POWRR Institute, Day One

Technology Instruction Module: Operating Systems and File Formats Speaker Notes

# Slide 1 - Title Slide

* In this section, we’ll be looking at some of key issues you need to consider in relation to Operating Systems and File Formats.
* File formats in particular are a key consideration of digital preservation.

# Slide 2 – What is an Operating System?

* Operating Systems are the layer of system software that operates between the software programs we interact with and the computer’s hardware.
* The Operating System manages both the software programs and hardware, scheduling tasks relating to each to ensure the most efficient use of the computers capabilities.
* Operating Systems are required on all forms of computer platform including PCs and laptops, smart phones and tablets, and servers.
* Although all Operating Systems carry out similar functions, they have their individual quirks and it is important to be familiar with these to aid your digital preservation efforts.

# Slide 3 – Many Flavours: OS Family Tree

* Many different types of Operating Systems exist, as can be seen in the family tree diagram here. But the majority you will encounter are either Microsoft (DOS) or UNIX-based.
* For example the leading operating systems on the following platforms are:
	+ PCs/Laptops – More than 80% Windows, c. 10% MacOS (Unix) and c. 2% Linux (Unix)
	+ Smart phones/tablets – c. 90% Android and c. 10 iOS, both Unix
	+ Servers – Unix systems are dominant

# Slide 4 – Some Important Differences

There are many similarities and differences between Operating Systems, particularly between Microsoft (DOS)-based systems and UNIX-based systems. Some key differences to be aware of are:

* **Cost** – Microsoft Operating System software usually costs several hundred dollars, most UNIX-based systems are available free or very cheap as open source software.
* **Licences** – Microsoft OSs will be accompanied by strict commercial licences which restrict how they can be used and distributed. Licences for most UNIX-based systems are more open and allow for redistribution and reuse.
* **Customisation** – Microsoft OSs allow minimal customisation compared with UNIX-based systems. This is probably a negative for the user but a positive for digital preservation as it means there is more reliable consistency between systems.
* **Command line and GUIs** – Interaction with Microsoft OSs can be managed almost exclusively through graphical user interfaces, making the user experience more simple and consistent. Using UNIX-based systems will almost certainly require some use of the command line for actioning processes which can be intimidating for some users, but also (as with customisation) allows the user more power.
* **Storage** – Microsoft OSs organise information into files and folders, but the actual physical locations of the data can be spread across different parts of the storage, making it more difficult to copy to a new system with relying heavily on the OS. UNIX-based systems use the terminology files and directories for their storage structure and data is co-located so it is easier to find on the storage and to move and copy.
* One small but important detail to note if moving data between UNIX and Windows systems, and vice versa, is that Windows systems use backslashes for file locations while UNIX systems use forward slashes (like website URLs). These need to be converted if data is moved, thankfully there are free tools available to automate this process.

# Slide 5 – Getting to Know What’s Inside

* As already mentioned, one of the key differences between Microsoft and UNIX-based systems is the way files are structured.
* In Microsoft systems the main folders to the be aware of, and to look in for content for preservation, are C:/Windows, C:/Program Files, and C:/Users
* In UNIX-Based systems content is most likely to be in the /home (or /Users on Macs) or /mnt (/Volumes on Macs) directories, 🖰 the first being the users home directories and the second temporary mounted directories which may include shared content or external storage media.

# Slide 6 – Don’t Dear the Command Line

* The Command Line Interface was the primary way in which users interacted with early computers, before Graphical User Interfaces become commonly available.
* The command line is still used by many advanced users as it brings benefits such as:
	+ Quicker processing times as fewer system resources as used to execute an command line instruction
	+ Users have greater control over the processes actioned as well as more power and precision in issuing instructions to the computer
	+ It allows the automation of common processes through simple scripting, such creating folders/directories or moving data
* Becoming comfortable using some of the most simple commands is useful for digital preservation purposes as some tools only operate via the command line. But don’t panic, often only a few simple commands need to be used!
* There are differences between commands used in Microsoft and UNIX-based environments, but there are plenty of guides and introductions to their use.

# Slide 7 – File Formats: Just keep the bits…

* We’ll now move on to consider the main issues relating to file formats and digital preservation.
* There are 2 key issues relating to the preservation of file formats: retaining the original bitstream of the file, making sure it isn’t altered over time, and providing access to the file.

# Slide 8 – What’s in a File?

* To start it is important to understand what is in a file.
* 🖰 At a fundamental level, all digital data is stored as a series of 0s and 1s, as we can see on the left. These are “binary digits” or “bits”.
* These bits are interpreted by the computer to render the information we ultimately see onscreen.
* 🖰 Most files will contain a file header which contains information about the file. This can include the file format and version and information about the contents of the file.
* The column shows an example of information from a file header. Useful metadata can often be extracted automatically from this.
* 🖰 Using the bits and information in the file header (and sometimes a footer) the computer will render the file onscreen, be it as an image, document, spreadsheet etc.

# Slide 9 – What Are the Risks?

* Digital preservation is necessary because the files we keep face a variety of risks. If left alone, they are not likely to survive intact on into the future in the way that physical documents might
* So, what are the key risks faced? They include:
	+ Media obsolescence – this is when storage media, such as tape, floppy disks or CDs become obsolete and you no longer have the hardware needed to read them. Example – most new laptop don’t have a disc drive.
	+ Media failure – storage media is commodity product and tends to have a reasonably short lifespan. Most hard disks tend to have a reliable lifetime of around 5 years. A commonly cited example of media failure is ‘bit rot’ – though all forms of storage media are subject to different forms of decay, bit rot refers to the loss of data due to the small electronic charge of a bit being ‘flipped’ from 1 to 0 or vice versa, or alternatively, this can happen due to cosmic rays or other high energy particles.
	+ Disasters that damage digital data can come in many forms from fire, flood, etc. to human-caused issues such as viruses and malicious attacks.
	+ Finally, file formats themselves can become obsolete as the software they were created in goes out of use (backwards compatibility is not always guaranteed) or the file format itself is no longer used. This is a particular concern for proprietary formats as they are more difficult to reverse engineer.
	+ Worse still, any loss may not be entirely clear to the casual observer. In fact it may require massive manual effort just to work out if any of your data has become damaged. Unless some care is taken to manage and preserve the data properly.

# Slide 10 – What is the Result?

So what happens when these risks bite? The outcome is often unpredictable:

* Media degradation will often lead to a complete failure of the storage device. In other words, you can’t read back any of the data stored on it.
* In some cases, damage might be more subtle. Some of the bits in a bitstream might become lost or damaged. This might lead to an obvious result as in the case of this before (left) and after (right) screenshot of a digitised newspaper page.
* Alternatively, damage might be less difficult to recognise visually. Some of the newspaper pages from the same collection as the one here that were also damaged looked fine until you zoomed in, and they became fuzzy. Although the bitstream was damaged, the viewer software did it’s best to render the image without informing the user. Things are not always as they seem!

# Slide 11 – Stuff Happens

* It’s important to remember that the risks described are always present but also that whenever a digital collection is moved, processes, curate or altered in any way things can and will go wrong!
* This can include:
	+ Network dropouts at critical times, such as in the middle of moving a large number of files
	+ Without careful planning storage disks can get full and any subsequent data copied there can be lost
	+ Bugs in in software can lead to unexpected results, including changes to files and data copied to unknown locations
	+ And the biggest danger is often human error, a simple loss of concentration can sometimes lead to problems like files accidentally deleted.
* It’s also important to remember that these problems can be multiplied many times when doing things at scale!

# Slide 12 – How Do We Solve These Problems?

So how do we solve these problems?

* Keeping more than one copy of the data is essential. 2 is ok, 3 is great. Some organisations store 4 copies. It is also recommended to use more than one form of storage media.
* Note that digital preservation is a trade-off between risk and cost. The more copies the better, but more are costly. There is no precise answer for the perfect number of copies as the sweet spot is likely to depend on your own circumstances.
* Keep one copy in a different geographical location – provides some insurance against natural or unnatural disaster
* Storage media will degrade over time, so be prepared to periodically migrate data to new storage media. Having a refreshment plan is good practice.
* Understanding what you have in your collections is key, even if this is only a list of the number of files, their sizes, types and locations. There are tools to help generate this type of list (sometimes called a manifest).
* Things will still go wrong, so implement a process of integrity (or fixity) checking so you can automatically tell if your bitstreams are still intact.
* Encourage the use of open or stable file formats where you can. Many commonly used formats are open, and their specifications may even be international standards. PDF is a good example of an open standard, but that doesn’t necessarily mean it is the right format for you!
* Carrying out preservation actions will help ensure continued access to your data. Migration and emulation are the most common preservation actions and we will examine them in more detail shortly.

# Slide 13 – Making Sense of a Collection

* Knowing what you have in your collections is an important to first step to understanding them and planning for their preservation, allowing you to understand risks faced and to take appropriate actions.
* This process is commonly referred to as Characterisation in digital preservation and can answer questions such as:
	+ How many files at there?
	+ How big are the files in the collection?
	+ What file formats are included?
	+ Is any of the data dynamic or interactive?
	+ Does it contain personal information?
	+ Are any of the files encrypted?
	+ What risks are associated?
* If you are carrying out this process at any kind of scale, it is obviously useful to be able to automate it rather than having to analyse each file individually.
* Thankfully, there are a number of characterisation tools available for digital preservation.
* Which one you choose to use will depend on the type of collections and systems you have.

# Slide 14 – Characterisation Tools

* One of the most commonly used characterisation tools is DROID, developed by The National Archives in the United Kingdom.
* DROID produces detailed reports on the files in a particular folder of group of folders by harvesting information such at the file extension or from the file header and comparing this with the Pronom file format registry. Pronom contains information on a large number of file formats and their behaviours.
* DROID is just one of the many characterisation tools available, others include C3PO, JHOVE, Apache TIKKA and FITS.

# Slide 15 – Trust No One

* Whilst characterisation is a very useful process it is essential to note that these tools are not infallible and it is important to validate their data through quality control.
* 🖰 One way this can be done is using validation tools, a number of these are available. A recent example is a PDF validator produced by a project called VeraPDF.

# Slide 16 – What is a “checksum” or “hash value”

* Another process we mentioned that is core to digital preservation is integrity checking. This is a way to check that files remain unchanged over time using values called checksums or hash values.
* To perform integrity checking we need to use a simple technology called checksums or hashes. This animation shows how they work.

Let’s say we have a bitstream, or digital file, that we want to preserve.

1. 🖰 We begin by creating a checksum. This is simply a fairly unique short number derived from the file using a software tool.
2. 🖰 Think of a checksum as finger print.
3. 🖰 At some point in the future, we want to verify that our file remains exactly as it was, back when we first created the checksum at the top of the screen.
4. 🖰 We generate a new checksum from the file.
5. 🖰 We then compare the new checksum with the old one.
6. 🖰 In this case, the checksums are identical, so we know the file is undamaged and exactly as it was.
7. 🖰 However, if the file had become damaged, the checksum we would generate from it would be different.
8. 🖰 On comparing the two checksums, we can see that they are different from each other. This confirms that our file is no longer identical to how it was. Perhaps it has been damaged by media failure or “bit rot”.

# Slide 17 – Combined Strategies

Integrity checking can be combined with the strategy of keeping multiple copies of each file to provide a robust digital preservation approach. This works as follows:

1. 🖰 We make 3 copies of the file we want to preserve, ideally placing one file offsite to protect against natural disasters
2. 🖰 We generate checksums from each file, and we can see that they are all the same, and each file is good.
3. 🖰 Over time we can then recalculate our checksums, and see that the three copies of the file are still exactly as they were
4. 🖰 Until at some point in the future we recalculate our checksums and discover that one of them is different!
5. 🖰 Straight away we know that the middle copy has become damaged
6. 🖰 So we then discard the damaged file
7. 🖰 And replace it with a copy of one of the others

Using these techniques, we can dramatically reduce the chance of losing any of our data.

# Slide 18 – Integrity Checking Tools

Some characterisation tools include functionality for integrity checking but there are also software tools specifically for this process. These include:

* The tool “Fixity” is a good place to start for generating checksums.
* ACE is a more advanced tool, ideal for performing scheduled integrity checks.
* More options can be found on COPTR, a registry of digital preservation tools.

# Slide 19 – Approaches to Preservation

* The processes described in the last few slides combine to form one of the main form of preservation, often known as bit-level preservation.
* As mentioned earlier, this addresses the issue of preserving the original bit stream, but we must also address how to preserve access to that bit stream.
* There are several different approaches to this issue, including preserving original hardware (sometimes known as the computer museum approach), also using digital forensic techniques to uncover ‘lost’ data via digital archaeology.
* The most common approaches, however, are migration and emulation.

# Slide 20 – Migration

* There are two main forms of migration for digital preservation, and it is possible to use one or both.
* The first is a method often referred to as ‘normalisation’. This is where all files of a particular type (for example, text documents) are ‘normalised’ to one file format.
* The example on the slide shows Word documents being normalised to PDF. For images this could be JPEGs and GIFs normalised to TIFFs.
* The choice of normalised files format used will depend on the needs of the organisation and its users.
* The second method involves migrating old file formats to newer versions when they are at risk of becoming obsolete.
* This could be migrating an old .xls spreadsheet to a newer .xlsx format.
* Both methods have their positives and negatives:
* Normalisation creates homogenous, easier to manage collections and means that users need to know how to use fewer files types.
* Migrating to new versions means that files can be accessed in current computer environments.
* Both processes can be automated but quality control is incredibly important and careful consideration must be given to migration pathways to avoid loss of data and functionality.

# Slide 21 – Emulation

* Emulation is the process of recreating the original environment in which a file was created and used via a layer of specially written software: the emulator.
* 🖰 Emulation has been particularly successful in the world of computer games, where enthusiasts will create emulators to allow them to play older games.
* 🖰 It is also an increasingly popular preservation method and several projects have produced emulators for everything from early browsers to old versions of PowerPoint. Many of these emulators are freely available, either online or as software downloads.
* Emulation perhaps seems like the ideal version of digital preservation as it allows users to access the files in their originally environment, providing a more authentic experience.
* It is however, very resource intensive and emulators will require updates (or their own emulators) as computer environments change.
* It can also be difficult to confirm the emulator truly captures the original environment unless there is still access to an original example to compare.

# Slide 22 – Questions

Possible discussion points:

* Is there material in your collections created in different computing environments?
* What type of storage do you currently use?
* Do you think migration or emulation sounds like a better solution?