'DIY' DIGITAL PRESERVATION

For Software

PLAN

Plan and housekeeping

Part 1

10:45am-10:55am **Intros and housekeeping**

10:55am-11:15am **Digital preservation**
What it is and why it is important?

What material do you care about and hope to keep?

11:15am-11:25am **Obsolescence management**

Questions

11:25am-11:35am **[5-10 minute comfort break]**

Part 2

11:45am-12:55am **Software preservation**
What are the different approaches?

What kind of files are you working with? (Risks)

11:55am-12:05am **Case study**

Why are you looking to preserve software?

12:05am-12:15am **How to guide**

Discussions and Questions

WHO ARE WE

- **Bridging the Digital Gap**
  15month traineeship scheme

- UK National Archives
  (National Lottery Heritage Fund)

- Bringing ‘digital’ skills into the archives sector

THE NATIONAL ARCHIVES
TRANSPORT FOR LONDON
EVERY JOURNEY MATTERS
LONDON METROPOLITAN ARCHIVES
UNIVERSITY OF WESTMINSTER
ual:

WHY ARE WE DOING THESE WORKSHOPS

- Agitate the cultural record to reflect lived experience
- Embrace tools that support historical self-determination among non-specialist
- Raise awareness, share skills, knowledge exchange.

THE NATIONAL ARCHIVES

TRANSPORT FOR LONDON
EVERY JOURNEY MATTERS

LONDON METROPOLITAN ARCHIVES

UNIVERSITY OF WESTMINSTER

ual:

Bridging the Digital Gap

WHAT IS DIGITAL PRESERVATION

Digital material is vulnerable in different ways than analog material.

“a series of managed activities undertaken to ensure continued access to digital materials for as long as necessary.”

DPC, Digital preservation handbook

ARCHIVIST'S NIGHTMARE

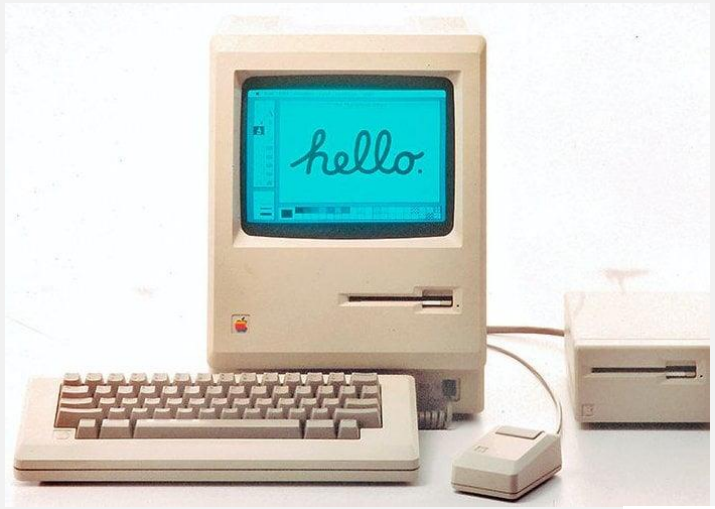
In real life

https://commons.wikimedia.org/wiki/File:Messy_storage_room_with_boxes.jpg

Bridging the Digital Gap

DIGITAL MATERIAL AND SOFTWARE RELIANCE

- Platform
- Software vendor
- Operating systems (MacOS Catalina no longer support 32-bit)
- Hardware requirement
- Browser support (Mainstream browers no longer support FLASH)
- Subscription



DIGITAL PRESERVATION AND BORN-DIGITAL

- Records that have been natively created in digital format
- Digital signal processing
- File formats
- Containers

WHY IS DIGITAL PRESERVATION NECESSARY?

Recover the past

Preserve our heritage

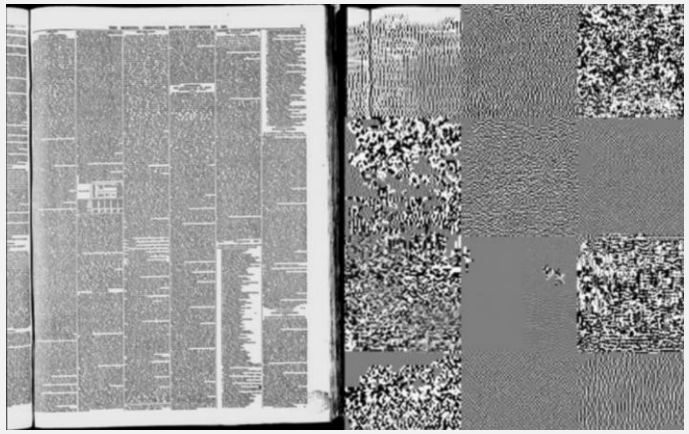
Share the knowledge

Prepare for the future

CHALLENGES FOR PERSONAL DIGITAL ARCHIVING

Technical

- Hardware failure
- Media failure (Bit rot)
- File corruption
- Virus/malware
- Media obsolescence(hardware, software, file format)



Non-technical

- Loss/theft, natural disasters.
- Unclear ownership/responsibility
- Lack of documentation
- Overdependence on third party solutions

EXAMPLES OF RISKS

From Nasa's Viking Project

From a Professional Photographer

Bridging the Digital Gap


WHAT MATERIAL DO YOU CARE
ABOUT AND HOPE TO KEEP?

Software Preservation

Why is it important?

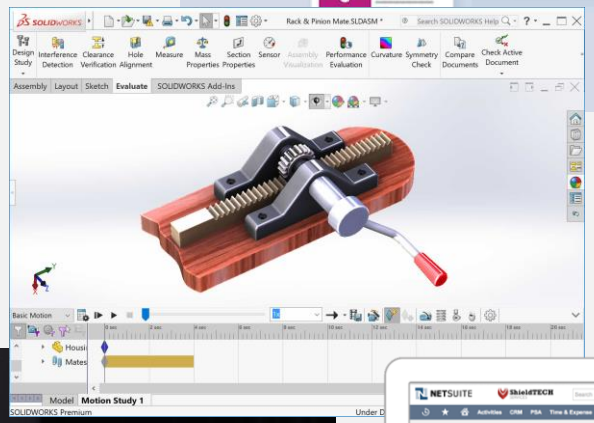
Bridging the Digital Gap

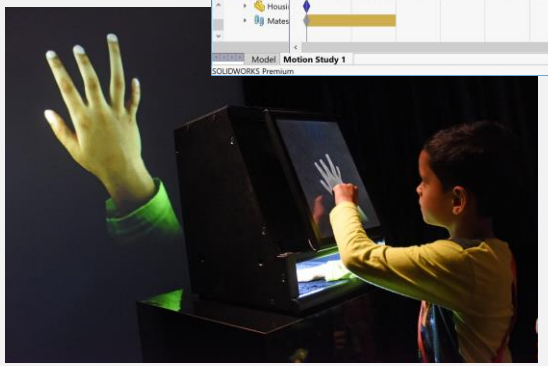
Software is a set of instructions, data or programs used to operate computers and execute specific tasks.

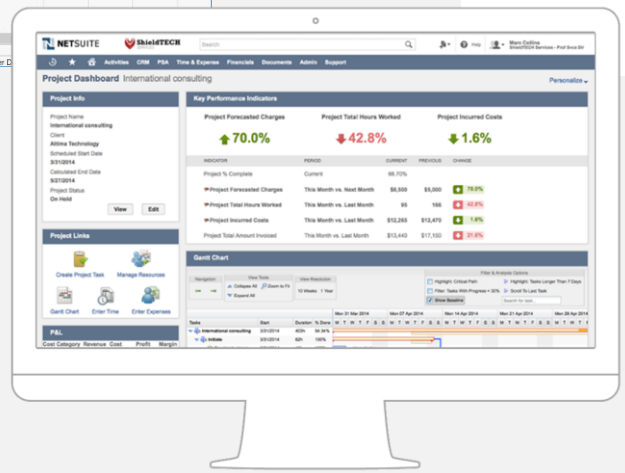
https://medium.com/@mohdfurqanhaiderrizvi/what-is-software-df18ef927e

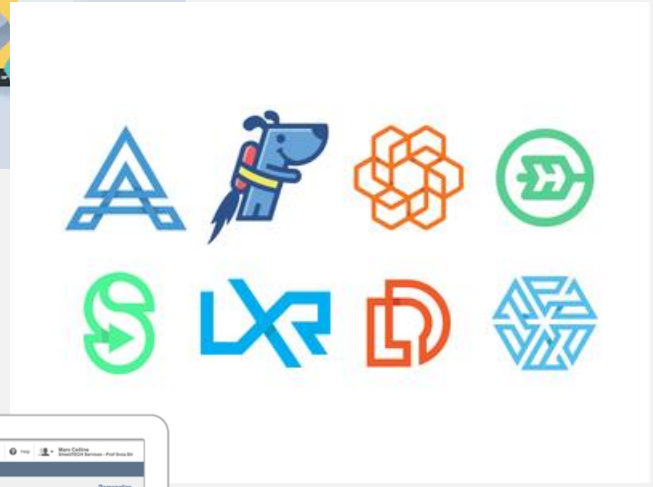
SOFTWARE IS COMPLICATED

- Audio
- Video
- Graphics
- Interface
- 3D objects
- Interaction









ABOUT YOU

Content Creator	Artist	Blogger	Designer
Developer	Web developer	Software developer	Data engineer
Researcher	Business analysist	Data scientist	User researcher
Digital Native	Tiktok	Gamer	Notetaker

WHY IS SOFTWARE PRESERVATION IMPORTANT

Encourage software reuse

- Reduced development cost
- Reduced development risk
- Accelerated development
- Increased quality and dependability
- Focused use of specialists
- Standards compliance
- Reduced duplication
- Learning from others
- Opportunities for comercialisation

Legal compliance and accountability

- Reduced exposure to legal risks
- Avoidance of liability actions
- Easily demonstrable compliance lessons audit burden
- Improved institutional governance.
- Enhanced reputation
- Social expectations met
- Sense of responsibility

Software Sustainability Institute, Software Preservation Benefits Framework

WHY IS SOFTWARE PRESERVATION IMPORTANT

Create heritage value

- **Create heritage value**
  - Heritage value is generally considered to be of intrinsic value

Enable continued access to data and services

For research and business

- Fewer unintentional errors due to increased scrutiny
- Reduced deliberate research fraud
- New insight and knowledge
- Increased assurance in results

For systems and services

- Current operations maintained
- Opportunity for improved operations via corrective maintenance
- Reduced vendor lock-in
- Improved disaster recovery response
- Increased organizational resilience
- Increased reliability

Software Sustainability Institute, Software Preservation Benefits Framework

DIFFERENT APPROACH APPROPRIATE TO DIFFERENT PURPOSE

	Technical preservation	Emulation	Migration	Cultivation	Hibernation
Achieve legal compliance and accountability	✓	✓	✓		
Create heritage value	✓	✓			
Enable continued access to data and services	✓	✓	✓	✓	✓
Encourage software reuse			✓	✓	✓

Software Sustainability Institute, Software Preservation Benefits Framework

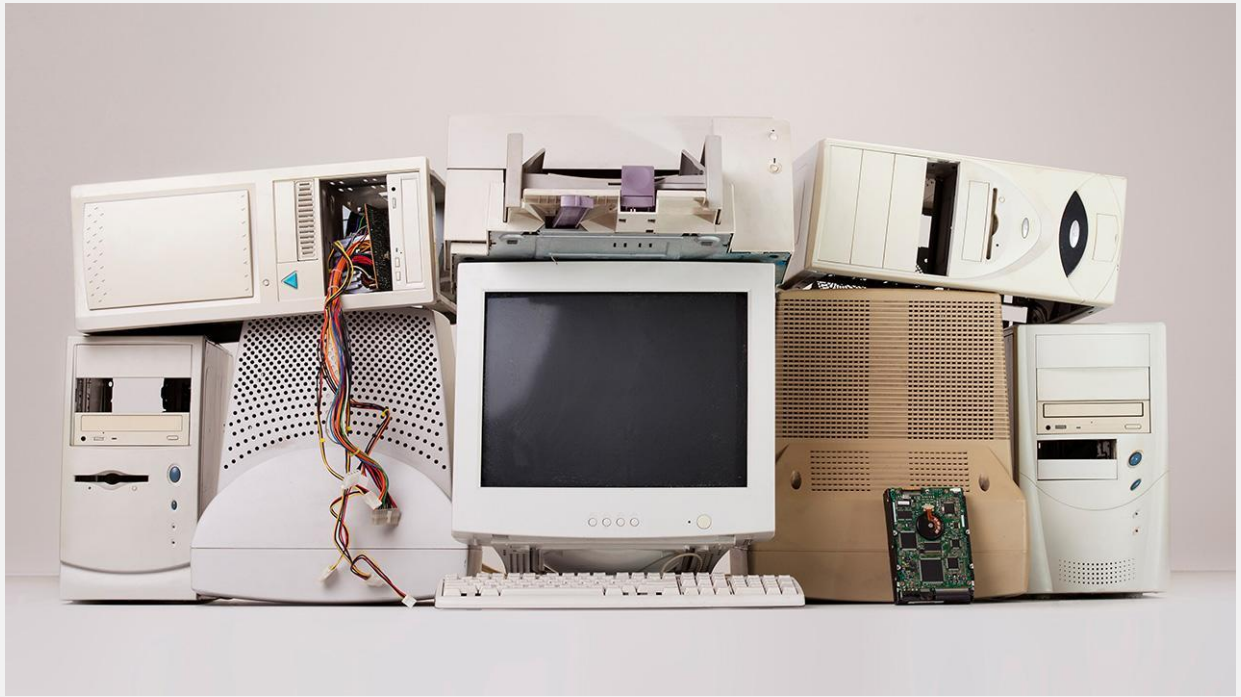
TECHNICAL PRESERVATION

Keeping original software and hardware in the same state. Works best when there is a known preservation period

- **Easy to do on your own**
  - Maintenance
  - Isolation

Things to do

- Purchase spares
- Regularly checking it still works
- Maintaining hardware
- Replacing hardware elements as they fail
- Scheduling review points in the calendar



No obsolete technology can be kept functional indefinitely

EMULATION

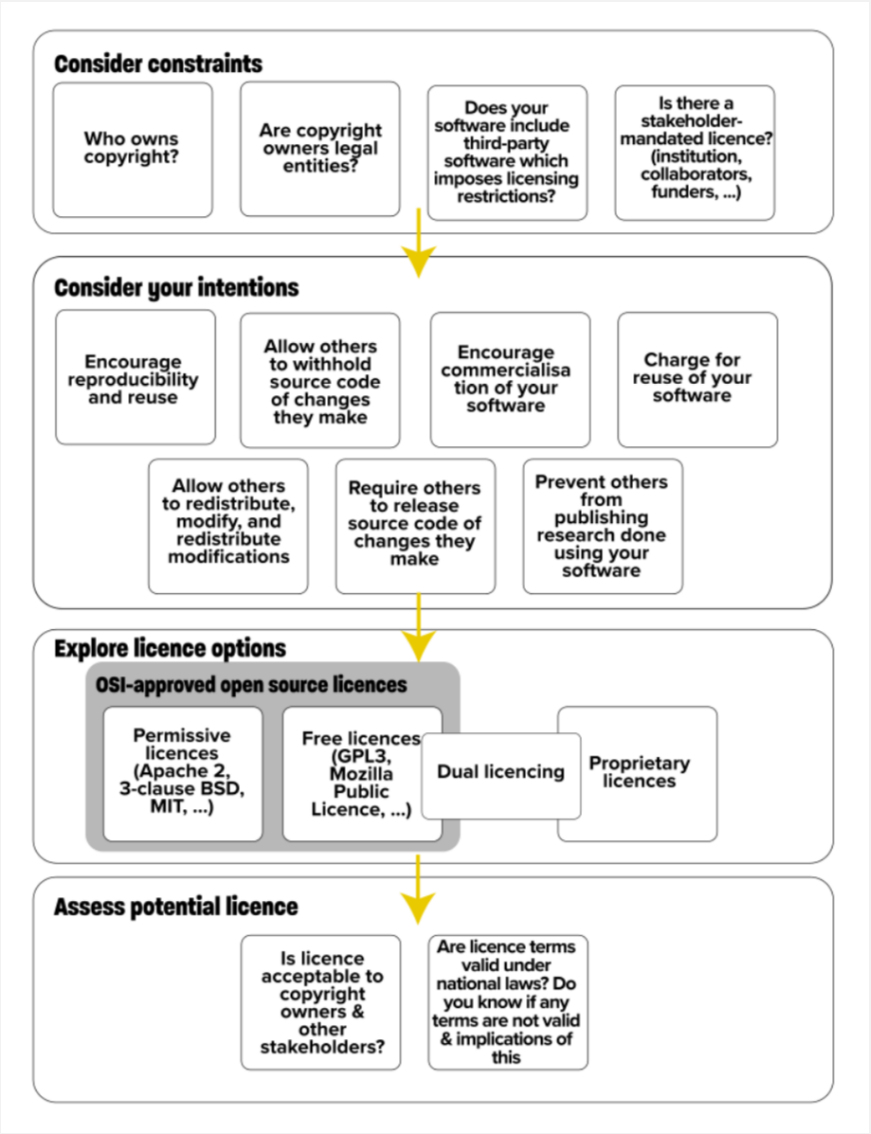
An emulator is a **software package that mimics your old hardware and operating environment.**

Flexibility to run on new hardware or cloud.

Readily available emulators and customised emulators available.

Things to do:

- Check licensing details
  - License type
  - License owner
  - License terms
  - Proof of license
- Testing
- Verifying and validating results
- Updating the emulator



Michael Jackson (ed.) (07 August 2018). Software Deposit: How to choose a software licence (Version 1.0). Zenodo. doi:10.5281/zenodo.1327316.

MIGRATION

Keeps the system functional with new technology.

The effort required for migration varies widely from small changes (e.g. reconfiguration), to major updates, or involving completely redeveloping the software from the original requirement.

Improved functionality, user experience with further development.
Improved performance with new hardware and platform.

Things to do

- Reconfiguring and recompiling
- Learning and using new programming languages

The cost is likely to match or exceed the initial development cost.

KEY QUESTIONS TO ASK YOURSELF

- Is there still knowledge and expertise to handle and run the software?
- How authentic does the preserved software need to be?
- How adequate does the preserved software need to be: should it perform exactly as the original, the same but with only minor deviations, or perform the core functionality only?
- How much access do you have? (Owner, developer, access to source code, access to hardware, user)
- Do you have the necessary Intellectual Property Rights (IPR)?
- What are you needing to preserve? (A few major pieces of functionality, Most of the functionality, but tolerant of minor deviations, All functionality, but fixing errors when found, Must perform exactly as original)
- What is your likely effort profile? (Something or nothing now, something or nothing in the future)
- What is the maintainability of underlying hardware?
- Is maintaining integrity and/or authenticity an important requirement?
- How long do you want to preserve it for?
- Can you afford it?
- Are you also interested in further development or maintenance?
- What development effort has been invested into the software so far?
- Is the software open source? Could it be made open source?

INFORMATION ABOUT YOUR SOFTWARE:

- Version of software
- Vendor/publisher
- Operating systems
- Hardware requirement
- Software installation guide
- Software specifications document
- License & Terms

ARCHIVE YOUR SOURCE CODE NOW

An initiative whose goal is to collect, preserve, and share software code—both freely licensed and not—in a universal software storage archive.

WHAT KIND OF FILES ARE YOU WORKING WITH?

CASE STUDY

~/code/src/github.com/nteract/nteract/applications/desktop/example-notebooks/altair.ipynb - idle

```
[7]  import altair as alt
     from vega_datasets import data

     cars = data.cars()
```

Faceted Scatter Plot with Linked Brushing

This is an example of using an interval selection to control the color of points across multiple facets.

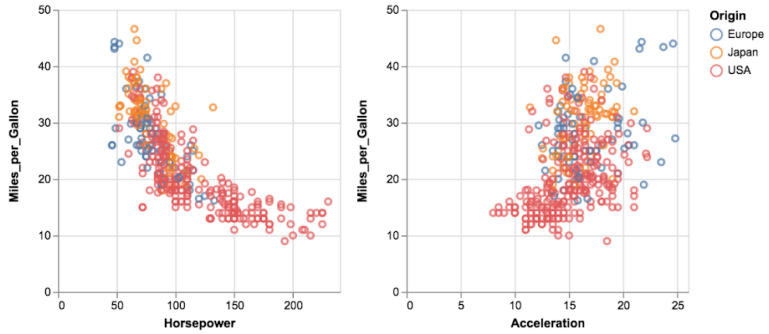
```
[8]  brush = alt.selection(type='interval', resolve='global')

     base = alt.Chart(cars).mark_point().encode(
         y='Miles_per_Gallon',
         color=alt.condition(brush, 'Origin', alt.ColorValue('gray'))
     ).add_selection(
         brush
     ).properties(
         width=250,
         height=250
     )

     print("Select a region in the chart below to try this out!")

     base.encode(x='Horsepower') | base.encode(x='Acceleration')
```

```
Select a region in the chart below to try this out!
```



python3 | idle

Last saved 2 minutes

https://netflixtechblog.com/notebook-innovation-591ee3221233


WHY ARE YOU LOOKING TO
PRESERVE SOFTWARE?

How to

Appraise, identify, organise, migrate, store

LOCATE YOUR MATERIAL

Hardware: floppy disks, CDs, USB/Flash drives, camera, mobile device

Shared drives: Google, DropBox, your institution/workplace

Other places: email attachments, chat history, social media

APPRAISE YOUR DIGITAL MATERIAL

Consider how you want to access your files:

- High quality exports only?
- Project files or linked files?
- Installations files?
- Process documentation?

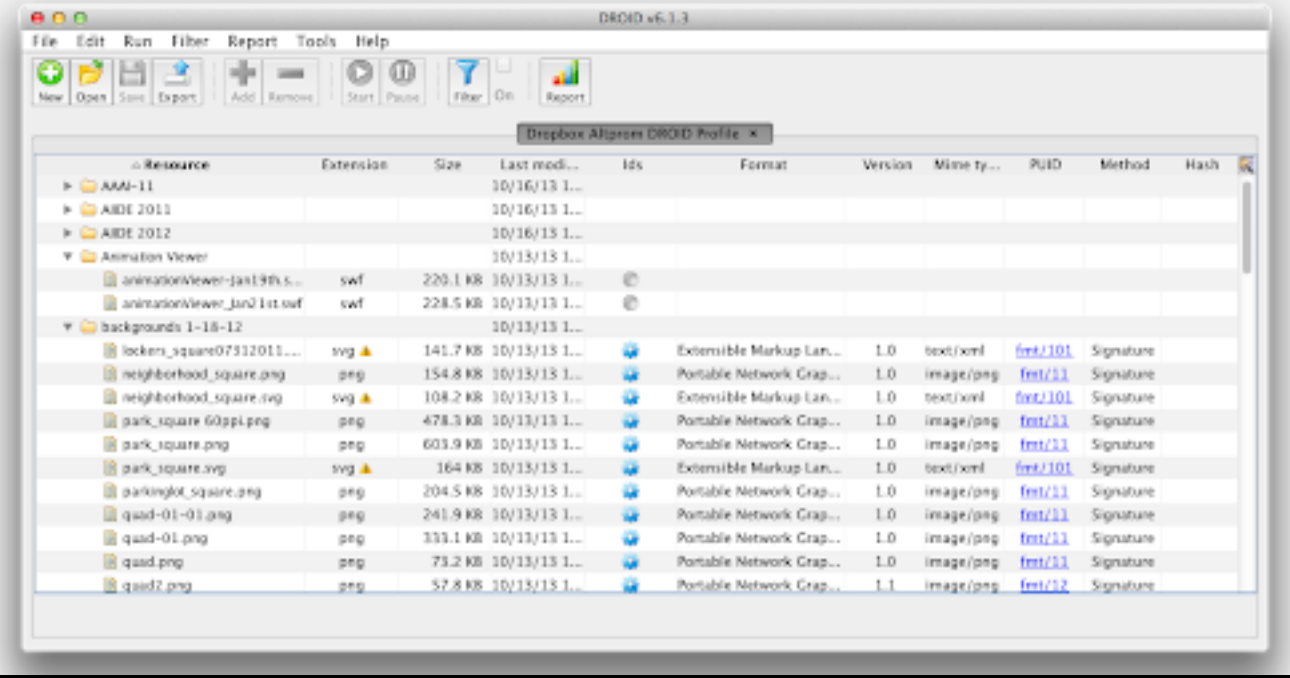
Suggestions:

- Deleting low-resolution duplicates
- **TreeSize Free** (Windows) or **GrandPerspective** (MacOS) can help visualise your files according to size and maintain any existing folders

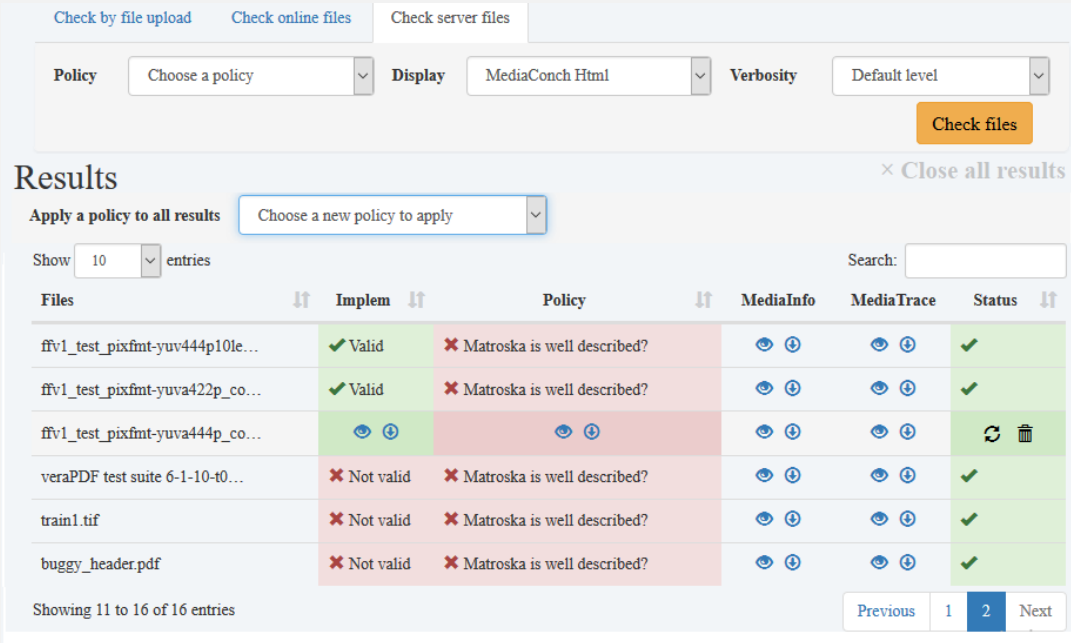
IDENTIFY WHAT YOU HAVE

- File extensions identification alone may not always be accurate
- Knowing what formats you have will support in determining how to care for them

Digital **R**ecord **O**bject **Id**entification **(DROID)**



MediaConch



ORGANISE YOUR FILES

- Use meaningful file names
- Avoid special characters
- Be consistent
- “Img_3081” vs. “2018_1087_1187” vs. “20181118_COL_nisha-voiceover_01”
- Many file-renaming tools on the web, e.g. **Bulk Rename Utility** (Windows)

EVALUATING FILE FORMATS AT-RISK OF OBSOLESCENCE

- **The following criteria should be considered by data creators when selecting file formats:**
  - Ubiquity
  - Support
  - Disclosure
  - Documentation quality
  - Stability
  - Ease of identification and validation
  - Intellectual Property Rights
  - Metadata Support
  - Complexity
  - Interoperability
  - Viability
  - Re-usability

Migrating to recommended preservation formats:

- Audio:
  WAV-PCM
- Video:
  FFmpeg
- Images:
  Jpeg 2000/TIFF

Look up PRONOM

bear in mind lossless versus lossy encoding for long-term preservation

https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

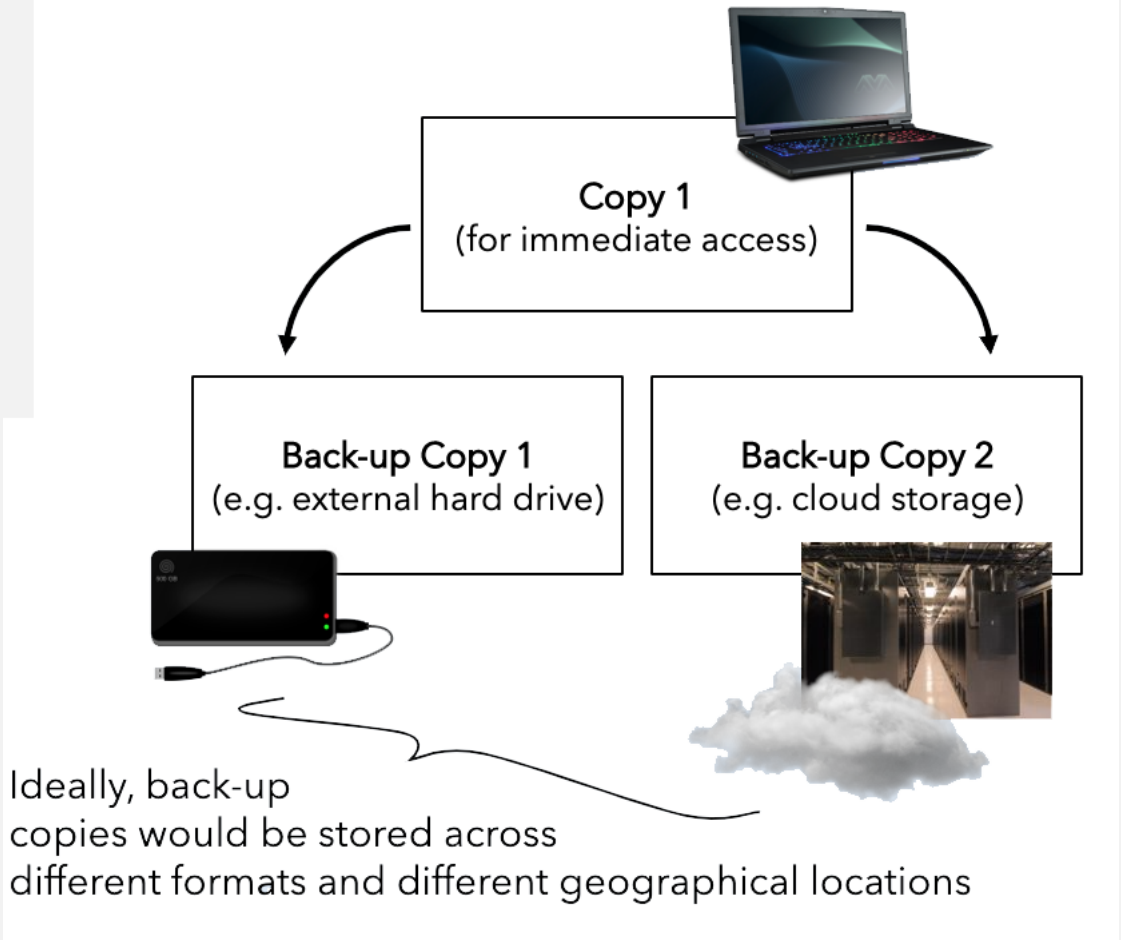
PRONOM

Summary

Name	Tagged Image File Format
Version	
Other names	TIFF
Identifiers	PUID: fmt/353 MIME: image/tiff Apple Uniform Type Identifier: public.tiff
Family	
Classification	Image (Raster)
Disclosure	Full
Description	The Tagged Image File Format (TIFF) is a raster image format originally developed by the Aldus Corporation, primarily for use in scanning and desktop publishing. When Adobe Systems Incorporated purchased Aldus in 1994, they acquired the rights to the TIFF format and have maintained it since then. TIFF files comprise three sections: an Image File Header (IFH), an Image File Directory (IFD), and the image data. TIFF files can contain multiple images (multi-page TIFF), and each image has a separate IFD. The IFH always appears at the beginning of the file, and is immediately followed by a pointer to the first IFD. The IFD contains metadata which describes the associated image, stored as a series of tags. The IFD also contains a pointer to the actual image data. TIFF supports colour depths from 1 bit to 24 bit (e.g. monochrome to true colour), and a wide range of compression types (RLE, LZW, CCITT Group 3 and Group 4, and JPEG), as well as uncompressed data.
Orientation	
Byte order	Little-endian (Intel) and Big-endian (Motorola)
Related file formats	Has lower priority than Exchangeable Image File Format (Uncompressed) (2.2) Has lower priority than Exchangeable Image File Format (Uncompressed) (2.1) Has lower priority than Exchangeable Image File Format (Uncompressed) (2.0) Has lower priority than Digital Negative Format (DNG) (1.1) Has lower priority than Tagged Image File Format for Electronic Photography (TIFF/EP) Has lower priority than Geographic Tagged Image File Format (GeoTIFF) Has lower priority than Tagged Image File Format for Internet Fax (TIFF-FX) Has lower priority than Sony ARW RAW Image File (1.x) Has lower priority than Kodak Digital Camera Raw Image File Has lower priority than Nikon Digital SLR Camera Raw Image File Has lower priority than Digital Negative Format (DNG) (1.0) Has lower priority than Digital Negative Format (DNG) (1.2) Has lower priority than Digital Negative Format (DNG) (1.3) Has lower priority than Canon RAW (2.0)

Technical Environment	
Released	01 Aug 1986
Supported until	
Format Risk	
Developed by	Aldus Corporation
Supported by	None.
Source	Digital Preservation Department / The National Archives
Source date	07 Jul 2011
Source description	PUID created for the TIFF format in response to the difficulties we have been having with multiple identification of the format and a consensus on a new interpretation of the standard from within The National Archives and outside with external stakeholders.
Last updated	13 Sep 2018
Note	

STORING YOUR DIGITAL FILES



- LOCKSS = Lots of Copies Keep Stuff Safe
- 3-2-1 back-up rule used by professionals
- Store copies in multiple locations, ideally in different formats

MAINTAIN YOUR DIGITAL FILES

- Refresh storage device (ideally every 5 years)
- Test access to files, spot check once a year
- Bear in mind digital preservation when creating files, particularly when using subscription software or emerging software/formats.
- Formats that are supported by a wide range of software or are platform-independent are most desirable.

THANK YOU

Corportatearchives@tfl.gov.uk

Bridging the Digital Gap

RESOURCES

- Workshop: https://software-carpentry.org/workshops/
- Technology watch report: https://www.dpconline.org/docs/technology-watch-reports/1460-twr15-01/file
- BitList: www.dpconline.org/our-work/bit-list
- Digital preservation handbook: https://www.dpconline.org/handbook/organisational-activities/creating-digital-materials#:~:text=Digital%20preservation%20refers%20to%20the,needs%20of%20the%20original%20creator.
- Recommended file formats statement: www.loc.gov/preservation/resources/rfs/audio.html
- Sustainability of digital Formats: https://www.loc.gov/preservation/digital/formats/
- Software Heritage: softwareheritage.org
- Benefits framework: https://www.software.ac.uk/sustainability-and-preservation-framework
- How to choose a software license: https://zenodo.org/record/1327316#.X2zNEshKgdU
- Droid: https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/
- PRONOM: https://www.nationalarchives.gov.uk/PRONOM/Default.aspx
- Case study on Netflix's use of Jupyter Notebook: https://netflixtechblog.com/notebook-innovation-591ee3221233