'DIY' DIGITAL PRESERVATION

For Software
## PLAN

### Plan and housekeeping

<table>
<thead>
<tr>
<th>Part 1</th>
<th>Part 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:45am-10:55am</td>
<td><strong>Intros and housekeeping</strong></td>
</tr>
</tbody>
</table>
| 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]** |
| 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
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.
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.”
ARCHIVIST’S NIGHTMARE

In real life
ARCHIVIST’S NIGHTMARE

In the digital world
DIGITAL MATERIAL AND SOFTWARE RELIANCE

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

<table>
<thead>
<tr>
<th>Technical</th>
<th>Non-technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Hardware failure</td>
<td>• Loss/theft, natural disasters.</td>
</tr>
<tr>
<td>• Media failure (Bit rot)</td>
<td>• Unclear ownership/responsibility</td>
</tr>
<tr>
<td>• File corruption</td>
<td>• Lack of documentation</td>
</tr>
<tr>
<td>• Virus/malware</td>
<td>• Overdependence on third party solutions</td>
</tr>
<tr>
<td>• Media obsolescence(hardware, software, file format)</td>
<td></td>
</tr>
</tbody>
</table>
EXAMPLES OF RISKS

From Nasa’s Viking Project

From a Professional Photographer
WHAT MATERIAL DO YOU CARE ABOUT AND HOPE TO KEEP?
Software Preservation

Why is it important?
**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
<table>
<thead>
<tr>
<th>ABOUT YOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Creator</td>
</tr>
<tr>
<td>Artist</td>
</tr>
<tr>
<td>Blogger</td>
</tr>
<tr>
<td>Designer</td>
</tr>
<tr>
<td>Developer</td>
</tr>
<tr>
<td>Web developer</td>
</tr>
<tr>
<td>Software developer</td>
</tr>
<tr>
<td>Data engineer</td>
</tr>
<tr>
<td>Researcher</td>
</tr>
<tr>
<td>Business analyst</td>
</tr>
<tr>
<td>Data scientist</td>
</tr>
<tr>
<td>User researcher</td>
</tr>
<tr>
<td>Digital Native</td>
</tr>
<tr>
<td>Tiktok</td>
</tr>
<tr>
<td>Gamer</td>
</tr>
<tr>
<td>Notetaker</td>
</tr>
</tbody>
</table>
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 commercialisation

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
<table>
<thead>
<tr>
<th></th>
<th>Technical preservation</th>
<th>Emulation</th>
<th>Migration</th>
<th>Cultivation</th>
<th>Hibernation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achieve legal compliance and</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>accountability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Create heritage value</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enable continued access to</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>data and services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encourage software reuse</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
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
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
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.

softwareheritage.org
WHAT KIND OF FILES ARE YOU WORKING WITH?
CASE STUDY
Faceted Scatter Plot with Linked Brushing

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

```python
import altair as alt
from vega_datasets import data
cars = data.cars()

brush = alt.selection(type='interval', resolve='global')

base = alt.Chart(cars).mark_point().encode(
y=alt.Y('Miles_per_Gallon', sort='x'),
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!](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 Record Object Identification (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

<table>
<thead>
<tr>
<th><strong>Name</strong></th>
<th>Tagged Image File Format</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Version</strong></td>
<td>TIFF</td>
</tr>
<tr>
<td><strong>Other names</strong></td>
<td>PUID: fmt/353</td>
</tr>
<tr>
<td></td>
<td>MIME: image/tiff</td>
</tr>
<tr>
<td></td>
<td>Apple Uniform Type Identifier: public.tiff</td>
</tr>
<tr>
<td><strong>Family</strong></td>
<td>Image (Raster)</td>
</tr>
<tr>
<td><strong>Disclosure</strong></td>
<td>Full</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Orientation</strong></td>
<td>Little-endian (Intel) and Big-endian (Motorola)</td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/rah/7cDigital%20Negative%20Format">Digital Negative Format (DNG) (1.1)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/Tagged_Image_File_Format">Geographic Tagged Image File Format (GeoTIFF)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/Tagged_Image_File_Format">Tagged Image File Format for Internet Fax (TIFF-FX)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/Tagged_Image_File_Format">Sony ARW RAW Image File (1.0)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/Tagged_Image_File_Format">Nikon Digital SLR Camera Raw Image File</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/rah/7cDigital%20Negative%20Format">Digital Negative Format (DNG) (1.0)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/rah/7cDigital%20Negative%20Format">Digital Negative Format (DNG) (1.2)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/Tagged_Image_File_Format">Digital Negative Format (DNG) (1.3)</a></td>
</tr>
<tr>
<td></td>
<td>Has lower priority than <a href="https://en.wikipedia.org/wiki/Tagged_Image_File_Format">Canon RAW (2.0)</a></td>
</tr>
</tbody>
</table>
### Technical Environment

<table>
<thead>
<tr>
<th>Released</th>
<th>01 Aug 1986</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported until</td>
<td></td>
</tr>
<tr>
<td>Format Risk</td>
<td></td>
</tr>
<tr>
<td>Developed by</td>
<td><a href="https://www.aldus.com">Aldus Corporation</a></td>
</tr>
<tr>
<td>Supported by</td>
<td>None.</td>
</tr>
<tr>
<td>Source</td>
<td><a href="https://www.archives.gov">Digital Preservation Department / The National Archives</a></td>
</tr>
<tr>
<td>Source date</td>
<td>07 Jul 2011</td>
</tr>
<tr>
<td>Source description</td>
<td>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.</td>
</tr>
<tr>
<td>Last updated</td>
<td>13 Sep 2018</td>
</tr>
<tr>
<td>Note</td>
<td></td>
</tr>
</tbody>
</table>
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

Ideally, back-up copies would be stored across different formats and different geographical locations
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
RESOURCES

- Workshop: https://software-carpentry.org/workshops/
- Technology watch report: https://www.dpconline.org/docs/technology-watch-reports/1460-twtr15-01/file
- BitList: www.dpconline.org/our-work/bit-list
- 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
- PRONOM: https://www.nationalarchives.gov.uk/PRONOM/Default.aspx
- Case study on Netflix’s use of Jupyter Notebook: https://netflixtechblog.com/notebook-innovation-591ee3221233