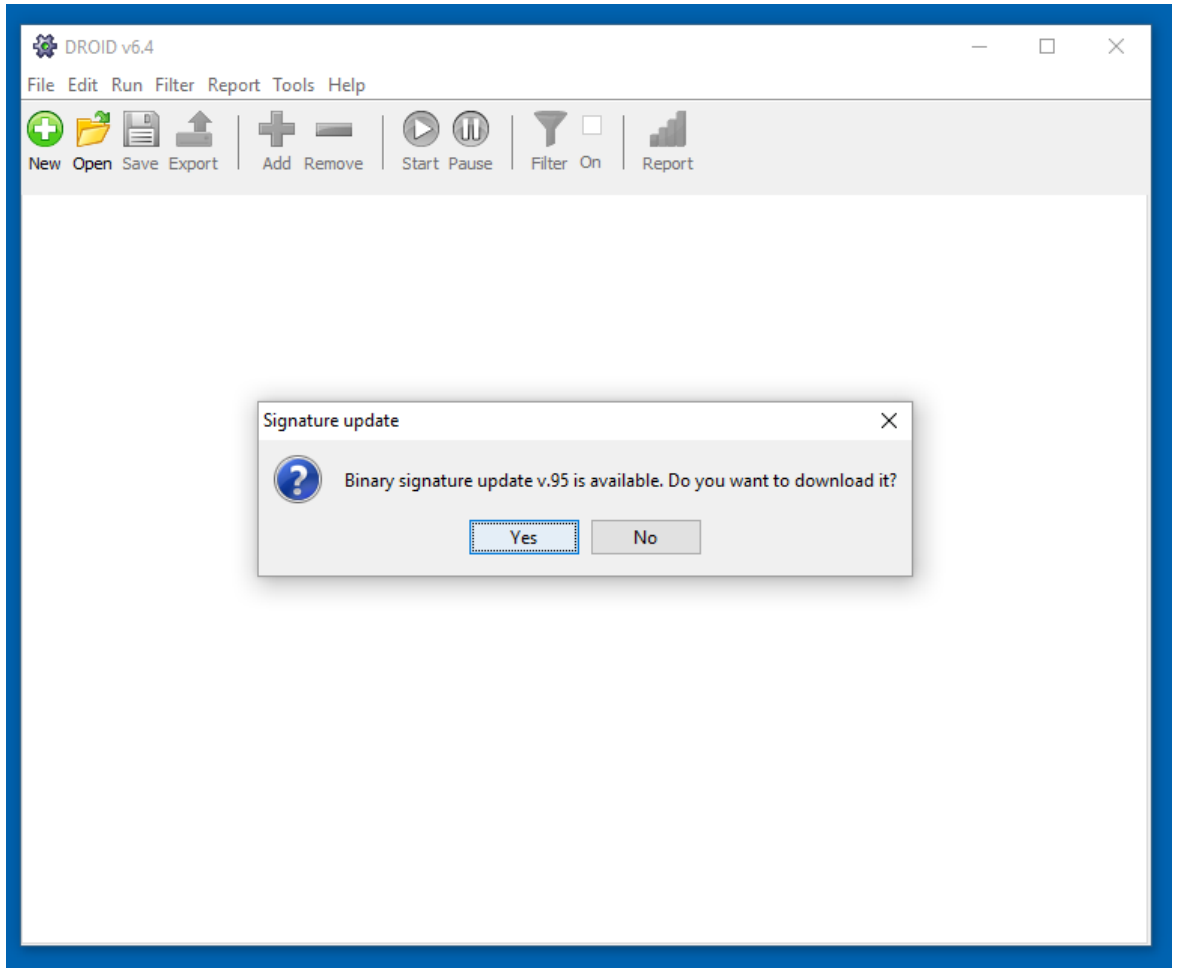
- Generate test signatures
- Check for any clashes with existing signatures
- Assign any priorities needed
- Release!

THE NATIONAL ARCHIVES

Priorities

Orientation	Binary
Byte order	Big-endian (Motorola)
Related file formats	Has lower priority than Acrobat PDF/A - Portable Document Format (1a) Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1:1999 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1:2001 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1a:2003 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 2:2003 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 3:2003 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 1a:2001 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange 3:2002 Has lower priority than Acrobat PDF/A - Portable Document Format (2a) Has lower priority than Acrobat PDF/A - Portable Document Format (2b) Has lower priority than Acrobat PDF/A - Portable Document Format (2u) Has lower priority than Acrobat PDF/A - Portable Document Format (3a) Has lower priority than Acrobat PDF/A - Portable Document Format (3b) Has lower priority than Acrobat PDF/A - Portable Document Format (3u) Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4p Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5g Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5pg Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5n Has lower priority than Acrobat PDF/E - Portable Document Format for Engineering PDF/E-1 Has priority over TrueType Font Is previous version of Acrobat PDF 1.1 - Portable Document Format (1.1)
Technical Environment	

Updates - Always use the latest release



THE NATIONAL ARCHIVES


Photo by Lee Cartledge on Unsplash

Digital preservation systems and PRONOM updates

- May take longer for digital preservation systems to add the latest PRONOM signature release
- PRONOM gets better over time, if storing material in a digital preservation systems you may want to run file format identification tools over already ingested files to get the latest format identifications ‘re-characterisation’

THE NATIONAL ARCHIVES

Homework!

What are we looking for in a DROID report?

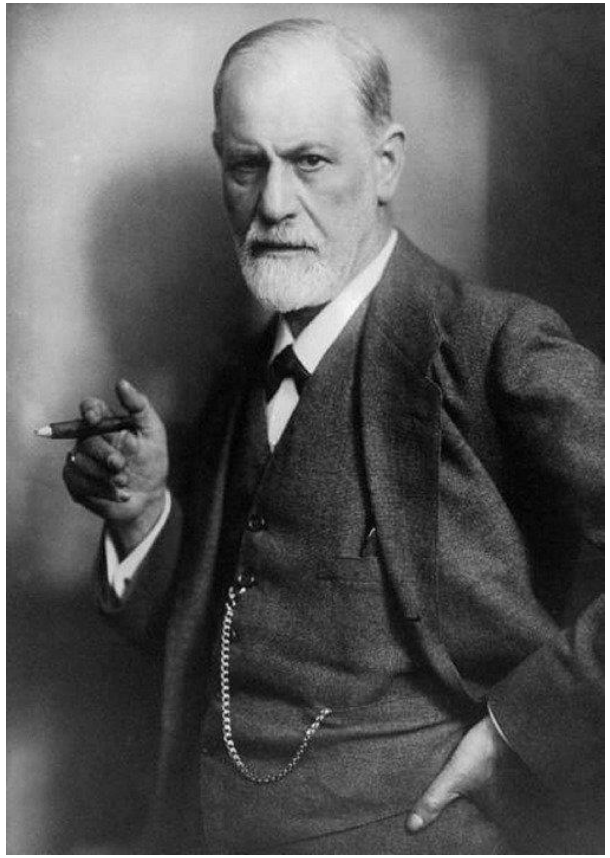
We analyse the report for:

- **Unidentified files**
- **Files identifying by extension only**
- **Files identifying as multiple formats**
- **Extension mismatches**
- **Zip files**
- **Identified files not on our white list**
- **Password protected files**
- **Duplicates**

THE NATIONAL ARCHIVES

Freud

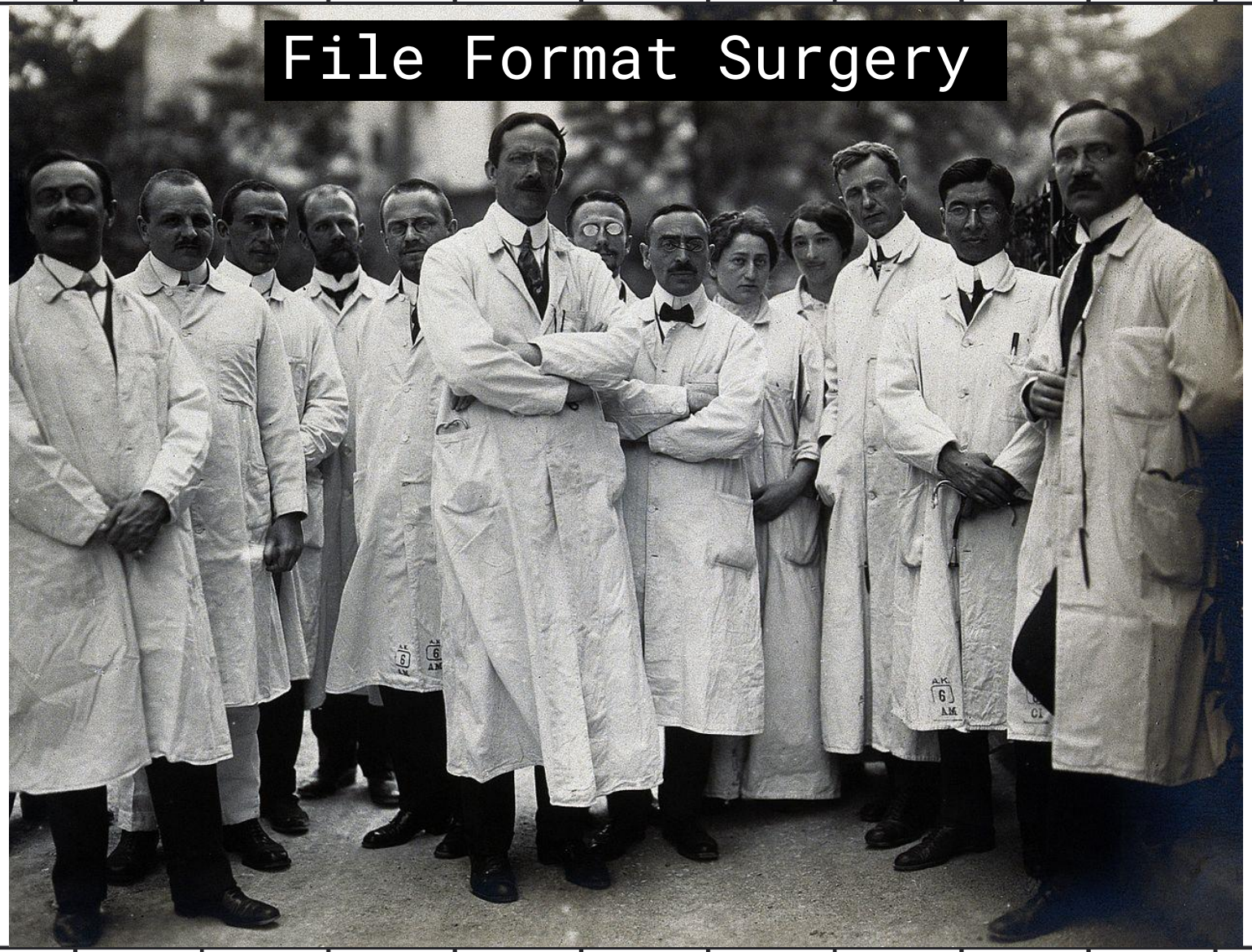
Freud is a tool we use to identify potential file format issues which we can flag from a DROID CSV export



https://github.com/digital-preservation/freud

THE NATIONAL ARCHIVES

File Format Surgery



THE NATIONAL ARCHIVES

Photo by Dan Gold on Unsplash

Homework 1 of 2

Review following materials (2-3)

- How to research and develop signatures for file format identification (TNA, 2011)
- Hacking the Droid Signature File for Characterization (Ross Spencer, 2012)
- Making the Switch from User to User AND contributor: My First File Format Signature (Andrea Byrne, 2016)
- My first file format signature (Jenny Mitcham, 2016)
- DROID Container Signature Files: What they are and how to create them: A template and an example, or few... (Ross Spencer, 2016)

Photo by JESHOOTS.COM on Unsplash

THE NATIONAL ARCHIVES

Homework 2 of 2

Check technical specification ‘7zFormat.txt’ for any sections which will confirm signature pattern

- Think about key words which may help to identify correct sections – eg. ‘header’ or ‘signature’

https://fastapi.metacpan.org/source/BJOERN/Compress-Deflate7-1.0/7zip/DOC/7zFormat.txt

THE NATIONAL ARCHIVES

Photo by Morvanic Lee on Unsplash

ArchiveSchool@nationalarchives.gov.uk

paul.young@nationalarchives.gov.uk

@pmyoung84

Feedback Survey for Session 2

https://tinyurl.com/y3vo5h8t

THE NATIONAL ARCHIVES

Next session - Describing what you have

31 January 2020

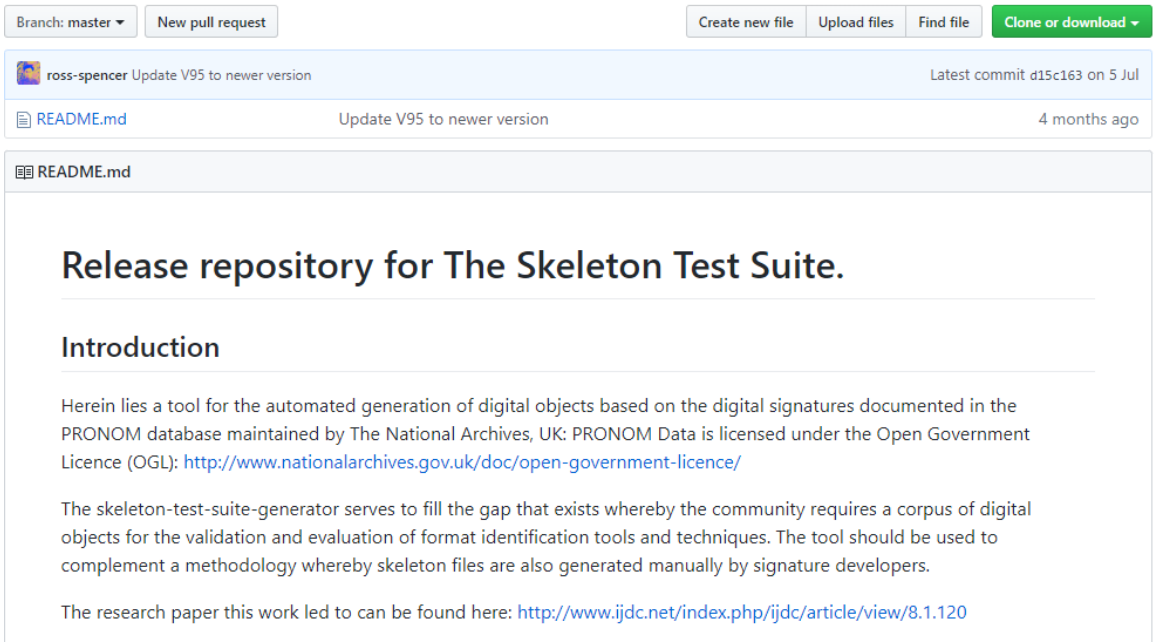
How to use tools to extract metadata from your digital files and an overview of the types of metadata validation carried out at The National Archives. A practical session on the use of a metadata extraction tool.

THE NATIONAL ARCHIVES

Skeleton Test Suite

Test against a further file format tool which automatically generates sample files based on the digital signatures documented in the PRONOM database we don't want new signatures to clash with old ones

You probably don't have examples of all 2000 PRONOM formats, so Ross Spencer cooked up some samples

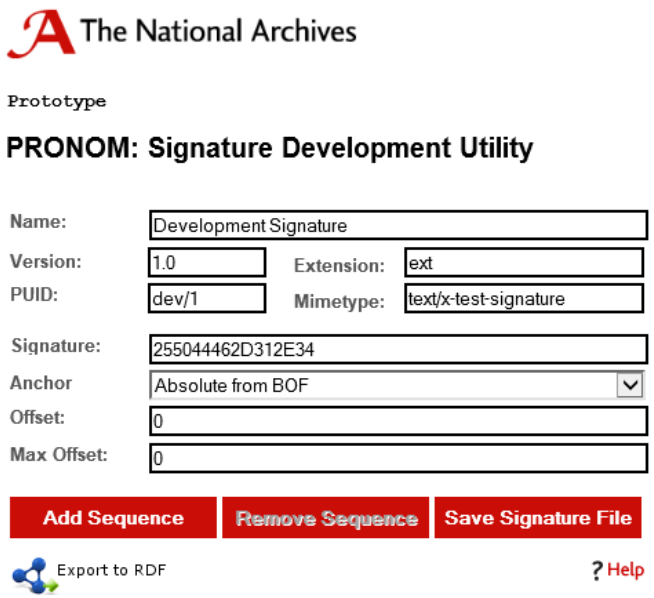


https://github.com/exponential-decay/pronom-archive-and-skeleton-test-suite

THE NATIONAL ARCHIVES

Test signatures

- Generate test signatures using - http://www.nationalarchives.gov.uk/pronom/sigdev/index.htm
- Install signature file and test over sample files



THE NATIONAL ARCHIVES