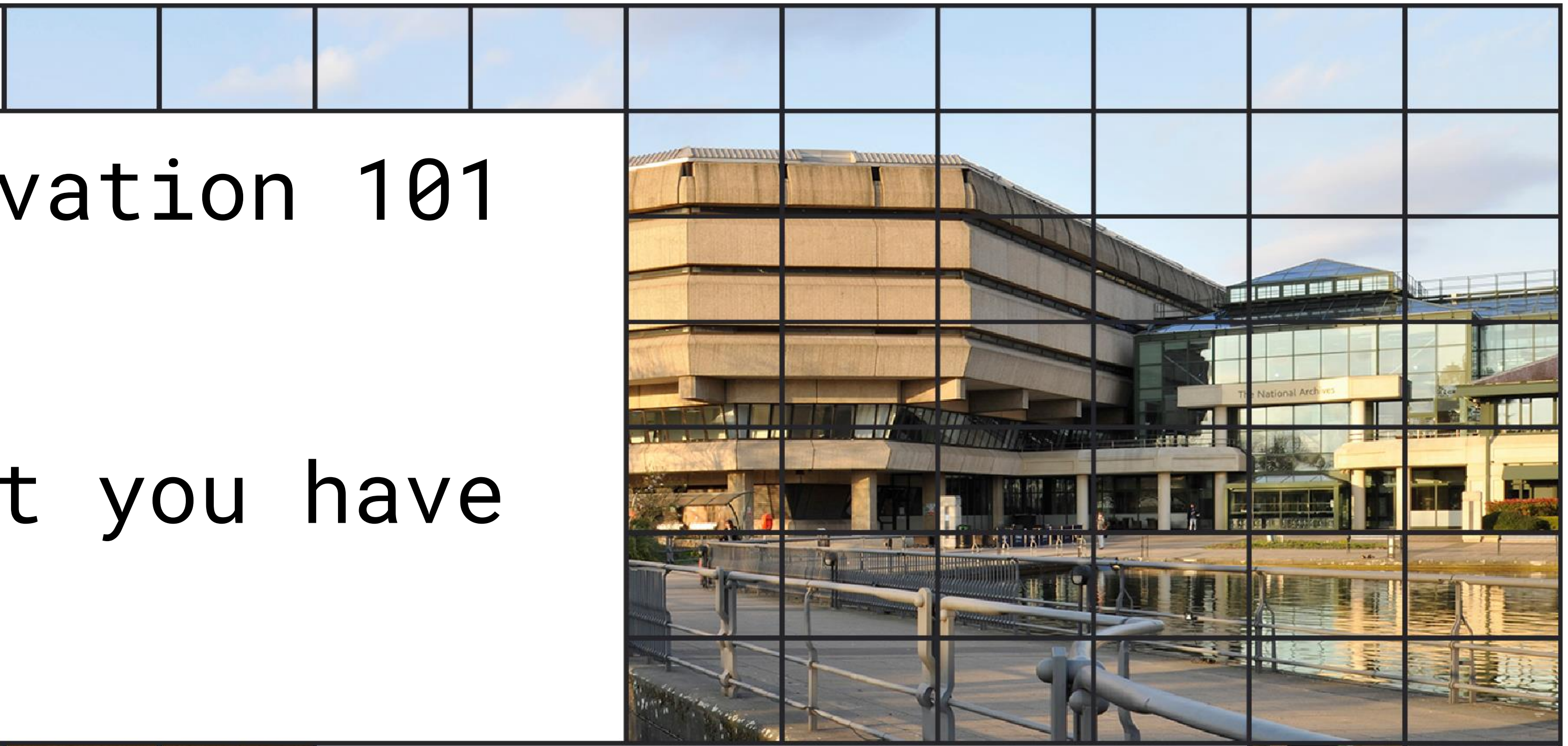
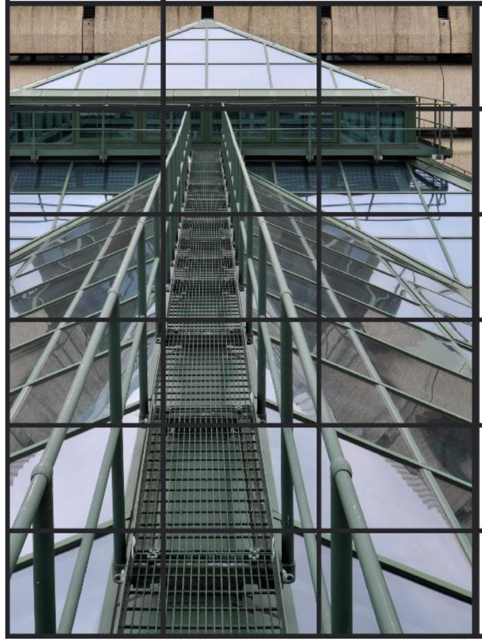
Digital Preservation 101

Session 3: Describing what you have





Anna de Sousa & Paul Young

THE NATIONAL ARCHIVES

Photo by Drew Beamer on Unsplash

Schedule

- What is metadata, questions to ask before you act, examples of metadata captured for digital preservation and access.
- **Break**
- Validating metadata. Generating DROID metadata, running the CSV validator. Editing metadata and schemas.
- **Lunch**
- Metadata tools (looking at JHOVE, FITS, Apache Tika)
- Guest speaker (Mark Bell – Automated descriptions)
- **Break**
- Continuation of metadata tools along with Homework
- **Finish**

If you want to be the best, and you want to beat the rest, metadata's what you need.

Metadata is commonly described as being data about data*.

*not that Data

Metadata – think before you act

- Metadata can cover technical, administrative, structural, preservation and descriptive elements and should assist in the preservation of your digital files, and in the provision of access to them. Information such as file name, checksum, filepath, date information, copyright status, closure information, hardware used to create the digital record and software used to render it are all types of metadata.
- There should be a sense of proportionality regarding metadata capture and generation - why are you generating the metadata? What function does it or could it serve?
- Metadata should enable you to both ask and answer questions:
    - how to produce the records?
    - how to mitigate preservation risk?
    - how to reliably create copies over time?
    - how to make decisions to provide access?
    - how to describe the records?
    - how to relate the records to other records or events?

THE NATIONAL ARCHIVES

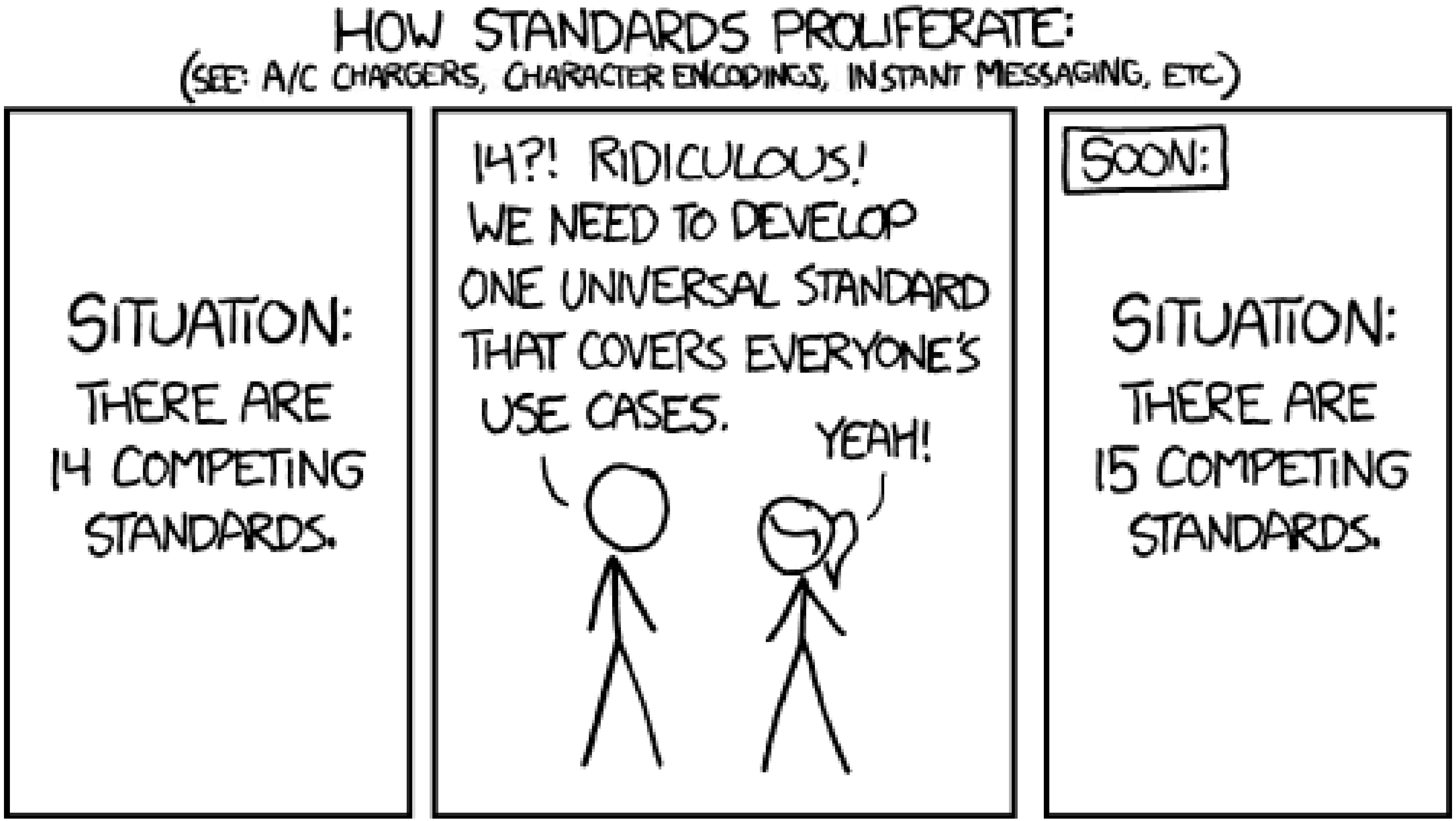
Metadata – think before you act

- When determining what metadata to extract and generate it is a valuable exercise to first determine:
  - The function of the metadata – what core information do you require to manage your digital records, preserve them and make them accessible?
  - Do you want to adhere to an existing metadata standard or do you need to extend existing standards to meet your needs?
  - What format do you want to capture your metadata in e.g. csv, xml?
  - How will you store your metadata – how will you maintain the relationship between the metadata and the digital records they relate to?
  - How will you utilise your metadata to assist access?
  - How will you validate your metadata?

Standards for digital objects

PREMIS: ‘The PREMIS Data Dictionary and its supporting documentation is a comprehensive, practical resource for implementing preservation metadata in digital archiving systems.’ : http://www.loc.gov/standards/premis/v3/

METS: ‘The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium’: http://www.loc.gov/standards/mets/


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

https://xkcd.com/927/

Photo by Ludemeula Fernandes on Unsplash

Metadata – automated extraction

- In the realm of digital records, you can extract metadata using free tools.
- Although you will need to point the tool to the files and start the process, this approach falls under automated metadata extraction.
- If costing metadata generation, this approach would be free except for staff time to run the software (which would be minutes of dedicated time for each instance).
- Tools such as DROID and Apache Tika enable automated metadata extraction.
- Just running DROID means you'd already have metadata that allows you to both make digital preservation decisions and describe your digital records, for example:
    - File name
    - Checksum
    - Filepath
    - File type
    - Date last modified

Metadata – manual generation

- Some metadata cannot be automatically generated. Often the record creator or depositor is best placed to provide this additional metadata due to their knowledge of the records. In some cases valuable metadata can be created by archivists or other third parties.
- As with all metadata, prior to proceeding you should determine what metadata you need to capture and why.

Costing the generation of metadata in this case will depend on factors such as:

- Are you using staff, volunteers, or a third party company to generate the metadata?
- How many records require metadata generation?
- How many individual fields of metadata are you asking for?
- Are you asking for quality assurance to be carried out in addition to generation of the metadata?

Digitisation is not Digital Preservation

- Often in the Digital Preservation realm you will hear the phrase ‘digitisation is not digital preservation’

- As we have learnt so far, simply storing digital records is not the same as actively carrying out digital preservation but it is remiss to ignore that the outputs of digitisation are digital records that also require digital preservation.

- As with born digital records, digitisation projects require metadata to be generated and captured in order to make the records accessible and enable them to be preserved over time.

- Simply digitising a record will not make it automatically findable or metadata rich in relation to the contents of the original record. Often digitisation requires more manual metadata generation than for born digital records, as the automatically generated metadata will tell you only about the digital image you have created and not about the original record.

Metadata for born digital records at TNA

Metadata is captured in csv format. The minimum metadata requirements for a born digital records transfer to TNA are:

1. Identifier (URI generated by DROID)
2. File_name (NAME generated by DROID)
3. Date (LAST MODIFIED generated by DROID)
4. Folder or file (TYPE generated by DROID)
5. Checksum (SHA256_HASH generated by DROID)
6. Copyright – usually ‘Crown Copyright’
7. Legal Status – usually ‘Public Record’
8. Held By – always ‘The National Archives, Kew’

We will gladly accept additional metadata fields that depositors want to provide, some previous examples include:

- Description – this is always gratefully received and is often beneficial as file names can be very non descriptive e.g. A1021023.docx
- Former reference
- Department

Metadata for born digital records at TNA

We also have mandatory closure fields to enable us to appropriately manage access to records. These are completed manually by the depositor and are based on the outcome of Advisory Council schedule applications:

1. Closure start date (date last modified of the record)
2. Closure period
3. Closure type
4. Foi exemption code(s)
5. Foi exemption asserted (date of Advisory Council meeting)
6. Description public
7. Title public
8. Description alternate
9. Title alternate

In addition we require that a checksum file is generated for the metadata csv – this ensures the metadata file has not been amended intentionally or via corruption or faulty copying. This file contains the name of the metadata csv two whitespaces and the checksum for the metadata csv.

Metadata for digitised records at TNA

- For digitised and surrogate records we require metadata around the creation of the images that provide both provenance and technical information.
- The technical acquisition metadata covers information such as the image resolution, format, filename, date of creation, colourspace information, checksum of the image and actions taken on the image e.g. deskew and crop.
- The technical environment metadata covers the hardware and software used to create the images and any software used to take action on the images e.g. deskew, crop, split.
- The transcription metadata includes all descriptive metadata about the content of the image which varies dependent on the project. For example a poor law project includes folio numbers and summary of the content of letters, a seal mould project includes information on the colour of the original seal and the dimensions of the original seal. These metadata fields are determined by the project owner in conjunction with the cataloguing team. The ensures consistency with existing series available in the catalogue.
- The closure csv provides us with the same fields as the born digital closure metadata and allows us to manage access.

It’s good to talk

- On your table, talk about the types of metadata you currently capture, or think you need to capture in the future if not already doing so.
- What concerns you most about metadata capture and storage?
- How do/will you store your metadata in relation to the digital records they relate to?
- How do/will you validate your metadata?

THE NATIONAL ARCHIVES

Photo by Dan Gold on Unsplash

Validating metadata

- At TNA all metadata we receive is in csv format. In order to ensure the quality of the metadata we receive, we validate it all before proceeding with our pre-ingest and ingest processes.

- We utilise a free tool developed at TNA called ‘CSV Validator’ which is open source and available to all.

- CSV validator allows the user to point to a metadata csv and a schema with rules for the metadata, and it reports if the metadata adheres to the rules (PASS) or does not (FAIL with errors).

- The schemas used by the csv validator adhere to the CSV schema language: https://digital-preservation.github.io/csv-schema/csv-schema-1.2.html

CSV Schema & schema language

- A CSV Schema is a rules based language which defines how data in each cell of a csv should be formatted.
- A schema rule is written in order for each column of the CSV file. Each set of column rules are asserted against each row of the CSV file. If validation of a column is not desirable, then an empty or optional rule is used.
- Always begin your schema with the version number of the schema language you're using (current version is 1.2) and the number of columns in the metadata csv you wish to validate e.g.

```
version 1.2
@totalColumns 23
```

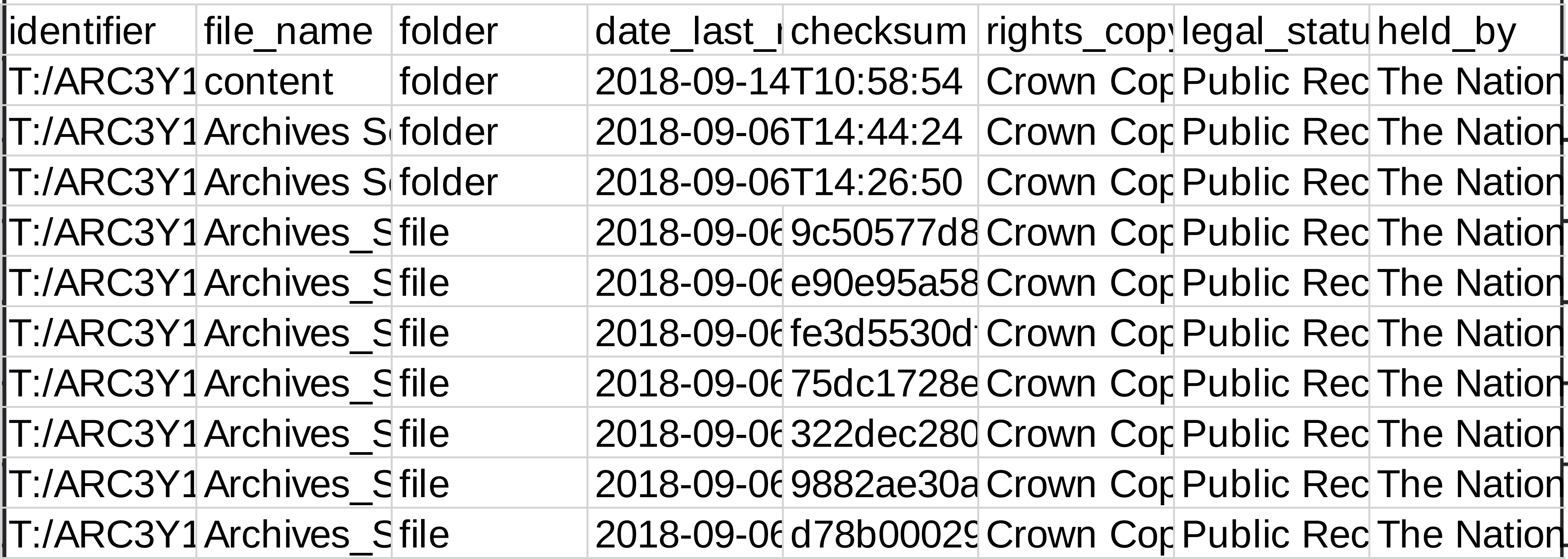
- A column rule may express constraints based on the content of other columns in the same row. It is possible to check that a cell entry is unique within that column in the CSV file (or that the value of a combination of cells is unique)

CSV Schema & schema language

The CSV schema can be used by the CSV Validator to:

- analyse the contents of the metadata CSV file you have completed to ensure it is consistent and accurate e.g. check that data falls within in expected numeric or character ranges.
- scan the original files to check their integrity
- check that all files mentioned in the metadata are included in the folder structure in your preparation area/on your hard drive
- check if there are any additional files in your preparation area/on your hard drive which are not represented in the metadata csv file

Snapshot of a metadata csv



identifier	file_name	folder	date_last_r	checksum	rights_copy	legal_statu	held_by
T:/ARC3Y1	content	folder	2018-09-14T10:58:54		Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives S	folder	2018-09-06T14:44:24		Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives S	folder	2018-09-06T14:26:50		Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	9c50577d8	Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	e90e95a58	Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	fe3d5530d	Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	75dc1728e	Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	322dec280	Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	9882ae30a	Crown Cop	Public Rec	The Nation
T:/ARC3Y1	Archives_S	file	2018-09-06	d78b00029	Crown Cop	Public Rec	The Nation

Schema for metadata on previous slide

```
version 1.2
@totalColumns 8
identifier: uri fileExists unique if($folder/is("folder"),ends("/")) integrityCheck ("includeFolder")
file_name: length(1,*)
folder: is("folder") or is("file")
date_last_modified: xDateTime
checksum: if($folder/is("file"),checksum(file($identifier),"SHA-256"),is(""))
rights_copyright: is(“Crown Copyright”)
legal_status: is("Public Record(s)")
held_by: is("The National Archives, Kew")
```

CSV schema language – some popular expressions

- is(“insert text or value here") e.g. is(“21) or is(“Crown Copyright”)

is allows you to specify exact text or number that should appear in that field - if it differs at all from what you specify within quotation marks, it will fail validation.

- range(lowest number,highest number) e.g. range(1,299)

range allows you to specify the lowest and highest value expected, anything within that specified range would pass validation.

- if($columnheadername/empty,empty,is(“text or value") e.g. if($ordinal/empty,empty,is("6")

if allows you to specify dependencies on the content of other columns in the same row. $ need to be placed before the column header name. empty allows for no metadata.

- any(“insert text or value"," insert text or value “ e.g. any(TRUE","FALSE“)

any allows you to specify any text or values that should be expected in the column. They must be within quotation marks and any variations must be comma separated. If you have long lists you should look to use regex instead of any.

- regex(“insert regex here") e.g. regex("[0-9]{1,4}")

You can use regex within a csv schema, you must declare it by using the word regex followed by round brackets and all your regex held within quotation marks.

- length(1,*) e.g. length(1,*)

length checks that the number of characters in the column meets the supplied definition.

Photo by JC Gellidon on Unsplash

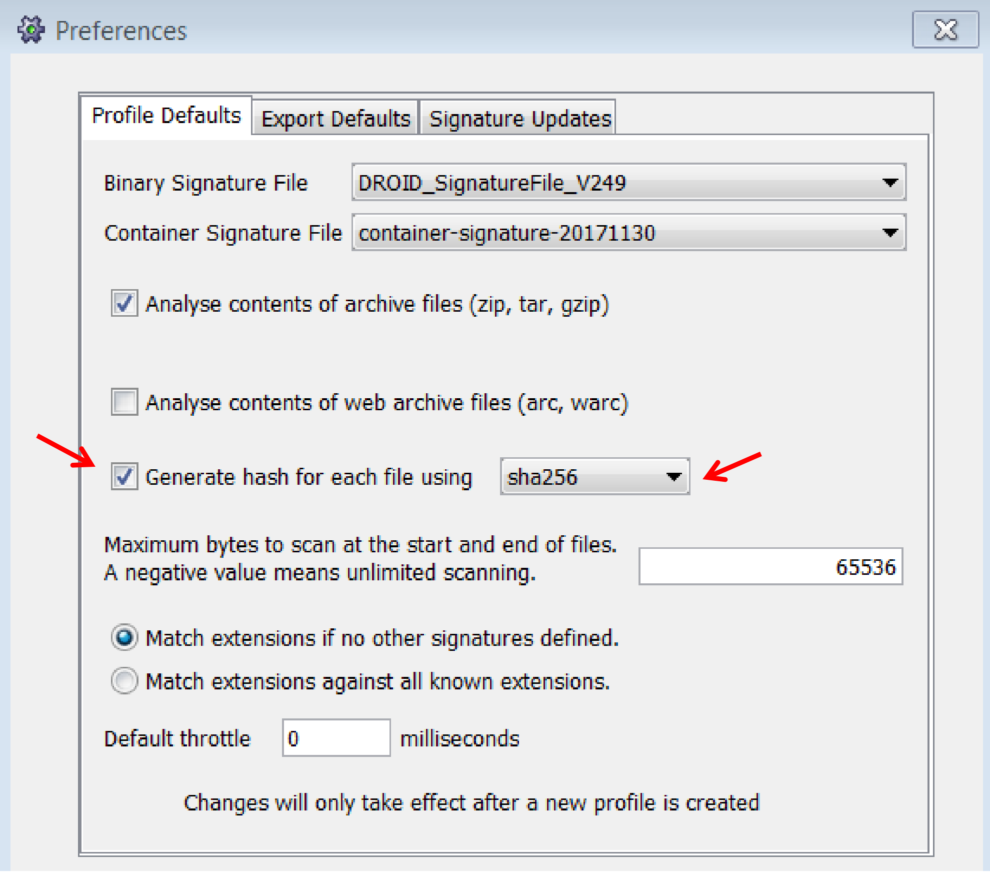
Practical Activity

Screen shots and exact instructions on following slides.

1. Run DROID to generate metadata
2. Run CSV validator pointing at metadata csv and schema
3. Edit metadata csv and schema
4. Run csv validator again with updated csv and schema

Run DROID

- Go to your desktop and click on the file named droid.bat. This will launch DROID
- Select ‘preferences’ from the ‘Tools’ menu and ensure the box next to ‘Generate hash for each file using’ is ticked, and the drop down box next to it shows ‘sha256’. If instead of sha256 it is giving the option to create an md5 checksum, click on the drop down box to select sha256. Click OK.



Preferences

Binary Signature File: DROID_SignatureFile_V249

Container Signature File: container-signature-20171130

☑ Analyse contents of archive files (zip, tar, gzip)

□ Analyse contents of web archive files (arc, warc)

☑ Generate hash for each file using: sha256

Maximum bytes to scan at the start and end of files.
A negative value means unlimited scanning. 65536

● Match extensions if no other signatures defined.

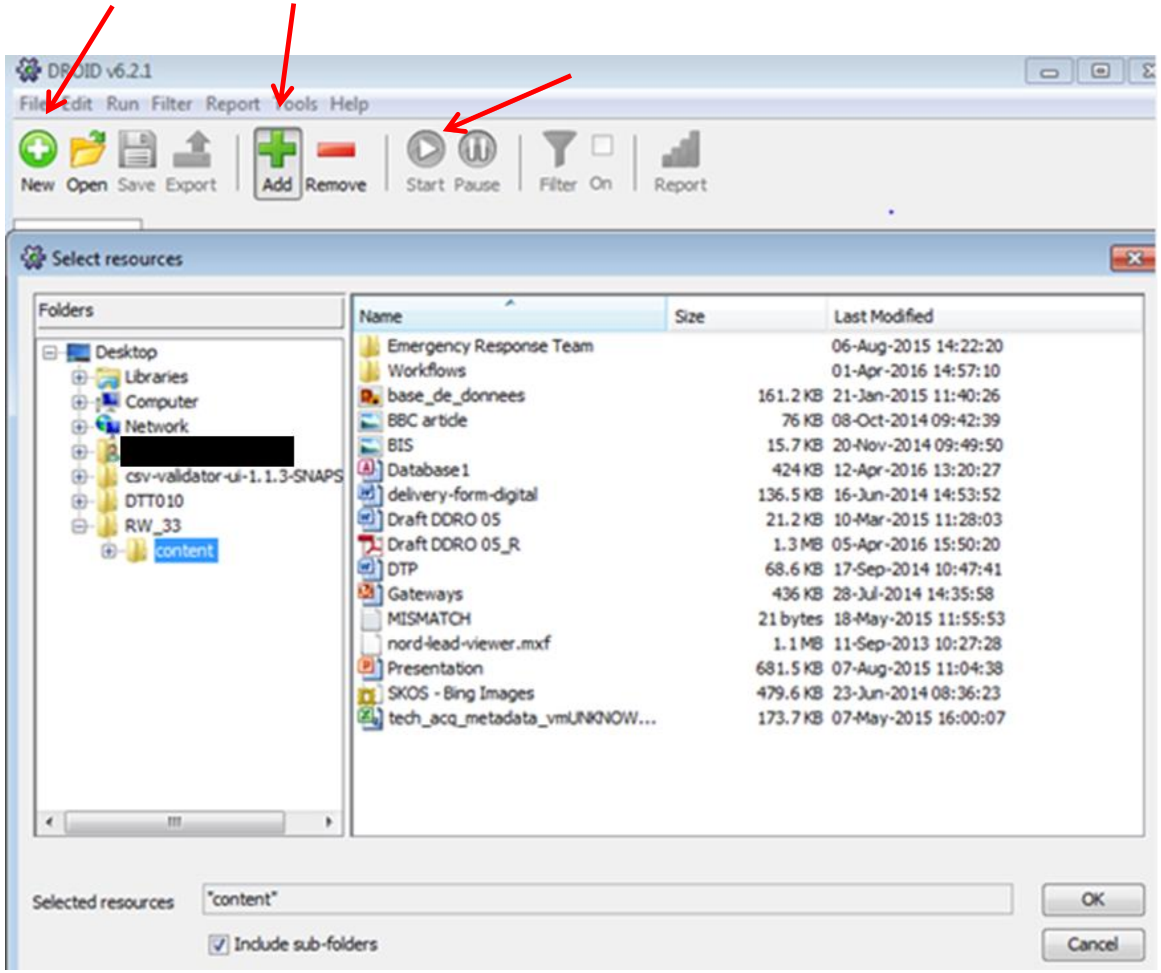
○ Match extensions against all known extensions.

Default throttle: 0 milliseconds

Changes will only take effect after a new profile is created

Run DROID

- Click on the New icon (Circle with a plus sign in its centre) to create a new profile, this ensures that that sha256 setting will take effect.
- Select the green ‘Add’ icon on the main screen. This will open Windows Explorer. At this point, navigate to your Documents folder then to: ArchivesSchoolSession3\1stexercise\content
- Click on the content folder. Once you’ve selected the folder, it will then appear on the main DROID screen. Click OK.
- Press the ‘Start’ icon to run DROID (this will turn blue after you click OK).



Name	Size	Last Modified
Emergency Response Team		06-Aug-2015 14:22:20
Workflows		01-Apr-2016 14:57:10
base_de_donnees	161.2 KB	21-Jan-2015 11:40:26
BBC article	76 KB	08-Oct-2014 09:42:39
BIS	15.7 KB	20-Nov-2014 09:49:50
Database1	424 KB	12-Apr-2016 13:20:27
delivery-form-digital	136.5 KB	16-Jun-2014 14:53:52
Draft DDRO 05	21.2 KB	10-Mar-2015 11:28:03
Draft DDRO 05_R	1.3 MB	05-Apr-2016 15:50:20
DTP	68.6 KB	17-Sep-2014 10:47:41
Gateways	436 KB	28-Jul-2014 14:35:58
MISMATCH	21 bytes	18-May-2015 11:55:53
nord-lead-viewer.mxf	1.1 MB	11-Sep-2013 10:27:28
Presentation	681.5 KB	07-Aug-2015 11:04:38
SKOS - Bing Images	479.6 KB	23-Jun-2014 08:36:23
tech_acq_metadata_vmUNKNOW...	173.7 KB	07-May-2015 16:00:07

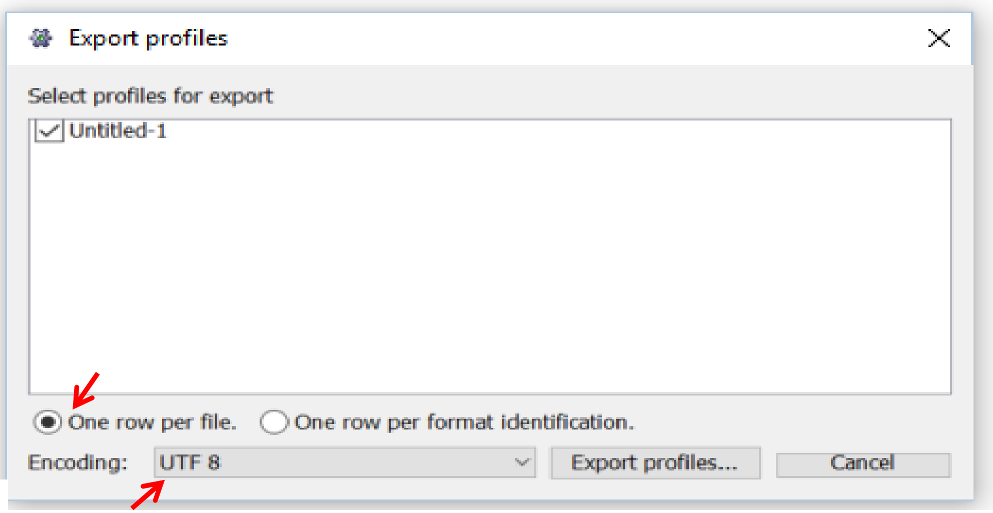
Selected resources "content"

☑ Include sub-folders

THE
NATIONAL
ARCHIVES

Export DROID report

- Once DROID has finished running, you will be able to export the results as a CSV (comma separated values) file. To do this, first select the ‘Export’ icon and tick the box labelled ‘untitled 1’ (or whichever profile you wish to export). Ensure the encoding at the bottom is set to UTF 8. Then, click ‘Export profiles’.

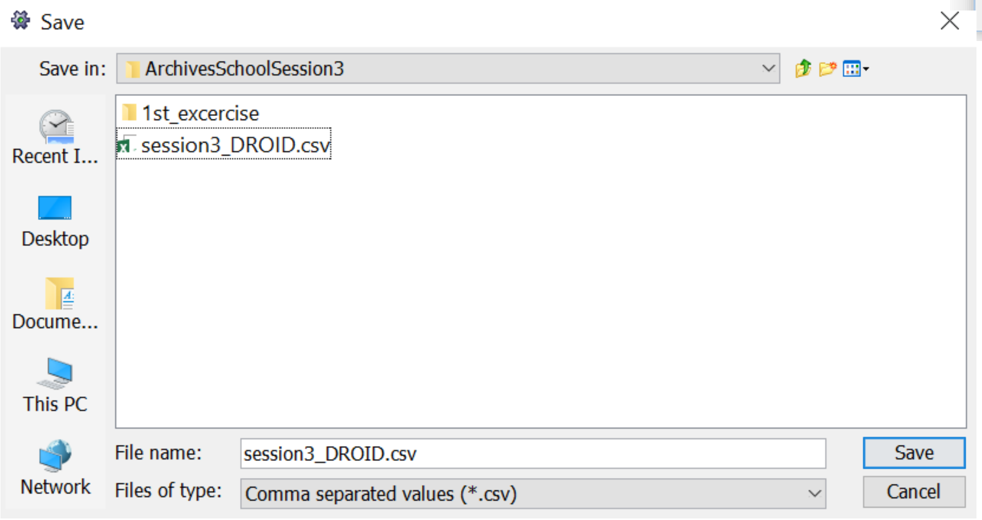


Select profiles for export

☑ Untitled-1

● One row per file. ○ One row per format identification.

Encoding: UTF 8



Save in: ArchivesSchoolSession3

1st_excercise

session3_DROID.csv

File name: session3_DROID.csv

Files of type: Comma separated values (*.csv)

- You will then be given the option to save the report. Select csv from the ‘files of type’ drop down underneath the filename text box. Save alongside the 1stexcercise folder

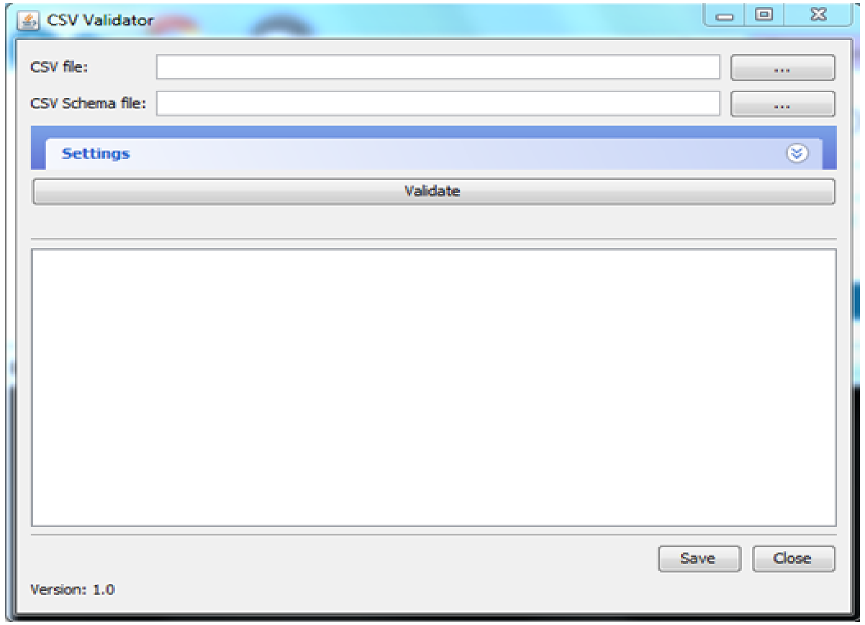
THE
NATIONAL
ARCHIVES

Schema

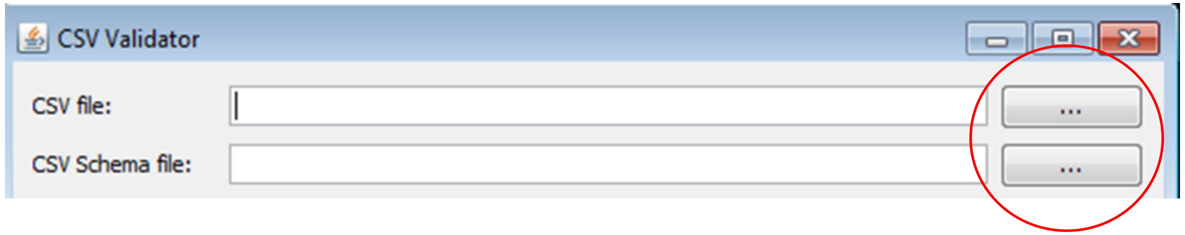
- Go to ArchivesSchoolSession3 in your Documents
- Open ArchivesShool_schema.csvs in Notepad++
- Let’s talk through what you see!

Run the CSV Validator

In order to run the CSV validator, go to your desktop and double click validate-gui.bat - Shortcut.
A command prompt will open and then this screen should appear

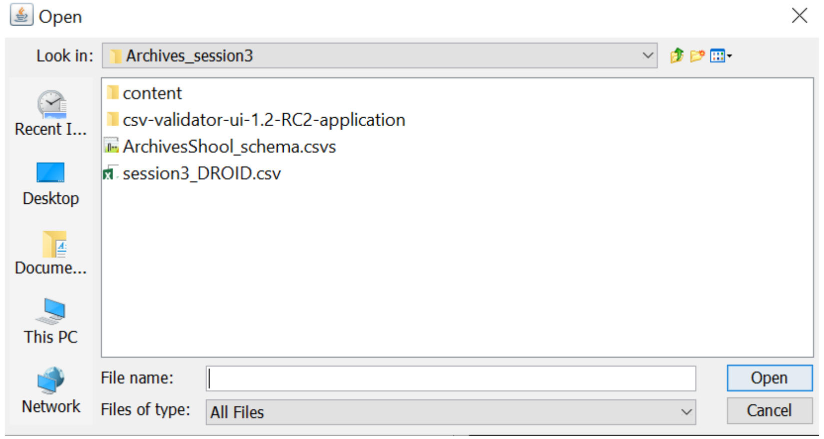


Click the '…' buttons to navigate to the location of your metadata.csv file and ArchivesShool_schema.csvs in the ArchiveSchoolSession3 folder in your documents.



Run the CSV Validator

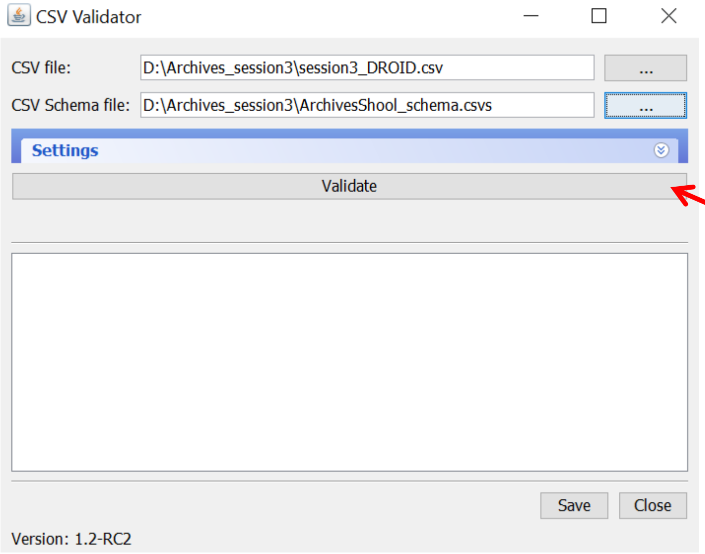
This will open up a standard ‘File open’ dialogue, allowing you to navigate to and select the relevant file in the file system.



Look in: Archives_session3

content
csv-validator-ui-1.2-RC2-application
ArchivesShool_schema.csvs
session3_DROID.csv

Files of type: All Files



CSV file: D:\Archives_session3\session3_DROID.csv

CSV Schema file: D:\Archives_session3\ArchivesShool_schema.csvs

Version: 1.2-RC2

Once you have selected your csv and Schema then click on the Validate button.
Once the CSV validator has run it will either show the word PASS in the white text box or it will say FAIL and report errors.

Photo by freestocks.org on Unsplash

Edit the metadata & schema

- Open up the DROID report csv you've been running the CSV Validator against.
- The final column heading you see should be 'FORMAT_VERSION'
- In the next column along, in the same row as the column headings, type: held_by
- In the row below the column heading type Archive School. Copy that down so you now have Archive School in that column for every row that contains data.
- Open the schema you've been running the CSV Validator against. At the top of the schema change @totalColumns 18 to @totalColumns 19 (as you've just added an additional column!)
- At the end of the schema you'll see the final row is currently FORMAT_VERSION. Hit enter to create a new row below this.
- In this new row type held_by: is ("Archive School")
- Save!

THE

NATIONAL

ARCHIVES

Run CSV validator again

You are now running the csv validator again exactly the same as before you just have additional metadata and an additional schema rule to check your new metadata!

Photo by Lee Cartledge on Unsplash

Tools

Photo by Cesar Carlevarino Aragon on Unsplash

Tools!

DPC Handbook:

- Sustainability of tools and community participation
- Finding digital preservation tools: tools registries
- Open source versus commercial software
- Enterprise-level solutions versus micro-services
- Where will it sit in your workflow

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

https://dpconline.org/handbook/technical-solutions-and-tools

Tool sustainability

- Is the tool regularly updated?
- How large is the user base?
- Who supports the tool?
    - TNA supports DROID and CSV Validator
    - Open Preservation Foundation (OPF has taken up support for several tools) https://openpreservation.org/technology/products/

Open Preservation Foundation

JHOVE

jpylyzer

fido

veraPDF

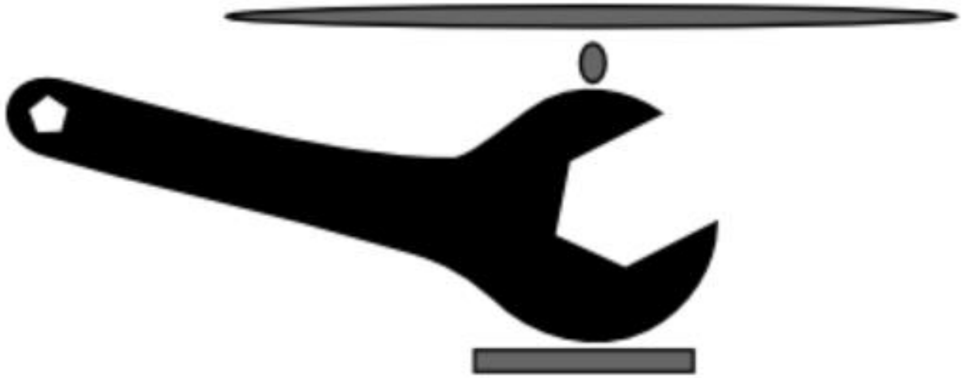
Digital Preservation Tool Registries

Community Owned digital Preservation Tool Registry (COPTR)

This register collates information from several independent registers e.g.

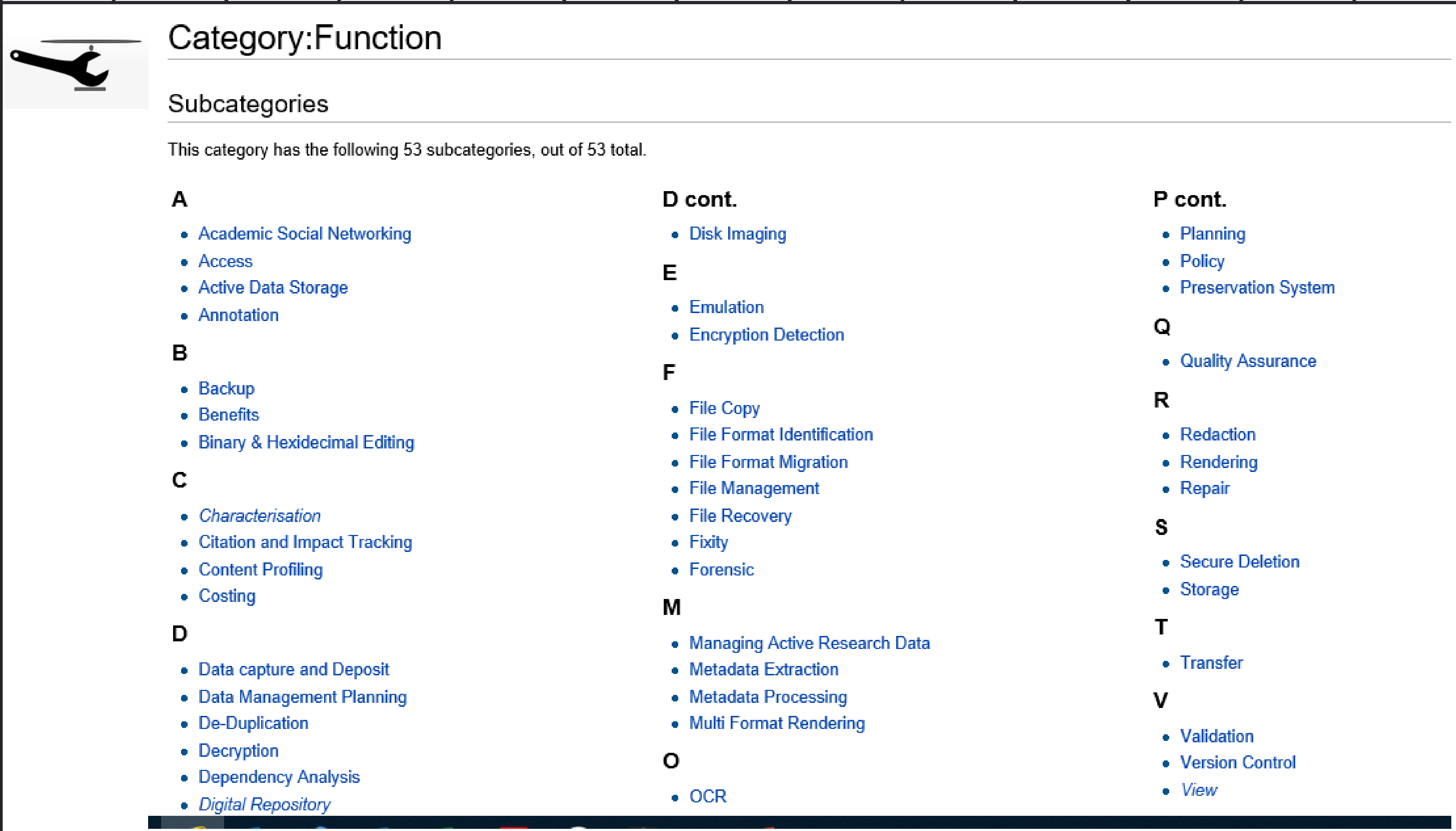
- Digital Curation Centre
- Library of Congress

462 tools currently, open for anyone to add information about tools they know about



https://coptr.digipres.org/

THE NATIONAL ARCHIVES



Category:Function

Subcategories

This category has the following 53 subcategories, out of 53 total.

A

- Academic Social Networking
- Access
- Active Data Storage
- Annotation

B

- Backup
- Benefits
- Binary & Hexidecimal Editing

C

- *Characterisation*
- Citation and Impact Tracking
- Content Profiling
- Costing

D

- Data capture and Deposit
- Data Management Planning
- De-Duplication
- Decryption
- Dependency Analysis
- *Digital Repository*

D cont.

- Disk Imaging

E

- Emulation
- Encryption Detection

F

- File Copy
- File Format Identification
- File Format Migration
- File Management
- File Recovery
- Fixity
- Forensic

M

- Managing Active Research Data
- Metadata Extraction
- Metadata Processing
- Multi Format Rendering

O

- OCR

P cont.

- Planning
- Policy
- Preservation System

Q

- Quality Assurance

R

- Redaction
- Rendering
- Repair

S

- Secure Deletion
- Storage

T

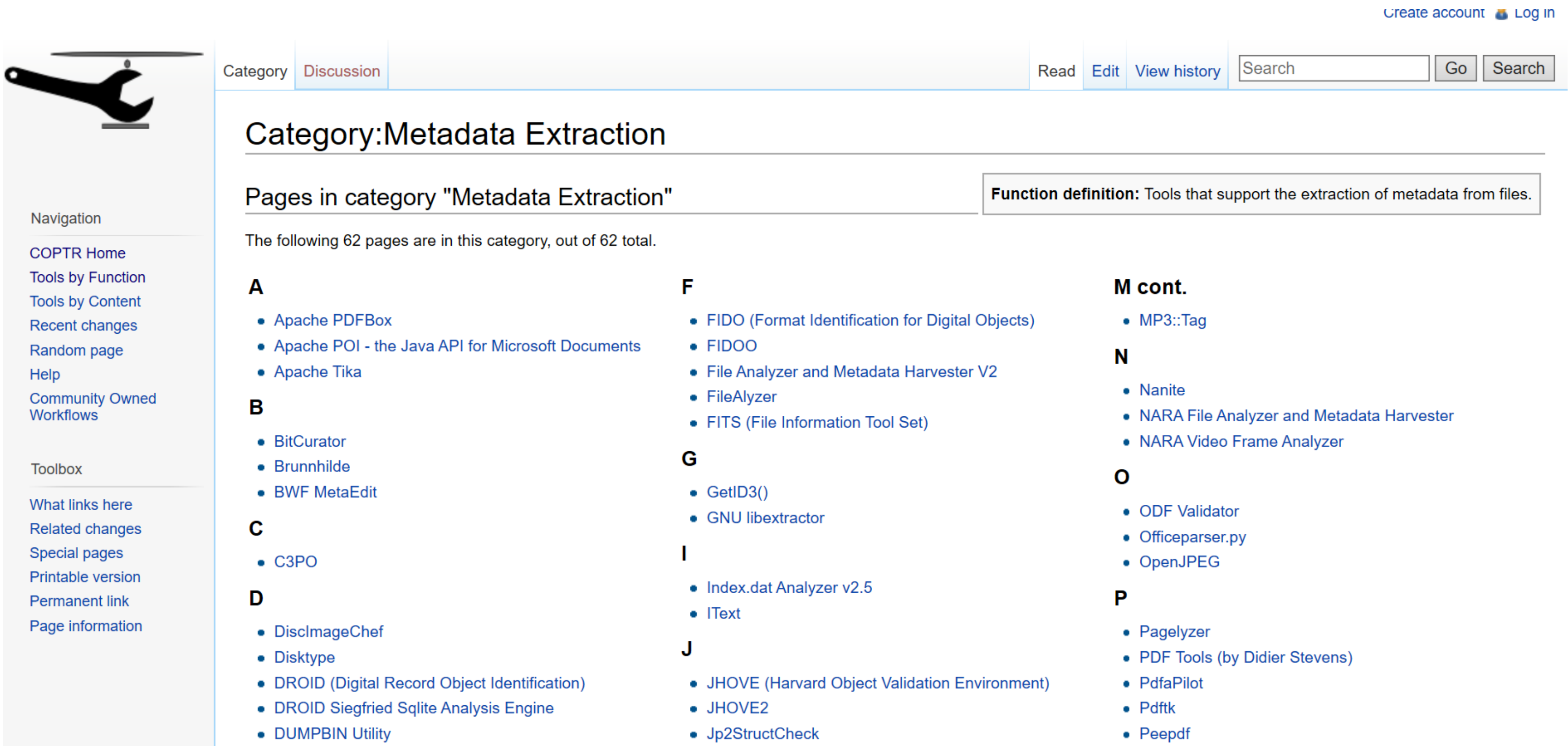
- Transfer

V

- Validation
- Version Control
- *View*

THE
NATIONAL
ARCHIVES

Metadata Extraction tools -



Navigation

- COPTR Home
- Tools by Function
- Tools by Content
- Recent changes
- Random page
- Help
- Community Owned Workflows

Toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Page information

Category:Metadata Extraction

Pages in category "Metadata Extraction"

Function definition: Tools that support the extraction of metadata from files.

The following 62 pages are in this category, out of 62 total.

A

- Apache PDFBox
- Apache POI - the Java API for Microsoft Documents
- Apache Tika

B

- BitCurator
- Brunnhilde
- BWF MetaEdit

C

- C3PO

D

- DiscImageChef
- Disktype
- DROID (Digital Record Object Identification)
- DROID Siegfried Sqlite Analysis Engine
- DUMPBIN Utility

F

- FIDO (Format Identification for Digital Objects)
- FIDOO
- File Analyzer and Metadata Harvester V2
- FileAlyzer
- FITS (File Information Tool Set)

G

- GetID3()
- GNU libextractor

I

- Index.dat Analyzer v2.5
- IText

J

- JHOVE (Harvard Object Validation Environment)
- JHOVE2
- Jp2StructCheck

M cont.

- MP3::Tag

N

- Nanite
- NARA File Analyzer and Metadata Harvester
- NARA Video Frame Analyzer

O

- ODF Validator
- Officeparser.py
- OpenJPEG

P

- Pagelyzer
- PDF Tools (by Didier Stevens)
- PdfaPilot
- Pdftk
- Peepdf

https://coptr.digipres.org/Category:Metadata_Extraction

THE NATIONAL ARCHIVES

JHOVE

FITS FITS FITS

FITS FITS FITS

Apache™ Tika

JHOVE

JSTOR (Journal Storage)/Harvard Object Validation Environment

- File format identification, validation and characterisation (representation information)
- Includes modules for:
  - AIFF (audio interchange format)
  - ASCII (text)
  - GIF (image)
  - GZIP (compressed)
  - HTML (web)
  - JPEG (image
  - JPEG 2000 (image)
  - PDF (document)
  - TIFF (image)
  - UTF-8 (text)
  - WARC (web archive)
  - WAVE (audio)
  - XML (data)
  - EPUB (book)
  - MP3 (audio)
  - ZIP (compressed)

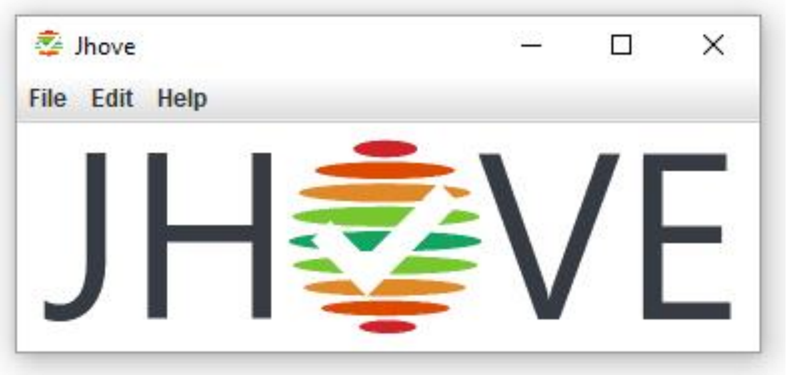
- **Bytestream (default)**

https://openpreservation.org/technology/products/jhove/

THE NATIONAL ARCHIVES

Exercise: JHOVE

- Open JHOVE GUI by going to ‘ArchivesSchoolSession3’ then ‘Software’ folder, open folder ‘jhove’ click on ‘jhove-gui.bat’
- Click ‘File’ select ‘open file’ and select a file from folder ‘TestFiles’ and use the JHOVE GUI to compare outputs



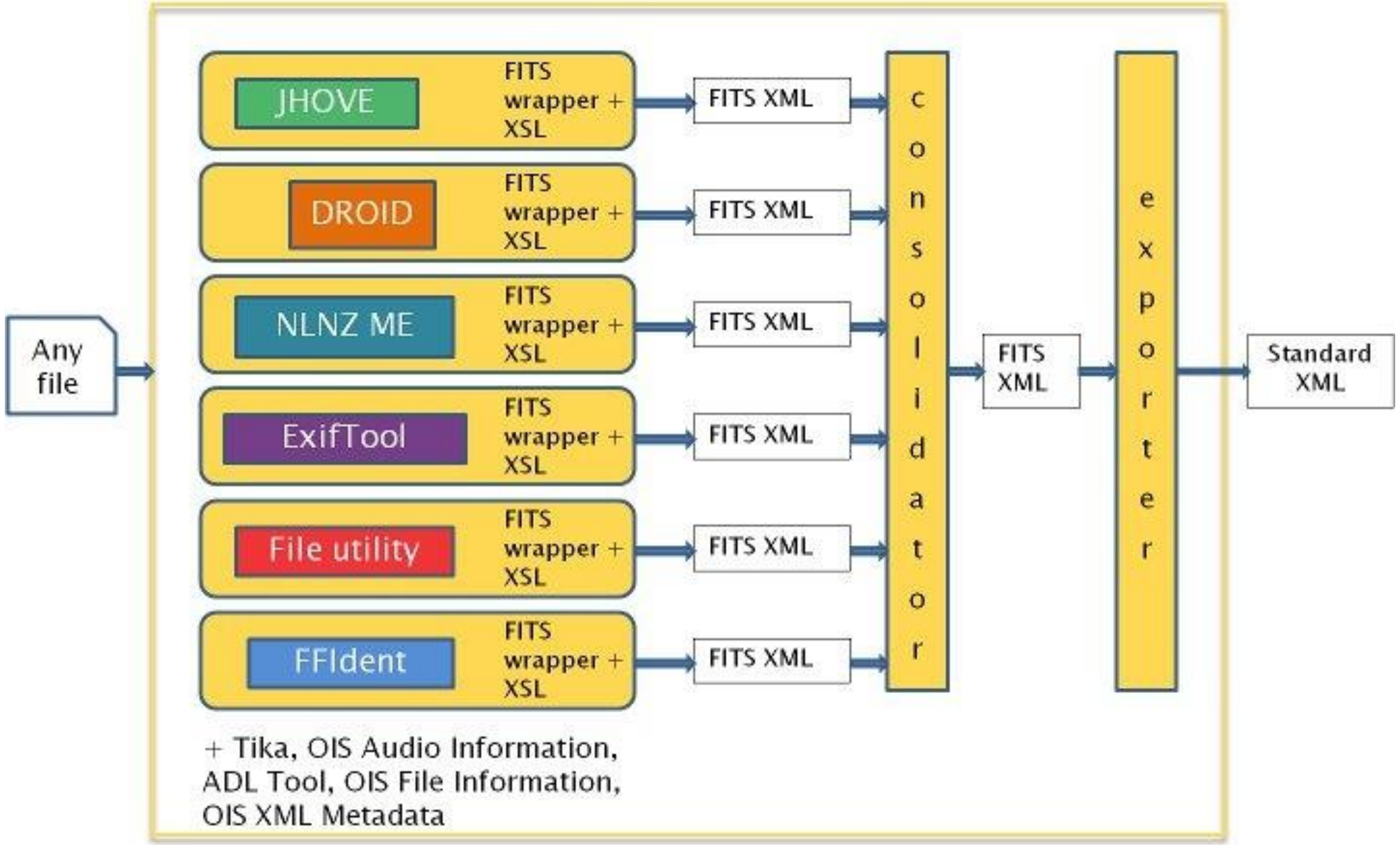
FITS (File Information Tool set)

https://projects.iq.harvard.edu/fits

The File Information Tool Set (FITS) identifies, validates and extracts technical metadata for a wide range of file formats.

It acts as a wrapper, invoking and managing the output from several other open source tools.

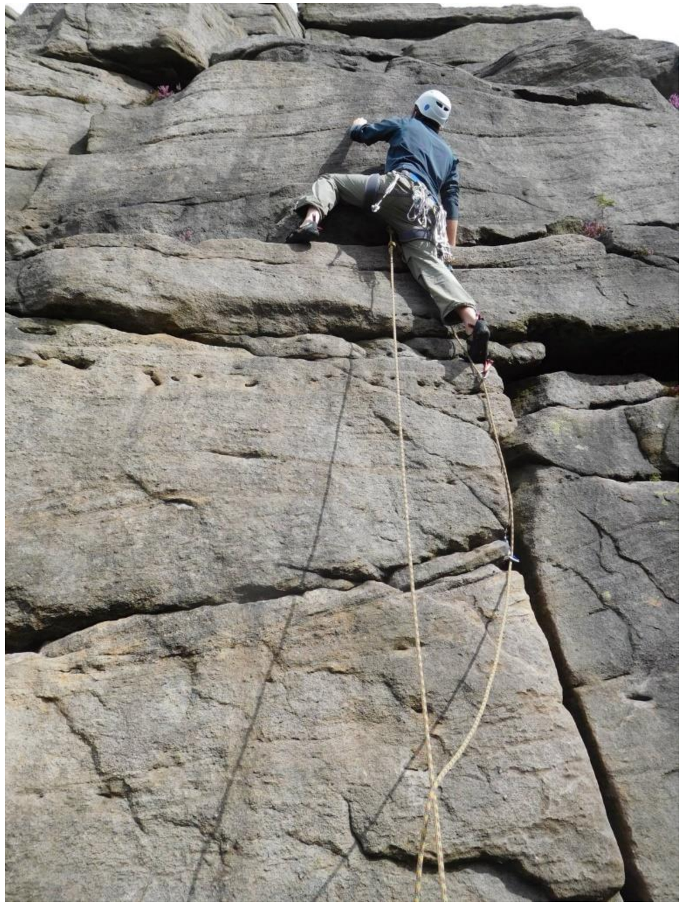
- ADL Tool
- Apache Tika
- DROID
- Exiftool
- FFIdent
- File Utility (windows port)
- Jhove
- MediaInfo
- National Library of New Zealand Metadata Extractor
- OIS Audio Information
- OIS File Information
- OIS XML Information



+ Tika, OIS Audio Information, ADL Tool, OIS File Information, OIS XML Metadata

Exercise: FITS

- Open a terminal window – type ‘cmd’ in windows search bar
- Navigate to folder which includes FITS – ‘cd Documents\Software\ArchiveSchoolSession3 \Software\fits-1.5.0’
- Run FITS over a single file in ‘TestFiles’ with following cmd – e.g. fits.bat –I "C:\Users\ASD1\Documents\ArchiveSchoolSession3\TestFiles\Draft DDRO 05.docx"
- Run FITS over directory ‘TestFiles’ with following cmd – fits.bat –i C:\Users\ASD1\Documents\ArchiveSchoolSession3\TestFiles –o “C:\Users\ASD1\Documents\ArchiveSchoolSession3\FitsOutput”



THE NATIONAL ARCHIVES

Apache™ Tika

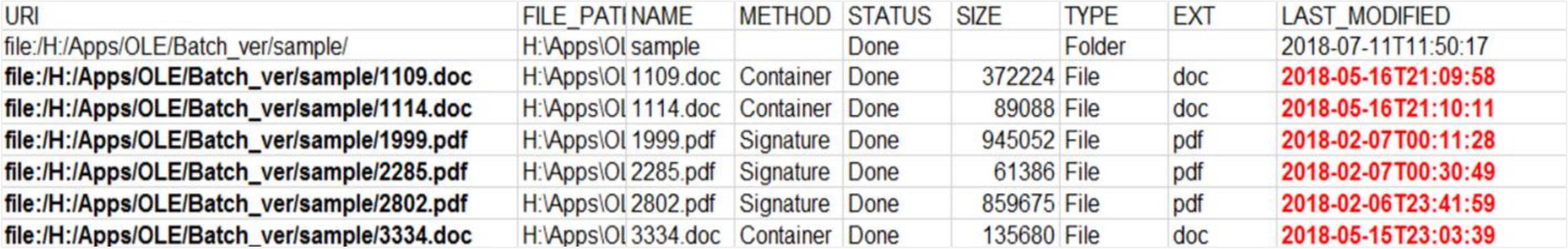
Detects and extracts metadata and text from over a thousand different file types.

The date problem

"The problem we are encountering relates to dates captured as Last Modified. I would welcome any advice on how dates can be different, how best to complete a more accurate analysis." - NRW

"missing a lot of dates so I think the only option is to open each document and see if there is some kind of date in them" - CoRWM

THE
NATIONAL
ARCHIVES

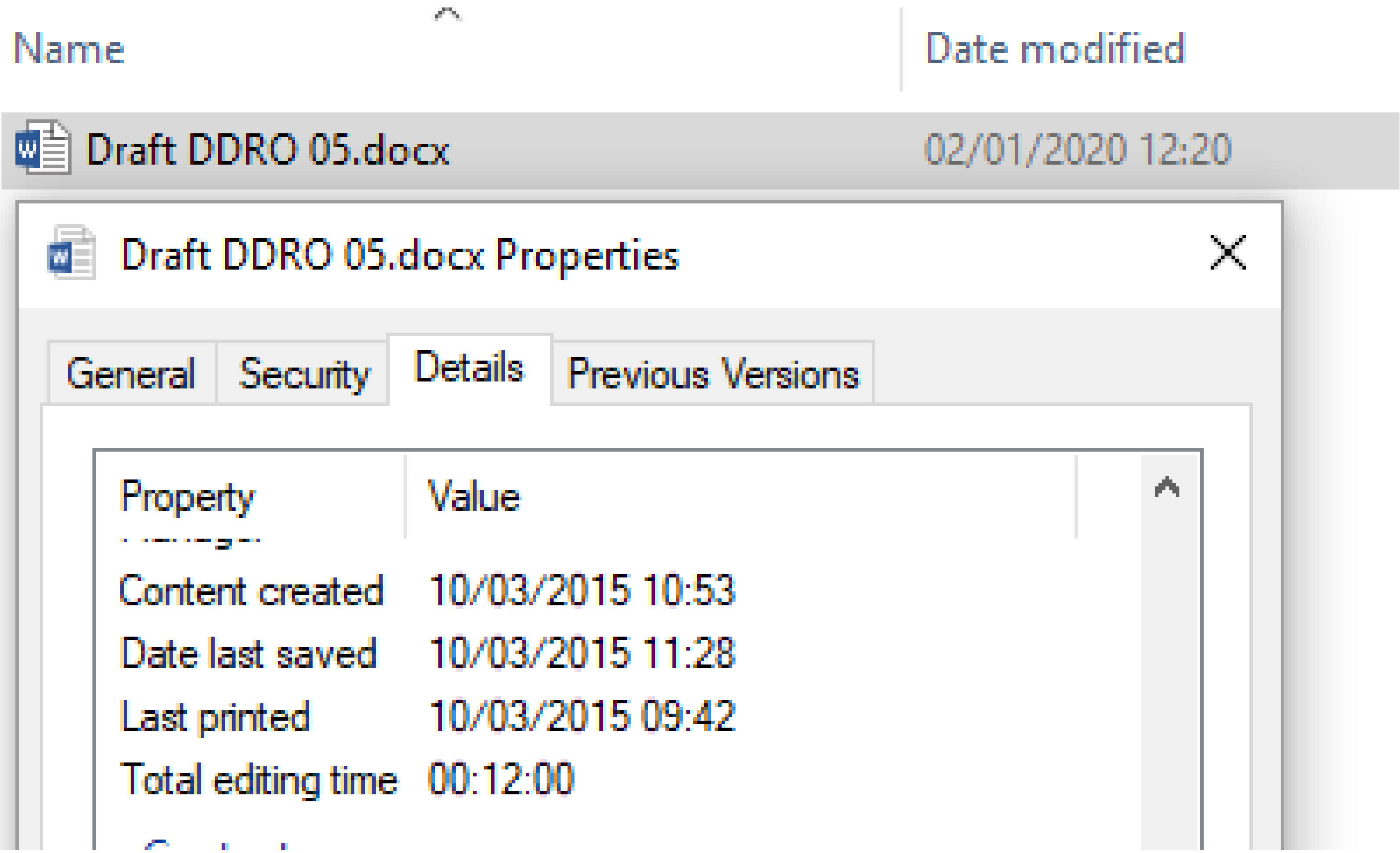


URI	FILE_PATI	NAME	METHOD	STATUS	SIZE	TYPE	EXT	LAST_MODIFIED
file:/H:/Apps/OLE/Batch_ver/sample/	H:\Apps\Ol	sample		Done		Folder		2018-07-11T11:50:17
file:/H:/Apps/OLE/Batch_ver/sample/1109.doc	H:\Apps\Ol	1109.doc	Container	Done	372224	File	doc	2018-05-16T21:09:58
file:/H:/Apps/OLE/Batch_ver/sample/1114.doc	H:\Apps\Ol	1114.doc	Container	Done	89088	File	doc	2018-05-16T21:10:11
file:/H:/Apps/OLE/Batch_ver/sample/1999.pdf	H:\Apps\Ol	1999.pdf	Signature	Done	945052	File	pdf	2018-02-07T00:11:28
file:/H:/Apps/OLE/Batch_ver/sample/2285.pdf	H:\Apps\Ol	2285.pdf	Signature	Done	61386	File	pdf	2018-02-07T00:30:49
file:/H:/Apps/OLE/Batch_ver/sample/2802.pdf	H:\Apps\Ol	2802.pdf	Signature	Done	859675	File	pdf	2018-02-06T23:41:59
file:/H:/Apps/OLE/Batch_ver/sample/3334.doc	H:\Apps\Ol	3334.doc	Container	Done	135680	File	doc	2018-05-15T23:03:39

File System Dates

- The ‘date last modified’ field was originally chosen because it was deemed the most reliable out of the file system dates. Creation dates can change when copying files (often to *after* the last modified date).
- After seeing several collections and scenarios where the ‘date last modified’ did not provide an accurate date for the file we investigated other methods for extracting accurate dates for born-digital records.
- Date last modified is not immune to change, especially when uploading or downloading to cloud storage or during migrations.

Windows Docx File

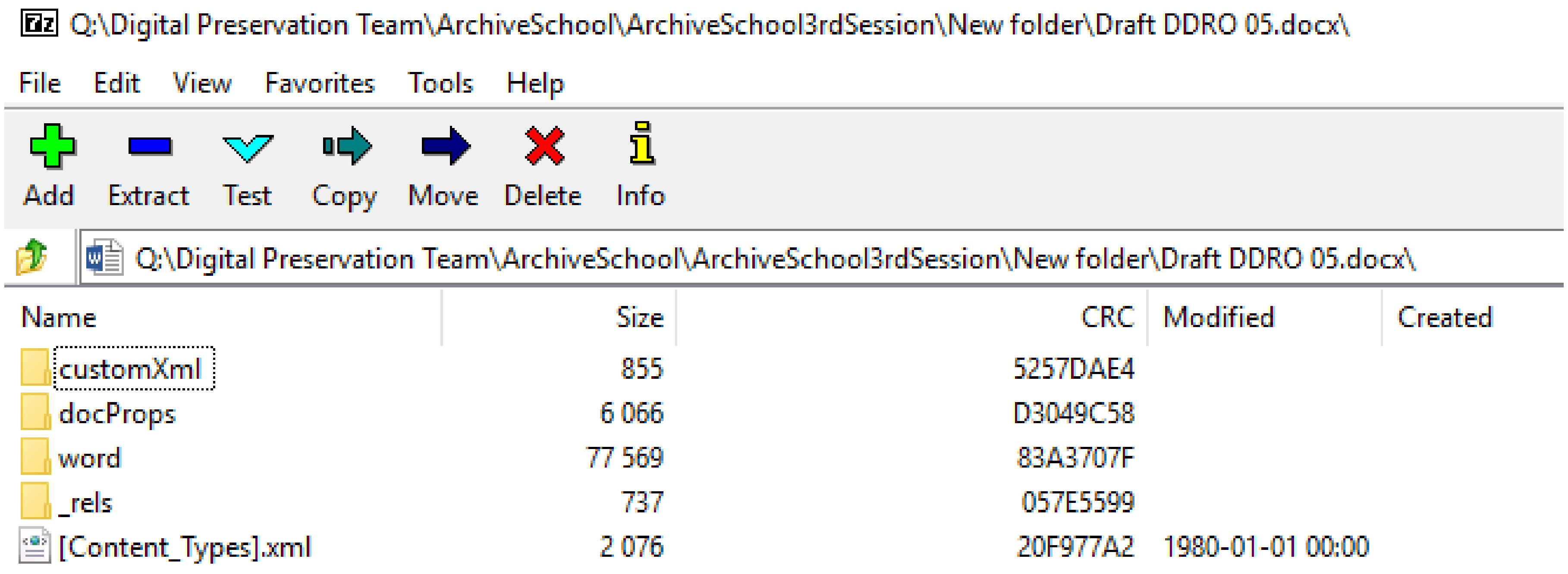


Name	Date modified
Draft DDRO 05.docx	02/01/2020 12:20

Draft DDRO 05.docx Properties

Property	Value
Content created	10/03/2015 10:53
Date last saved	10/03/2015 11:28
Last printed	10/03/2015 09:42
Total editing time	00:12:00

Windows Docx file

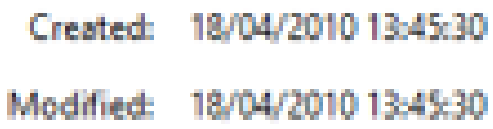


Q:\Digital Preservation Team\ArchiveSchool\ArchiveSchool3rdSession\New folder\Draft DDRO 05.docx\

Q:\Digital Preservation Team\ArchiveSchool\ArchiveSchool3rdSession\New folder\Draft DDRO 05.docx\

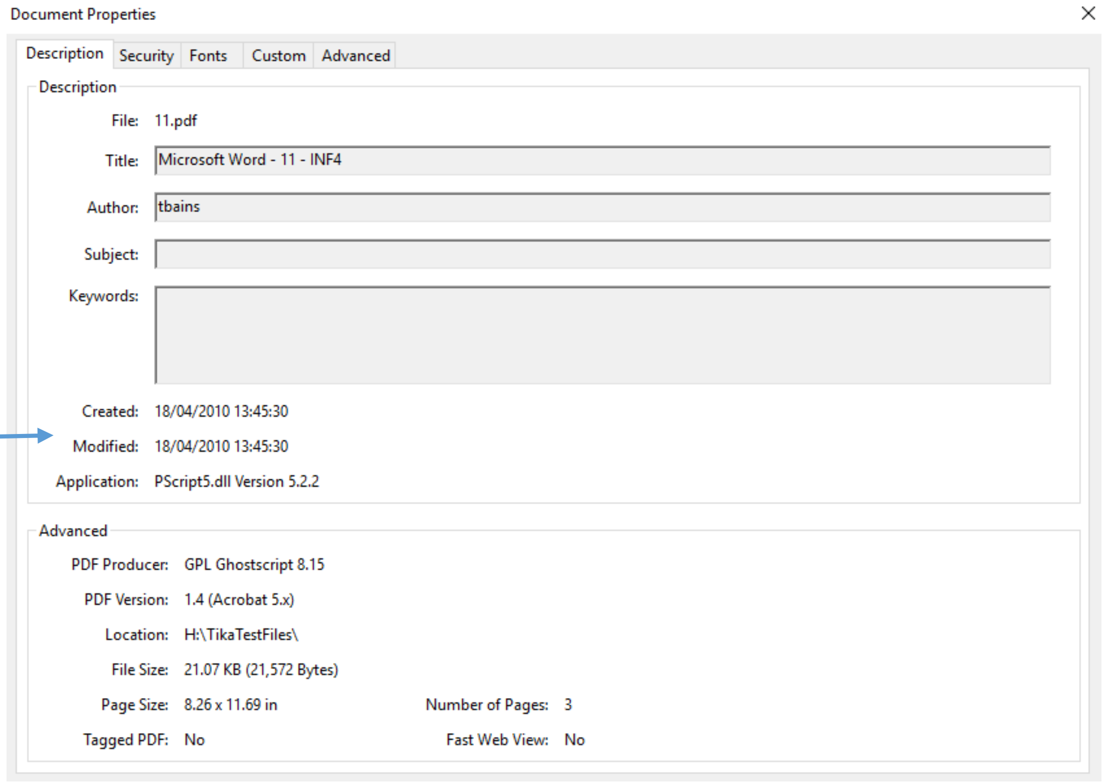
Name	Size	CRC	Modified	Created
customXml	855	5257DAE4		
docProps	6 066	D3049C58		
word	77 569	83A3707F		
_rels	737	057E5599		
[Content_Types].xml	2 076	20F977A2	1980-01-01 00:00	

Adobe PDF



Created: 18/04/2010 13:45:30

Modified: 18/04/2010 13:45:30



Document Properties

Description

File: 11.pdf

Title: Microsoft Word - 11 - INF4

Author: tbains

Subject:

Keywords:

Created: 18/04/2010 13:45:30

Modified: 18/04/2010 13:45:30

Application: PScript5.dll Version 5.2.2

Advanced

PDF Producer: GPL Ghostscript 8.15

PDF Version: 1.4 (Acrobat 5.x)

Location: H:\TikaTestFiles\

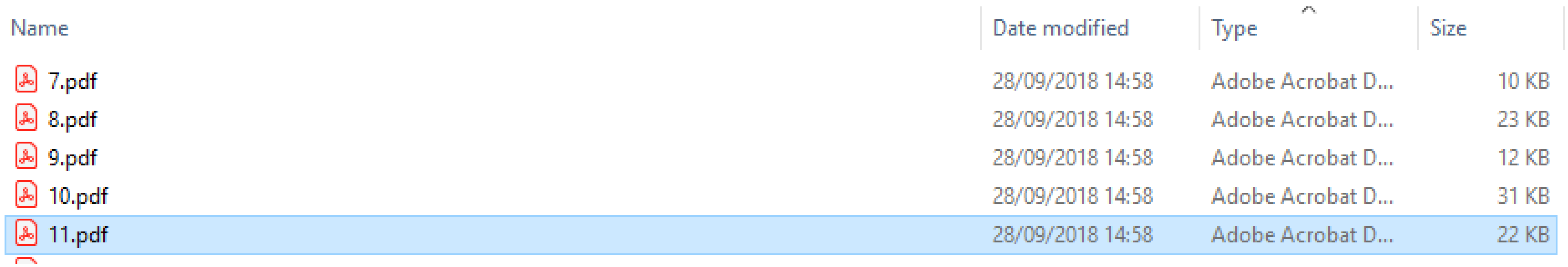
File Size: 21.07 KB (21,572 Bytes)

Page Size: 8.26 x 11.69 in

Number of Pages: 3

Tagged PDF: No

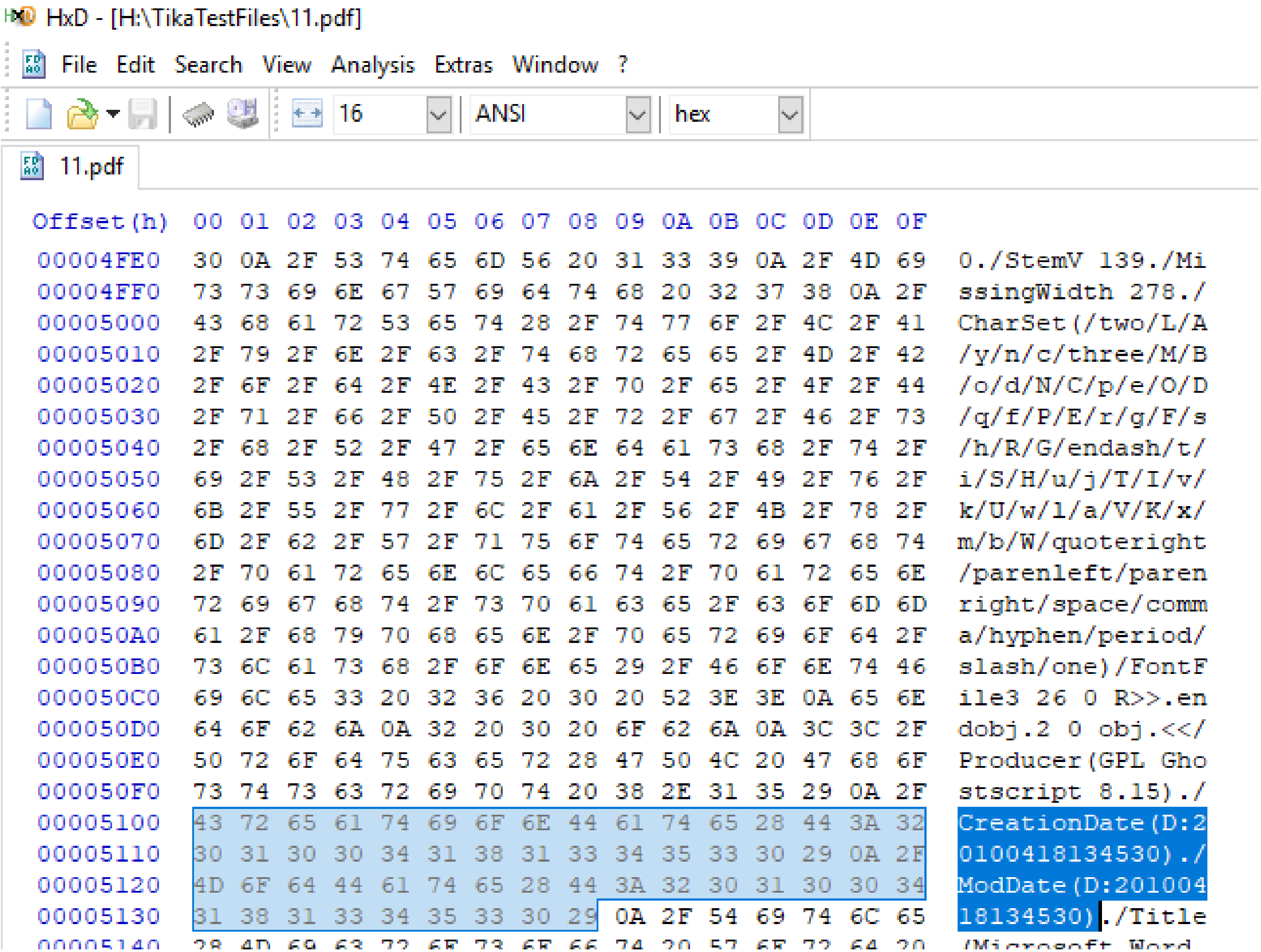
Fast Web View: No



Name	Date modified	Type	Size
7.pdf	28/09/2018 14:58	Adobe Acrobat D...	10 KB
8.pdf	28/09/2018 14:58	Adobe Acrobat D...	23 KB
9.pdf	28/09/2018 14:58	Adobe Acrobat D...	12 KB
10.pdf	28/09/2018 14:58	Adobe Acrobat D...	31 KB
11.pdf	28/09/2018 14:58	Adobe Acrobat D...	22 KB

THE NATIONAL ARCHIVES

Adobe PDF



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

00004FE0  30 0A 2F 53 74 65 6D 56 20 31 33 39 0A 2F 4D 69  0./StemV 139./Mi
00004FF0  73 73 69 6E 67 57 69 64 74 68 20 32 37 38 0A 2F  ssingWidth 278./
00005000  43 68 61 72 53 65 74 28 2F 74 77 6F 2F 4C 2F 41  CharSet(/two/L/A
00005010  2F 79 2F 6E 2F 63 2F 74 68 72 65 65 2F 4D 2F 42  /y/n/c/three/M/B
00005020  2F 6F 2F 64 2F 4E 2F 43 2F 70 2F 65 2F 4F 2F 44  /o/d/N/C/p/e/O/D
00005030  2F 71 2F 66 2F 50 2F 45 2F 72 2F 67 2F 46 2F 73  /q/f/P/E/r/g/F/s
00005040  2F 68 2F 52 2F 47 2F 65 6E 64 61 73 68 2F 74 2F  /h/R/G/endash/t/
00005050  69 2F 53 2F 48 2F 75 2F 6A 2F 54 2F 49 2F 76 2F  i/S/H/u/j/T/I/v/
00005060  6B 2F 55 2F 77 2F 6C 2F 61 2F 56 2F 4B 2F 78 2F  k/U/w/l/a/V/K/x/
00005070  6D 2F 62 2F 57 2F 71 75 6F 74 65 72 69 67 68 74  m/b/W/quoteright
00005080  2F 70 61 72 65 6E 6C 65 66 74 2F 70 61 72 65 6E  /parenleft/paren
00005090  72 69 67 68 74 2F 73 70 61 63 65 2F 63 6F 6D 6D  right/space/comm
000050A0  61 2F 68 79 70 68 65 6E 2F 70 65 72 69 6F 64 2F  a/hyphen/period/
000050B0  73 6C 61 73 68 2F 6F 6E 65 29 2F 46 6F 6E 74 46  slash/one)/FontF
000050C0  69 6C 65 33 20 32 36 20 30 20 52 3E 3E 0A 65 6E  ile3 26 0 R>>.en
000050D0  64 6F 62 6A 0A 32 20 30 20 6F 62 6A 0A 3C 3C 2F  dobj.2 0 obj.<</
000050E0  50 72 6F 64 75 63 65 72 28 47 50 4C 20 47 68 6F  Producer(GPL Gho
000050F0  73 74 73 63 72 69 70 74 20 38 2E 31 35 29 0A 2F  stscript 8.15)./
00005100  43 72 65 61 74 69 6F 6E 44 61 74 65 28 44 3A 32  CreationDate(D:2
00005110  30 31 30 30 34 31 38 31 33 34 35 33 30 29 0A 2F  0100418134530)./
00005120  4D 6F 64 44 61 74 65 28 44 3A 32 30 31 30 30 34  ModDate(D:201004
00005130  31 38 31 33 34 35 33 30 29 0A 2F 54 69 74 6C 65  18134530)./Title
00005140  28 4D 69 63 72 6F 73 6F 66 74 20 57 6F 72 64 20  (Microsoft Word
```

THE
NATIONAL
ARCHIVES

Apache™ Tika

- Wanted a tool to extract metadata from file formats
- Tika can detect and extracts metadata and text from over a thousand different file types.

Example:

```
java –jar “C:\Users\ASD1\Documents\ArchiveSchoolSession3\Software\TikaMetadata\tika-app-1.22.jar” –m “C:\Users\ASD1\Documents\Archive School Session 3\corrupted_Dates\Draft DDRO 05.docx”
```

And without –m

```
java -jar H:\DigitalLab\tikameta\tika-app-1.22.jar “C:\Users\ASD1\Documents\ArchiveSchoolSession3\corrupted_dates\Draft DDRO 05.docx”
```

Created a script to run TIKA over a directory of folders

https://github.com/paulyoung84/TikaMetadata

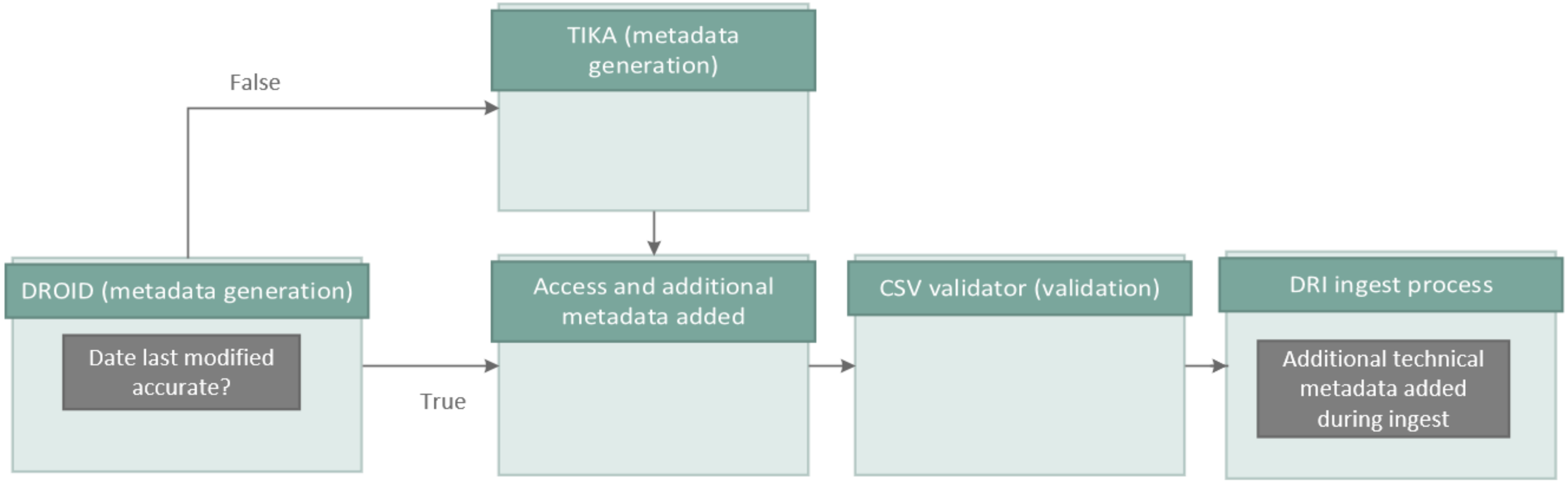
THE
NATIONAL
ARCHIVES

Exercise: TikaMetadata

- Open folder ‘Software’ and then open ‘TikaMetadata’ click on TikaMetadata.bat
- This will open a cmd terminal window which prompts you to add the folder you want to scan, drag folder ‘corrupted_dates’ into screen, this should populate with location of that folder
- Check ‘TikaMetadata’ folder for output csv file ‘TikaMetadata-Output_corrupted_dates.csv’, open and compare metadata grabbed from different formats

THE NATIONAL ARCHIVES

Tika metadata workflow


TIKA (metadata generation)
False
DROID (metadata generation)
Date last modified accurate?
True
Access and additional metadata added
CSV validator (validation)
DRI ingest process
Additional technical metadata added during ingest

THE NATIONAL ARCHIVES

Photo by Marcel Friedrich on Unsplash

Photo by Dan Gold on Unsplash

UTF8 Validator

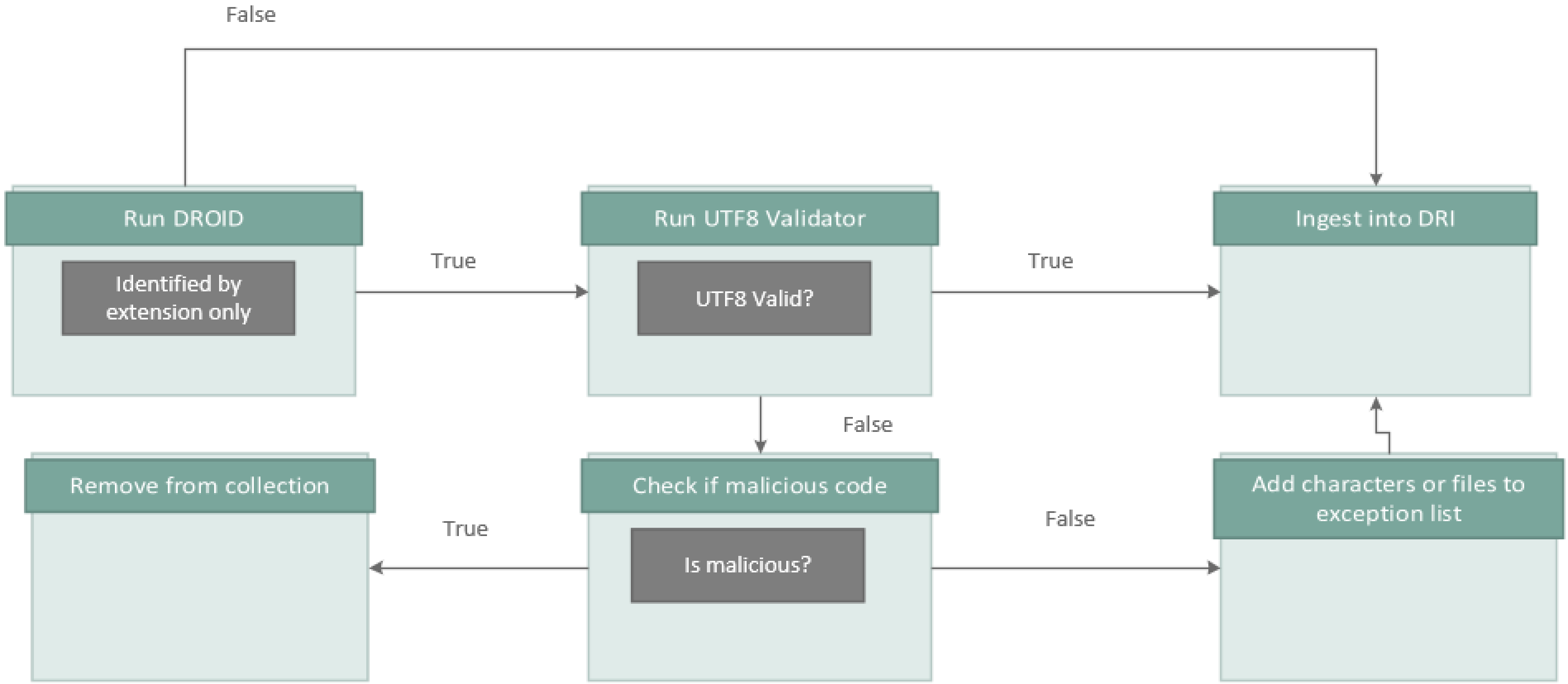
- Used to determine if files which identify be Extension only contain characters which are not UTF8 valid

- This is to highlight any possible files which may contain malicious binary code. Most extension only files are text based and will commonly have UTF8 characters.

- Once we are aware of which files contain invalid UTF8 characters we can check if they appear to be malicious or not, most can be ingested safely on an exception list

- We also ensure our metadata is completely UTF8 valid so we can ensure it can be presented correctly

UTF8 Validator workflow

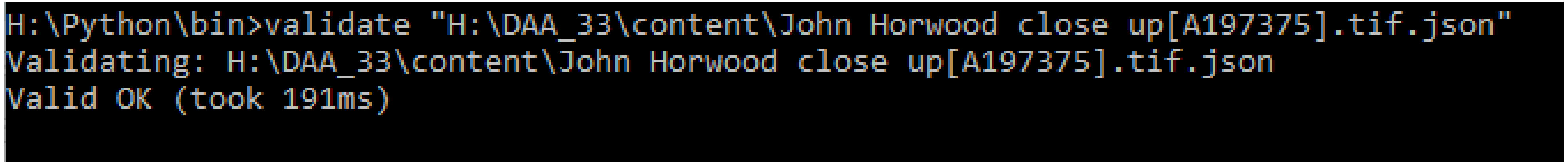


THE NATIONAL ARCHIVES

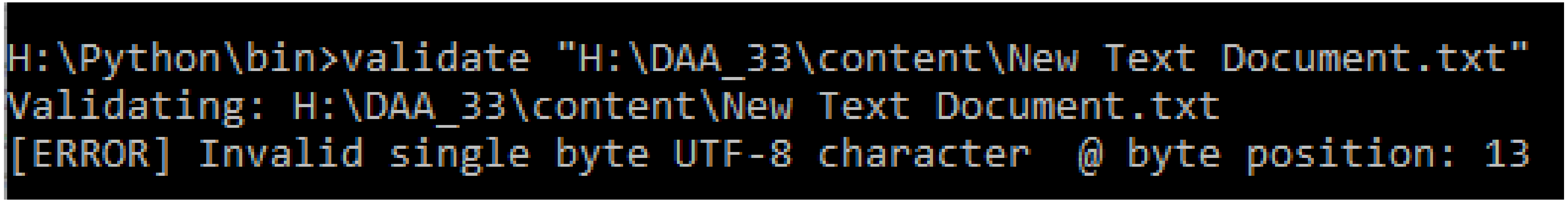
UTF8 Validator

https://github.com/digital-preservation/utf8-validator

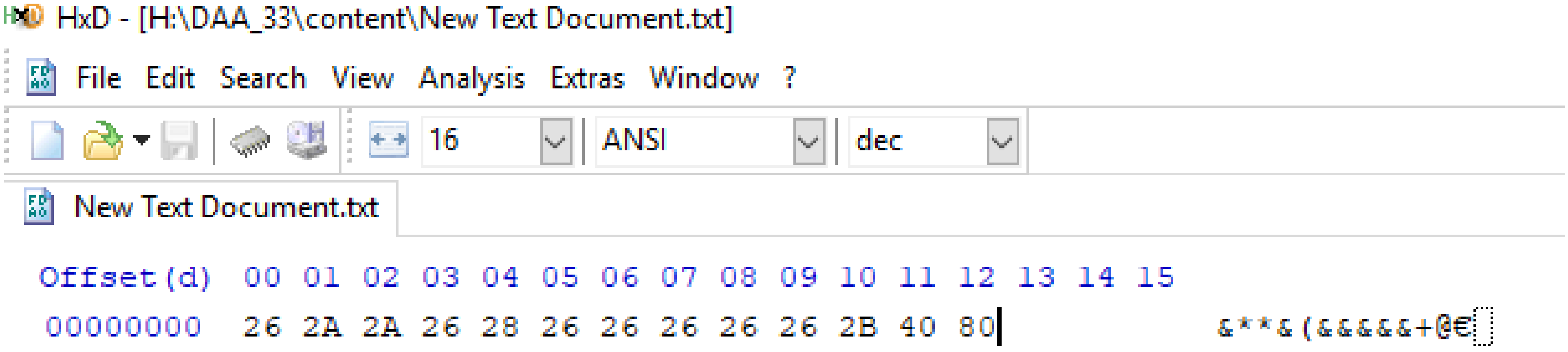
- Cmd line tool – validate.bat file.txt



```
H:\Python\bin>validate "H:\DAA_33\content\John Horwood close up[A197375].tif.json"
Validating: H:\DAA_33\content\John Horwood close up[A197375].tif.json
Valid OK (took 191ms)
```



```
H:\Python\bin>validate "H:\DAA_33\content\New Text Document.txt"
Validating: H:\DAA_33\content\New Text Document.txt
[ERROR] Invalid single byte UTF-8 character  @ byte position: 13
```



HxD - [H:\DAA_33\content\New Text Document.txt]

```
Offset(d)  00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15
00000000   26 2A 2A 26 28 26 26 26 26 26 2B 40 80                 &**&(&&&&&+@€
```

THE
NATIONAL
ARCHIVES

UTF8 validator batch

https://github.com/paulyoung84/utf8validatorbatch/blob/master/utf8batchvalidator.py

Batch script which runs from a DROID export CSV and validates all files which identified as extension only to determine if they contain invalid UTF8 characters.

- Prints to console all files which are not valid
- Saves to a CSV file all invalid filenames as well

Homework

- Use the DROID reports which you generated for your own material to create your own metadata CSV files.
- Think about what metadata fields you want to collect for your collection, what do you need to ensure the records can be preserved as well as findable and reusable?
- Remove columns which you do not require for long term preservation.
- Add in any additional columns which you feel are required.
- Think about (the logic) behind new csv schema rules which you could create to ensure any additional columns could be validated.
- Once you have created your metadata CSV, please email it to us at
- **ArchiveSchool@nationalarchives.gov.uk**

THE
NATIONAL
ARCHIVES

Bonus Homework (optional)

- Edit the schema used in this session, so that it can be used to validate your new metadata CSV
- Run the CSV validator over your collection and validate the metadata you have created using your new schema.
- Please contact us to discuss for help in developing schema rules and dealing with any validator errors. We are happy to help ☺
- **ArchiveSchool@nationalarchives.gov.uk**

Photo by Morvanic Lee on Unsplash

ArchiveSchool@nationalarchives.gov.uk

Feedback Survey for Session 3

https://tinyurl.com/y3vo5h8t

THE
NATIONAL
ARCHIVES