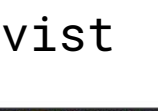
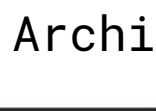
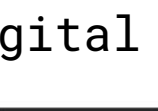
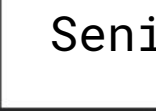


Anna de Sousa
Senior Digital Archivist



THE
NATIONAL
ARCHIVES



Photo by [Drew Beamer](#) on [Unsplash](#)

	<div>Schedule</div> <div><ul style="list-style-type: none">▪ Bits and bytes, character encodings, storage media▪ Break▪ Copying, checksums, digital preservation at TNA▪ Lunch▪ Knowing what you have, practical DROID exercise, discuss DROID report▪ Break▪ Normalisation, migration, feedback survey, homework▪ Finish</div>														

Aims of the course

To provide an overview of the core principles of digital preservation, give attendees the experience of hands on usage of some digital preservation tools and enable them to be able to apply the knowledge gained to their own place of work.

Course modules – session 1 & 2

Session 1 – Knowing what you have. Part 1 – **1 November 2019**

An overview of digital preservation and practical use of a file format identification tool to run over your digital files and provide information on your digital collections.

Session 2 – Knowing what you have. Part 2 – **29 November 2019**

In a continuation of knowing what you have, looking at file format research and how you can get your unidentified formats added to PRONOM and DROID.

THE

NATIONAL

ARCHIVES

Digital Preservation – one step at a time

Digital Preservation Coalition (DPC)

- <https://dpconline.org/handbook/glossary>
 'Digital Preservation Refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.'
- The core part of all digital preservation is the notion of action – it is not one process but rather a series of actions that allow for digital preservation to occur.
- Essentially it involves making decisions about which interventions to take over long periods of time in order to mitigate risk to the record.
- The upside is that even if you can only take one or two actions you are already carrying out digital preservation and can develop a more complete workflow over time.

THE

NATIONAL

ARCHIVES



Adventure so the
BEGINNS

Bits and bytes

- A bit is the smallest unit of storage and can be either a 0 or a 1. Bit is short for binary digit.
- Binary is a numbering system used by computers using only the digits 0 and 1.
- Computers are made up of digital circuits with switches that can only be on or off.
- All functions carried out by a computer are determined by which switches are on or off.
- 0 is off and on is 1.
- A byte is made up of 8 bits.
- One byte can store one character e.g. A or £, but one character may be made up of more than one byte.
- A byte can represent 256 different combinations
- All storage is measured in bytes e.g. Kilobytes → Megabytes → Gigabytes → Terabytes etc.

THE

NATIONAL

ARCHIVES

Character encoding

- When you type, the character encoding your computer is using maps the characters to bytes in computer memory. When you reopen the document your computer uses that encoding to read the bytes back into characters.
- If you create text using one character encoding and someone tries to read it on a computer or in a piece of software using a different character encoding, some of the original meaning may well be lost as the characters you chose may now not be properly displayed.
- It is sometimes obvious when this happens as the text is nonsensical or contains characters that do not make sense. Often however, the differences can be very subtle and hard to spot with the human eye, and may only become apparent when attempting computer processing of the data.

THE

NATIONAL

ARCHIVES

Character encoding

- UTF-8 is an attempt to create a universal character set, that incorporates every character glyph in existence and gives each character a unique and consistent code-point (byte value). This means that data interchange becomes more consistent and a person in Europe opening a document typed in Japanese will see each character as intended by the author, as long as they were both using UTF-8 encoding.
- At The National Archives we do not insist on a particular character encoding for born digital files transferred to us.
- We do insist that all accompanying metadata csv files are UTF-8 encoded.
- This allows us to ensure that all metadata is captured and stored in a consistent manner and can allow for us receiving metadata in any language without the need to convert the character encoding.

THE

NATIONAL

ARCHIVES

	<div><div><h3>UTF-8</h3><ul style="list-style-type: none">A bit is the smallest unit of storage and stores either a 0 or a 1. Bit is short for binary digit.Binary is a numbering system used by computers using only the digits 0 and 1.Computers are made up of digital circuits with switches that can only be on or off.All functions carried out by a computer are determined by which switches are on or off.0 is off and on is 1.A byte is made up of 8 bits.One byte can store one character e.g. A or £, but one character may be made up of more than one byte.A byte can represent 256 different combinationsAll storage is measured in bytes e.g. Kilobytes -> Megabytes -> Gigabytes -> Terabytes etc.</div><div><h3>Big5 (Chinese character encoding)</h3><ul style="list-style-type: none">* A bit is the smallest unit of storage and stores either a 0 or a 1. Bit is short for binary digit.* Binary is a numbering system used by computers using only the digits 0 and 1.* Computers are made up of digital circuits with switches that can only be on or off.* All functions carried out by a computer are determined by which switches are on or off.* 0 is off and on is 1.* A byte is made up of 8 bits.* One byte can store one character e.g. A or ?, but one character may be made up of more than one byte.* A byte can represent 256 different combinations* All storage is measured in bytes e.g. Kilobytes ;V> Megabytes ;V> Gigabytes ;V> Terabytes etc.</div></div>																	
																	THE	
																	NATIONAL	
																	ARCHIVES	



Types of storage media for digital files

- High Density Disk (HDD), Solid State Drive (SSD) = types of hard drive. Most computers have at least one hard drive inside, but you can also buy portable hard drives.
- HDD = Spinning plates inside, each holds data. Moving parts, slower than SSD but as a result a lot cheaper than SSD. Maximum today is around 10 TB.
- Portable hard drives used at TNA are HDD and hold up to 2TBs.
- SSD = no moving parts, faster ability for the computer to read or write to the drive (I/O – input/output). Maximum today is around 4 – 8 TB (but much more expensive than HDD)
- Flash drives (USB sticks) = teeny SSD! Maximum today about 1TB
- Optical storage = CDs/DVDs/Blu Rays – written and read by optic laser based devices
- Tape = usually stored on a cartridge, similar technology to that used by VHS or cassette but better storage data and digital. Current generation of LTO 8 tapes can hold up to 30 TB.

THE

NATIONAL

ARCHIVES

When all is not what it seems

- The way that folders and files are organised for viewing on a computer or portable hard drive replicate what we would expect to see in the paper world.
- In reality, the way that files are stored depends on the file system. This defines how data is written to and read from storage media by an operating system. For example NTFS is a Microsoft file system.
- Different file systems have different limitations.
- Folders are a representation of order rather than a digital object and tell your computer where files are located. Folders can also be referred to as directories e.g. T:\AdeSousa\Archive_School\Archive_School_Session1.pptx
- Files can be split and saved in multiple fragments across a drive.
- A master file table on the hard drive knows about all data and where it is stored. The master file table is stored in a hidden section on the drive reserved for this purpose. The master file table stores metadata about the files and directories on the hard drive.

THE

NATIONAL

ARCHIVES

Generating a directory list

- A directory list text file lists all the files and folders in a specified directory (screenshot examples on next 2 slides)
- Open a command prompt – on Windows type cmd when you click on the magnifying glass on bottom left of your taskbar to bring up the command prompt.
- Navigate to where the drive/folder is that you want to create a directory list by typing:
 - If for a drive, just type the drive letter followed by a colon and hit enter e.g. E:
 - If for a folder type cd followed by the folder name you want to generate a directory list for – type a backspace before the next folder name until you are at the required folder then hit enter e.g. cd Documents\Archive_School
- You will see a > appear after the final folder name you entered. After this type:
- dir /s /b /a /o:N > \directorylist.txt
- The directory list will now appear at the top level of the drive.
- If you want it to appear in the specific folder, repeat the filepath before the directory list name e.g. H:\Documents\Archive_School>dir /s /b /a /o:N
>H:\Documents\Archive_School\directorylist.txt
- Name the directory list in a way the reflects the collection so you don't end up with lots of directorylist.txt files e.g. HCA32directorylist.txt for a collection on series HCA 32

THE

NATIONAL

ARCHIVES

Generating a directory list – drive level

An example if generating a directory list for drive with the drive letter X

Command Prompt

```
Microsoft Windows [Version 10.0.17763.805]
(c) 2018 Microsoft Corporation. All rights reserved.

H:\>X:

X:\>dir /s /b /a /o:N >\directorylist.txt
```

THE

NATIONAL

ARCHIVES

Generating a directory list – folder level

An example if generating a directory list for the contents of the 'Archive_School' folder that is inside your Documents folder

Command Prompt

```
Microsoft Windows [Version 10.0.17763.805]
(c) 2018 Microsoft Corporation. All rights reserved.

H:\>cd Documents\Archive_School

H:\Documents\Archive_School>dir /s /b /a /o:N > H:\Documents\Archive_School>\directorylist.txt
```

THE

NATIONAL

ARCHIVES



Photo by [Dan Gold](#) on [Unsplash](#)

	<h1>Copying</h1>														
	<h2>Copying within a device</h2>														
	<ul style="list-style-type: none">▪ Telling the hard drive to make a copy of data and paste it somewhere else.▪ The computer consults the master file table to see where in the hard drive it can find that file (fragments can be in multiple locations).▪ It will copy bit for bit each block at an available empty address space (an unused part of the drive).▪ It will record the locations of the new fragments into the master file table. The user will be able to see a complete copy in the location they've specified.▪ Some file systems may just create a new reference to the same data i.e. you can see the copy in the location you've chosen but the file system has not created a duplicate copy.														
														THE	
														NATIONAL	
														ARCHIVES	

Copying

Copying across devices

- Storage device (internal hard drive or portable storage) will be connected to your computer via a bus – a communication system that transfers data within a computer.
- An example is USB – Universal Serial bus. When you move data from one storage device to another it goes via the bus (all aboard!)
- External portable hard drive connection: Motherboard in the computer has USB attached to the USB port, to which you attach the USB cable to the portable hard drive.
- Copying across devices = the motherboard data controller transmits data from itself via a cable to the portable hard drive and then the hard drive has to write the data and store it.

THE

NATIONAL

ARCHIVES

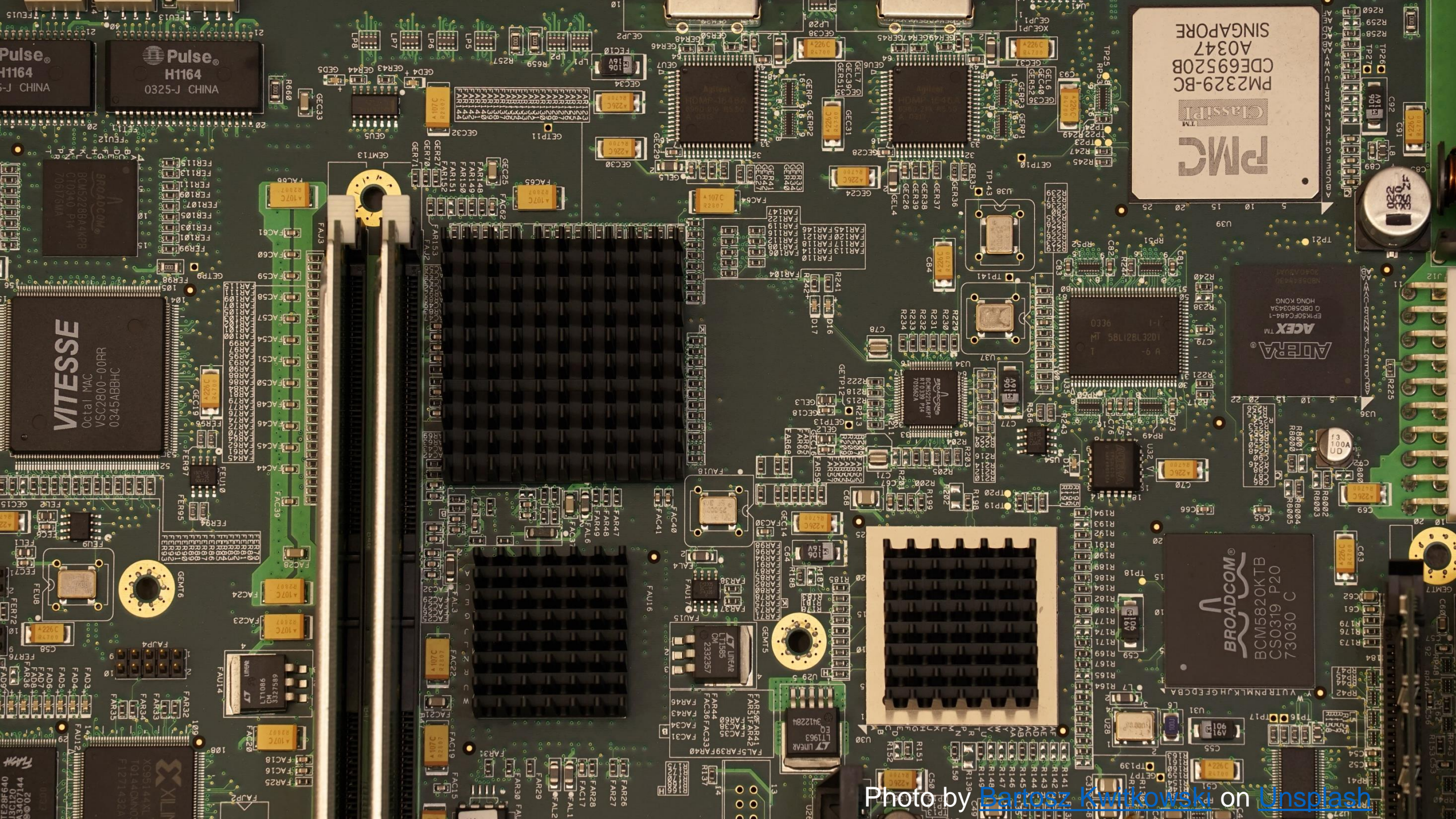


Photo by [Bartek Kwilowski](#) on [Unsplash](#)

Copying

Copying to or from the internet

- If connecting to the internet to copy to cloud or to download to your computer from a website or SFTP connection, Transmission Control Protocol/Internet Protocol (TCP/IP) - defines how computers communicate over networks.
- One of TCP/IP primary functions is to ensure the integrity of that communication, confirming that data packets are sent, received and assembled correctly.
- Transmitting over TCP involves inbuilt integrity checks that ensure the data of the end result matches what was sent from the originator.
- User datagram protocol (UDP) = 'fire and forget'. Unlike TCP/IP the web service firing out UDP does not carry out integrity checks on what is sent and received. This approach is most commonly used within web streaming – as you're watching a film on a streaming service there are probably a number of data packets getting lost along the way, but the viewer will rarely see an impact.

THE

NATIONAL

ARCHIVES

Copying – the hazards

- Power might glitch during copying.
- Cable connecting devices might have damage.
- Reading from hard drive to hard drive might go wrong in terms of the master file table – you could copy only fragments of a file rather than the entire file.
- The more component parts involved, the more things that can go wrong with the copy. So it is more risky to copy across devices than to copy within a device.
- The most common point for copying issues to occur is when copying from a device to portable hard drive or vice versa.
- Don't despair – you can use tools such as Teracopy to verify the success of a copy and also utilise checksums to ensure what you have copied is the same as the master version.
- If your TCP session abruptly ends, you'll be left with an incomplete transfer, which could result in corruption. There's various tools built into modern browsers and other applications now that will attempt to re-establish lost sessions, and resume in-progress downloads without loss, but things can still go wrong.

THE

NATIONAL

ARCHIVES

Checksums

- A 'fingerprint' for each file, made up of letters and numbers e.g.
c9e94209fb3e602d60fca3ec869051d444f42b49e8a4f22031f3c469b825d92d
- You can generate a new checksum for a file and check it against the original checksum, to ensure that the content of the file has not changed – during copying or over time.
- A change in checksum will mean either someone has changed the content of the file and then saved it or that the file has been corrupted in some way.
- **A change to the filename will not change the checksum.**
- Checksums are not a good way to determine the intellectual content of images. Images of the same object will often have different checksums due to automated image capture information such as timestamp of image generation.
- SHA 256 checksum for a 0 byte file is
e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855

THE

NATIONAL

ARCHIVES

Checksums

- No two checksums will be the same unless the content and format of a file are identical*:
 - If you copied and pasted a file and changed it's name it would still have the same checksum.
 - If you had the same word processed content in two word files the checksum could be different due to the stored hidden metadata e.g. username, creation date etc.
 - You can use checksums to identify duplicates but it cannot be used for content comparison.

*There's an infinitesimally small chance of two random data objects having the same checksum, but the chance is greater than zero. I've not verified it but this stackoverflow answer (<https://stackoverflow.com/a/288519>) claims the chance of a random MD5 clash to be '1 in 340 undecillion 282 decillion 366 nonillion 920 octillion 938 septillion 463 sextillion 463 quintillion 374 quadrillion 607 trillion 431 billion 768 million 211 thousand 456'

THE

NATIONAL

ARCHIVES



Photo by [Clément Falize](#) on [Unsplash](#)

Digital Preservation at The National Archives

- Engage with depositors early and often!
- Clean – virus scanning x 2 on receipt, identify and remove unnecessary system files (Knowing what you have & Keeping what you have safe)
- Authenticate via checksum = generated by depositor before arrival, confirmed upon receipt at The National Archives. We currently use secure hash algorithm 256 (SHA 256)
(Knowing what you have & Keeping what you have safe)
- Identify formats via DROID/PRONOM = generated by depositor before arrival, confirmed upon receipt at The National Archives
(Knowing what you have)

THE

NATIONAL

ARCHIVES

Digital Preservation at The National Archives

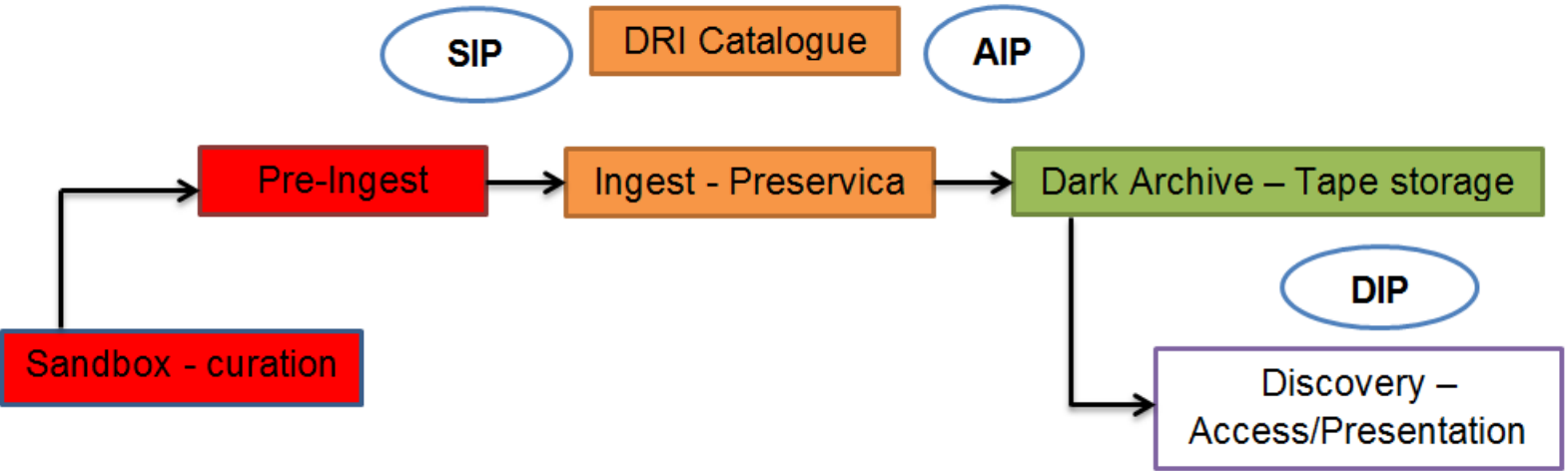
- Multiple copies for resilience
(Keeping what you have safe)
- Enforce metadata standards
(Describing what you have)
- Maintain our own infrastructure
(Keeping what you have safe)
- Keep preservation separate from presentation
(Keeping what you have safe & Access)

THE

NATIONAL

ARCHIVES

Digital Preservation at The National Archives



KEY

- Unsafe – data is unsanitised
- Safe – data is sanitised
- Super Safe – data is securely packaged

SIP = Submission Information Package

AIP = Archive Information Package

DIP = Dissemination Information Package

THE

NATIONAL

ARCHIVES

What have we observed?

- Media decay *does* happen – so move the content and keep it safe!
- Hardware obsolescence is rare in mainstream systems
- Format/data obsolescence is real, but minor. Formats have standardised: 99% of The National Archives' own records are made up of around 20 formats
- Storage failure/corruption – storage hardware failure can be catastrophic, but storage environments are managed to mitigate risk
- Incomplete/inadequate capture is our most immediate threat. We can only preserve what is captured and it only makes sense if described correctly.

THE

NATIONAL

ARCHIVES



Photo by [Lee Cartledge](#) on [Unsplash](#)

Knowing what you have. Part 1

As with any object you are looking to preserve, be it physical or digital, the first step is to know what it is!

In the digital world knowing what you have usually entails:

- Being able to identify its file format
- Knowing its file name
- Knowing its file size
- Knowing its checksum
- Knowing what software was used to create it (useful, but not always possible!)
- Knowing date related information about the file e.g. date last modified

THE

NATIONAL

ARCHIVES

Why each factor is important

- **Being able to identify its file format**
 - A file format is the method of storing digital content in a file in a structured consistent manner
 - Structure is based upon technical specifications that define the structure
 - The structure can be interpreted by software to facilitate rendering
 - Theoretically not bound to one particular software (not always the case!)
 - File format identification demonstrates the file conforms to the technical specifications of that file format - not just saved with an extension!

THE

NATIONAL

ARCHIVES

Why each factor is important

- **Knowing its file name**
- The file name has no impact on the preservation requirements of a file but it does allow for file management. With born digital files a duplicate file name does not necessarily indicate duplicate file contents!
- **Knowing its file size**
- Bytes -> Kilobytes -> Megabytes -> Gigabytes -> Terabytes -> Petabytes
- (To work out upwards x 1000 to work out downwards / 1000)
- Knowing the size of files can help you to plan storage requirements.
- It also indicates if there is something wrong with a file or if a file is empty – if you have a zero byte file, you can investigate further to determine the best course of action.

THE

NATIONAL

ARCHIVES

Why each factor is important

- **Knowing its checksum**
 - Knowing the checksum allows you to carry out fixity checking – regenerating the checksum and comparing it to the original to ensure no corruption or changes to the file have taken place.
 - It provides assurance you have received what your depositors intended to send you.
 - Checksums generated after copying can alert you to a failure in the copying process.
 - You can use checksums to discover exact duplicates.
- **Knowing date related information about the file**
 - With archival material the date last modified indicates when the record was last changed and saved. This can provide useful contextual information.
 - Be warned! Date last modified is not infallible, it can change when files are copied from one server to another, so this date may not reflect the last date a user actively modified the content of a file.

THE

NATIONAL

ARCHIVES

Magical* ways to know what you have

- With digital preservation when you have questions you want to answer, search for the free tool that enables you to find the answers you want!
- At The National Archives, to help us know what we have we use a tool called DROID (other tools are available!)
- DROID = Digital Record Object Identification:
 - Identifies file formats (using PRONOM)
 - Captures file name, file path, date last modified, checksum file size and more
 - Alerts you to an extension mismatch and any unidentified files.
 - Scans internal byte code of digital files
 - Uses PRONOM registry signature files at its core
 - Has command line and Graphical User Interface (GUI)

* Not really magic

THE

NATIONAL

ARCHIVES

DROID

- Developed by The National Archives, open source and freely available:
<https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/>
- User guide also available:
<https://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf>
- Embedded within Digital Preservation, Information Management, Digital Forensics and other tools e.g. Preservica

THE

NATIONAL

ARCHIVES

PRONOM

- File format registry
- Over 1600 entries (PUIDs)
- Format extensions, mime/media types, links to documentation
- File format identification signatures - for DROID! (Other file format identification tools are available!)
- <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

THE

NATIONAL

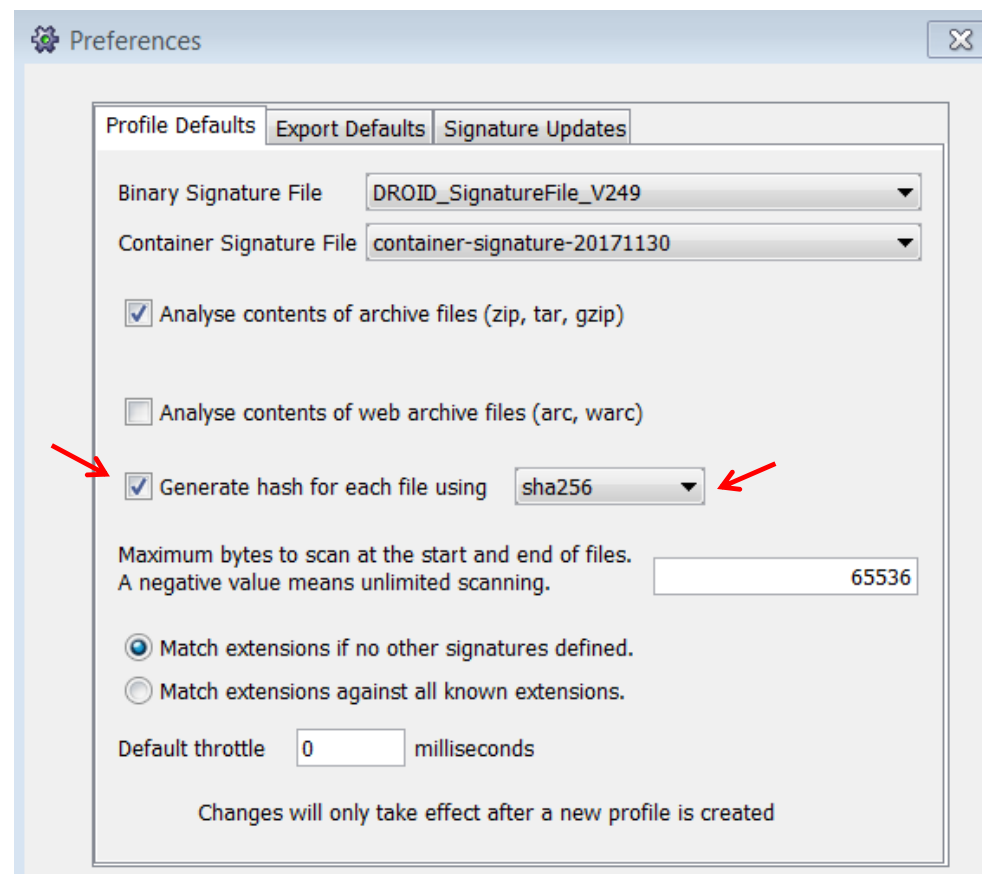
ARCHIVES



Photo by [Mikhail Vasilyev](#) on [Unsplash](#)

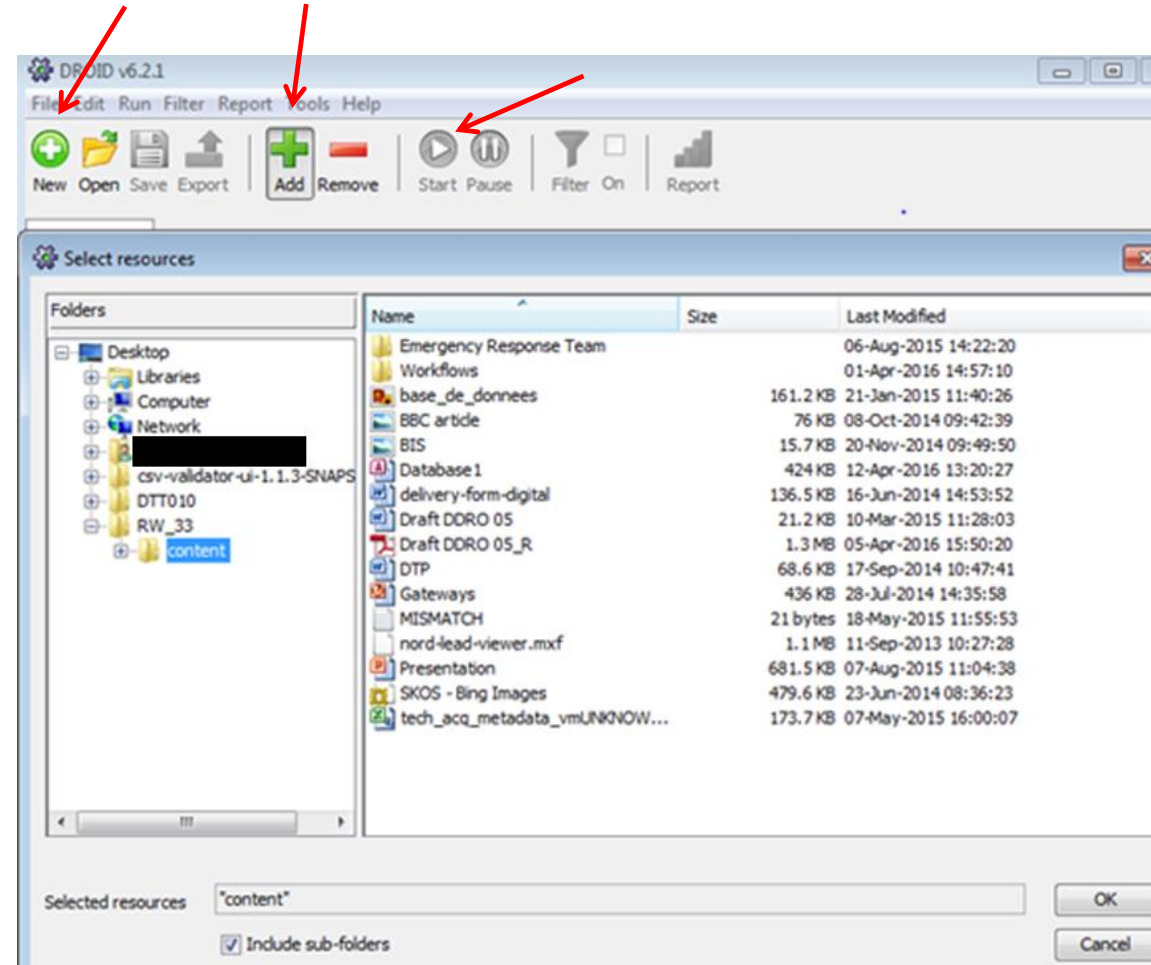
How to use the DROID GUI

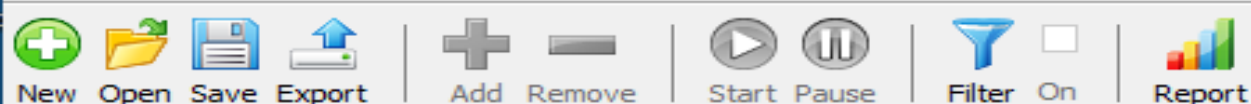
- Go into the DROID folder –right click on the file named droid.bat and select ‘send to’ and then ‘Desktop (create shortcut)’. This way you can go immediately to the droid shortcut on your desktop rather than having to locate the DROID files on your machine each time.
- Select ‘preferences’ from the ‘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.



How to use the DROID GUI

- 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 the top level folder that contains the files that you want DROID to scan. 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).










Untitled-1 x

Resource	Extension	Size	Last modified	Ids	Format	Version	Mime type	PUID	Method	Hash
Q:\Digital Preservati...			29/07/19 16:16							
MyFolder			29/07/19 16:16							
MyOtherDuplicat...	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange ...	1.01	image/jpeg	fmt/43	Signature	5536eea0cb9e62c13..
Thumbs.db	db	11 KB	22/05/14 14:14		OLE2 Compound Docu...			fmt/111	Signature	39da5573b74cf7241c..
My7ZippedCat.7z	7z	23.8 KB	05/12/13 15:03		7Zip format			fmt/484	Signature	6630132e6b67f9e61..
MyCat.jpg	jpg	23.8 KB	13/09/19 14:12		JPEG File Interchange ...	1.01	image/jpeg	fmt/43	Signature	5536eea0cb9e62c13..
MyZippedCat.zip	zip	23.7 KB	05/12/13 15:02		ZIP Format		application/zip	x-fmt/263	Signature	c23109bad49705c3e..
MyCat.jpg	jpg	23.8 KB	05/12/13 14:48		JPEG File Interchange ...	1.01	image/jpeg	fmt/43	Signature	5536eea0cb9e62c13..
MyBrokenCat.jpg	jpg	22.9 KB	05/12/13 14:49							1f0331db9f049a20af..
MyCat.jpg	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange ...	1.01	image/jpeg	fmt/43	Signature	5536eea0cb9e62c13..
MyCat.logfile	logfile	201 bytes	05/12/13 16:06							316602627da99c656..
MyCommaSeparate...	csv	24 bytes	05/12/13 14:52		Comma Separated Values		text/csv	x-fmt/18	Extension	d30d7bc85aa9ee6d1..
MyCorruptedPDFC...	pdf	15 bytes	05/12/13 14:57							0ec906063e9aac2dd..
MyDisguisedCat.jpg	jpg	8 bytes	05/12/13 14:51		Windows New Executable			x-fmt/410	Signature	aceb85682dd7b8c92..
MyDuplicateCat.jpg	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange ...	1.01	image/jpeg	fmt/43	Signature	5536eea0cb9e62c13..
MyEmptyCat.jpg	jpg	0 bytes	23/05/14 10:15							e3b0c44298fc1c149a..
MyExecutableCat...	exe	8 bytes	05/12/13 14:51		Windows New Executable			x-fmt/410	Signature	aceb85682dd7b8c92..
MyOtherDuplicateC...	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange ...	1.01	image/jpeg	fmt/43	Signature	5536eea0cb9e62c13..
MyPDFCat.pdf	pdf	26.5 KB	05/12/13 14:55		Acrobat PDF 1.5 - Port...	1.5	application/pdf	fmt/19	Signature	2c335681f09f689f20..
MyPlainTextCat.txt	txt	17 bytes	05/12/13 14:52		Plain Text File		text/plain	x-fmt/111	Extension	e41c2026639dda50d..

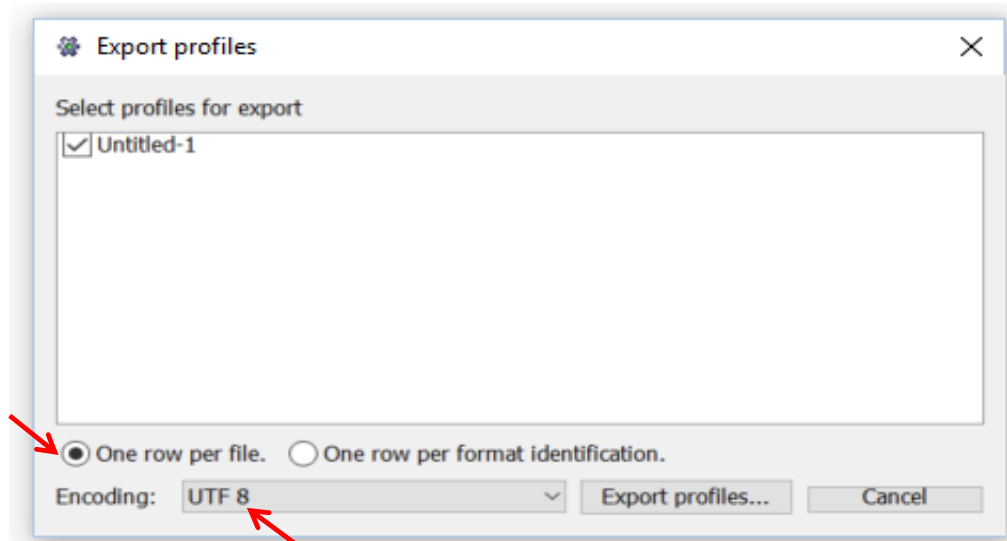
GUI results symbols explained

- In the extension column, a warning icon  means that there is an extension mismatch – specifically the extension of the file does not match any extensions listed in the related PUID entry.
- In the IDs column, the blue cog  means the file has matched with exactly one signature, or that it hasn't matched with a signature, but has an extension that matches exactly one PRONOM entry (e.g. .log, or .csv).
- The unlit bulb  means it has matched with no signatures or extensions at all.
- A blue, underscored number here (e.g. 8) indicates that there are multiple matches – usually this happens because it hasn't matched with any signature but has a common extension, such as 'doc' – in which case it'll list all PRONOM PUIDs that have the doc extension, but sometimes it can be that the file matches against multiple signatures – this is known as a 'signature clash' and is something we try to eliminate and should be notified of. The number represents the number of matches.
- On the resource column, the icon that looks like a small package  is an archival container (such as zip, tar, 7z, rar etc.) that DROID has attempted to recursively scan. Next to it should be a little 'plus' sign that allows you to expand the view to see what files are inside.
- Much rarer, on the resource column, an icon with a 'no entry' sign on it  means that DROID has been unable to scan the resource, normally because it lacks the permissions to read the contents.

THE
NATIONAL
ARCHIVES

How to use the DROID GUI

- 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'. You will then be given the option to save the report.



THE

NATIONAL

ARCHIVES

Time to talk through the DR0ID report

THE

NATIONAL

ARCHIVES



Photo by [Dan Gold](#) on [Unsplash](#)

Mitigating risk over time

- Digital preservation at it's core is making decisions about which interventions you need to take over time in order to mitigate risk to your collections.
- Which interventions are most appropriate will depend on the make up of your collections
- Two common interventions are:
- **Normalisation** – converting born digital file formats into a number of pre selected formats on ingest, rather than preserving them in their original format.
- **Migration** – converting the format of the file usually due to the threat of obsolescence of the original format or playback media.

THE

NATIONAL

ARCHIVES

Normalisation

- Normalising specific file types on ingest e.g. normalising all audio files to WAV, will mean that there are fewer variations of file types to preserve and can mean that the risk of file format obsolescence is easier to manage.
- The National Archives do not carry out normalisation, as we do not want to risk affecting the integrity and authenticity of the records we preserve. We preserve whatever our depositors provide to us in its original form.
- Normalisation carries some risk of losses in formatting information and loss of metadata held within the original file.
- Normalisation does not mean that obsolescence is no longer a threat, as you may need to take future normalisation actions on your previously normalised files if new risks related to your chosen file formats are identified.

THE

NATIONAL

ARCHIVES

Format Migration

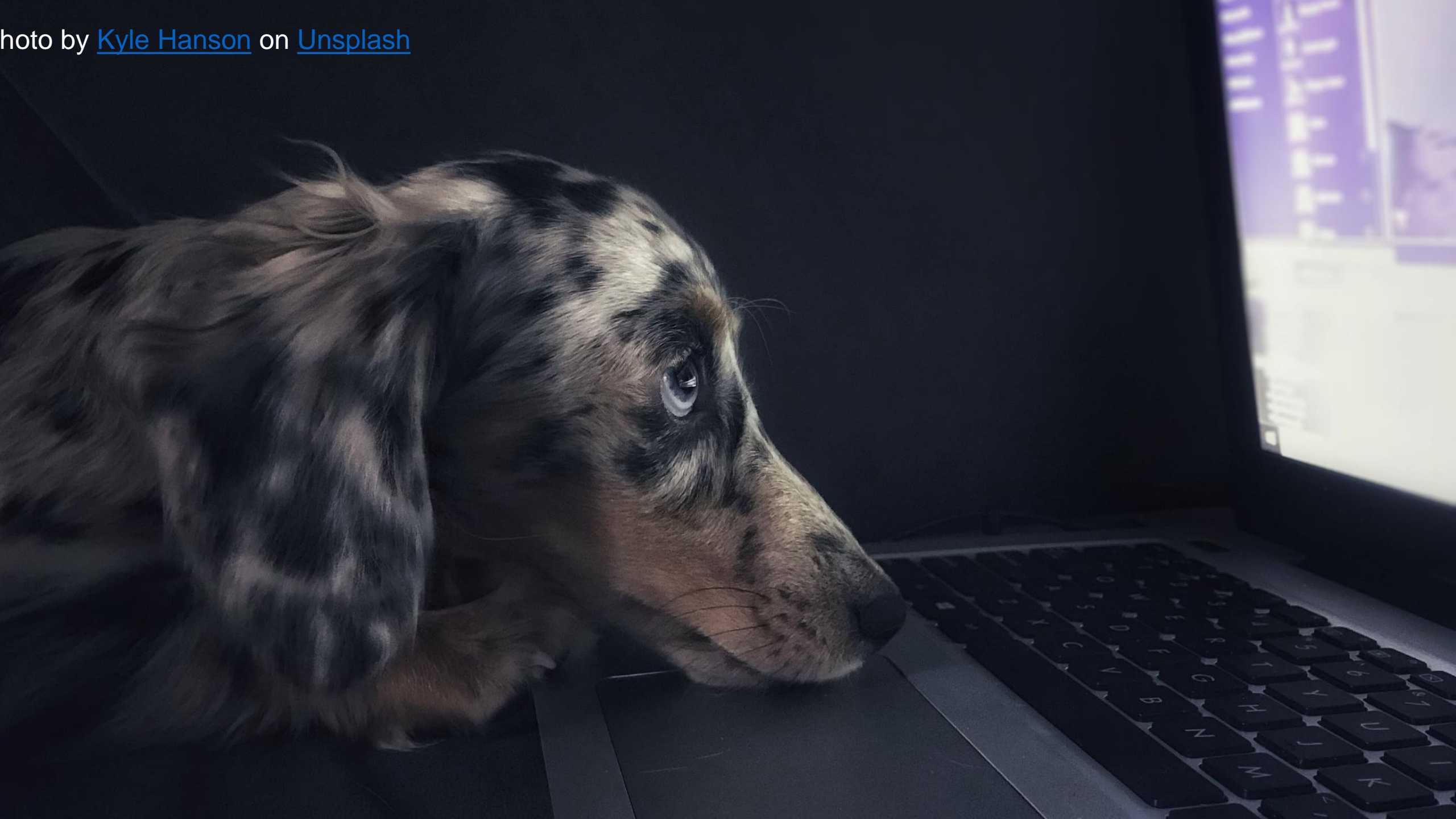
- Converting a file from one format to another, usually due to suspected threat of obsolescence of the file format. For example moving from MS Word Version 6 to MS Word for Windows 2010
- The National Archives carries out substitution of JP2 to JPEG and MXF to MP4 for access purposes, because JPEG and MP4 are much more accessible formats for general archive users. This action does not impact the master file, which we preserve in its original form in perpetuity.
- More information about preservation actions can be found in the DPC handbook: <https://dpconline.org/handbook/organisational-activities/preservation-action>

THE

NATIONAL

ARCHIVES

hoto by [Kyle Hanson](#) on [Unsplash](#)



Feedback Survey for Session 1

<https://tinyurl.com/y3vo5h8t>

THE

NATIONAL

ARCHIVES

Homework

- Run DROID over a collection of born digital files in your place of work.
- Ensure your preferences are set to generate SHA 256 checksums.
- Export your DROID report one row per file and UTF-8 encoded and save it as a csv.
- Look through your report and note down if you have any noticeable results – extension mismatch, unidentified formats etc.
- Bring samples of unidentified formats with you for Session 2.
- Email your DROID report to archiveschool@nationalarchives.gov.uk by 25 November 2019

THE

NATIONAL

ARCHIVES

Optional reading material

DROID

<https://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf>

DPC Handbook

<https://www.dpconline.org/handbook>

Character Encoding

<https://www.w3.org/International/questions/qa-what-is-encoding>

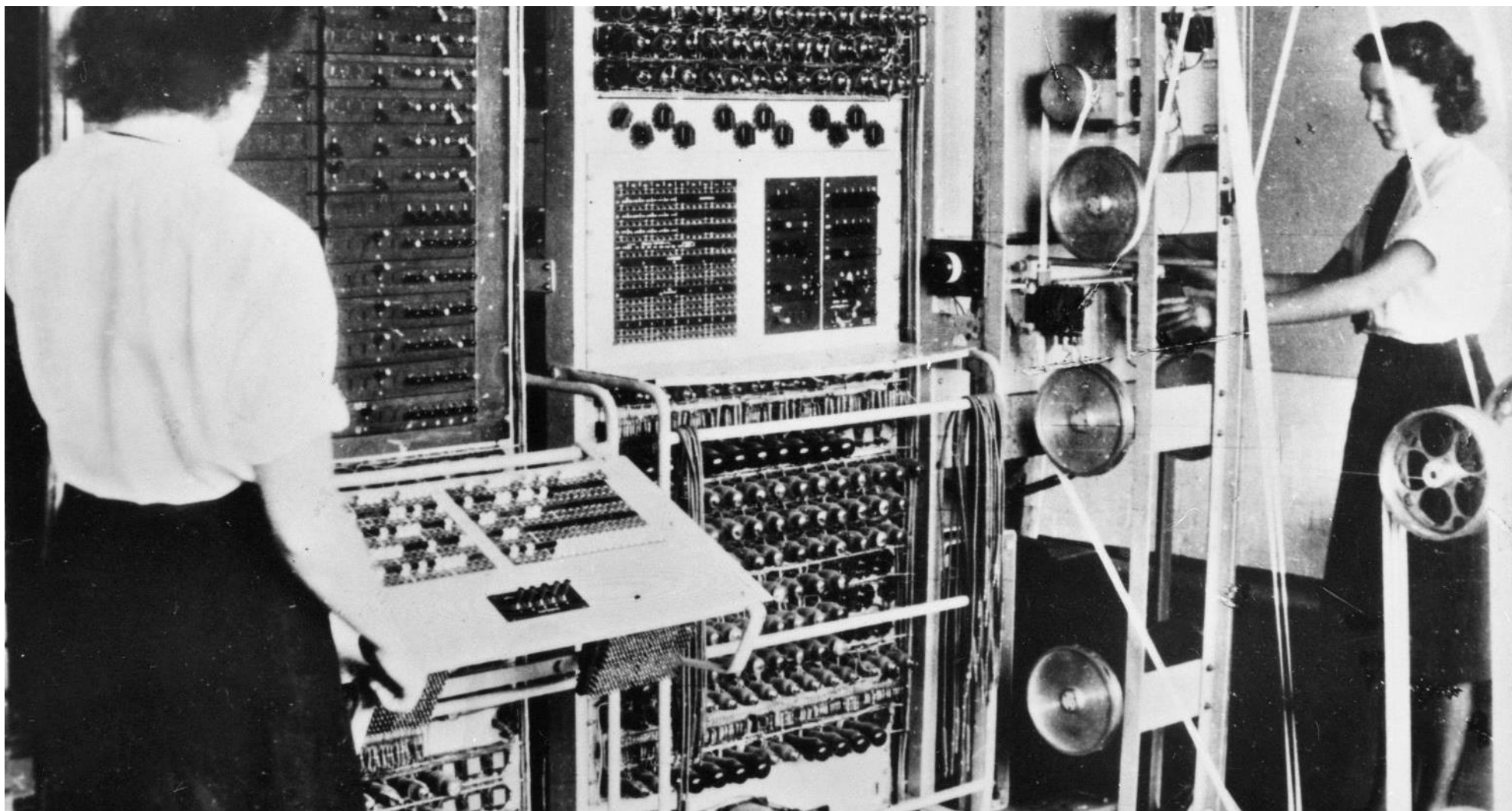
<https://www.joelonsoftware.com/2003/10/08/the-absolute-minimum-every-software-developer-absolutely-positively-must-know-about-unicode-and-character-sets-no-excuses/>

<http://kunststube.net/encoding/>

THE

NATIONAL

ARCHIVES



Colossus electronic digital computer

THE

NATIONAL

ARCHIVES