Digital Preservation for Community Archives

Version 1.4 JANUARY 2018

This guidance was produced for the Community Archives and Heritage Group by staff and students of the Department of Information Studies at University College London including Alexandra Eveleigh, Anna Sexton, Sophie Denman and Andrew Flinn. Thank you to all those others who kindly read and commented on the guidance.

January 2018.

How to use this guide

The aim of this guide is to provide a practical introduction to digital preservation for community archives.

The first section of the guide (Introduction to Digital Presentation) introduces the subject of digital preservation. It is broken into 'frequently asked questions' and covers the following questions:

- What is digital preservation?
- Digital preservation or digital curation?
- What are the characteristics of a digital object?
- What are the preservation risks for digital objects?
- What is necessary to preserve digital objects over time?
- What are the principal strategies for digital preservation?
- What is a digital repository? What are its basic components? How are they modelled?
- Should we develop our capacities for digital preservation in-house or look to partner with others?

The second section (Practical steps towards Digital Preservation) of this guide outlines practical steps towards digital preservation. This section is divided into an overview of steps towards digital preservation, followed by more detailed guidance on how to carry out those steps, and links to more advanced preservation guidance.

- The overview of steps towards digital preservation is broken down into:
    - Knowing your holdings
    - Aiming for prompt check in on receipt
    - Creating file lists of your digital holdings
    - Making copies of your digital objects
    - Planning where and how you will store your digital objects
    - Creating checksums of your digital objects so you can perform integrity checks over time
    - Refreshing your storage media
    - Keeping track of file accessibility
    - Always working from copies and documenting your actions

- The more detailed preservation guidance is broken down into:
    - Understanding your digital holdings
    - Assessing priorities for action in existing holdings
    - Removing legacy storage media from existing collections
    - Creating a dedicated workstation for check-in (capture)
    - Developing a workflow for preparing digital objects for check-in (capture)

- Developing procedures for gathering relevant information on handover of digital material from donors
- Selecting what to preserve
- Using third party data recovery services

The next section (More Advanced Guidance on Preservation) provides links to external sources which provide further advice on digital preservation. It includes some starting points for exploring more complex digital repository set-ups that move beyond just preserving the bits, to rendering digital objects as information objects in accessible forms to users.

The final section (Glossary of Terms) offers definitions of some key digital preservation terms.

Introduction to Digital Preservation

What is digital preservation?

Digital Preservation is about maintaining digital objects 'for as long as required, in a form which is authentic, and accessible to users'.[1]

Within your community archive you are likely to hold a range of digital objects that need preserving, including:

- Born digital objects
- Digitized surrogates

The term 'born digital' is a catch-all used to describe objects that are created in a digital format and exist only in an electronic form. Common examples of born digital objects are photographs taken with a digital camera, an email or a text document. The term 'digitized surrogates' relates to the outputs from the process of digitization when analogue material is reproduced in a digital form. For example, a digital scan of a printed photograph, or a WAV file created from reformatting the sound on a cassette tape are both digitized surrogates of originally non-digital items.

This guide provides useful starting points for managing both born digital objects and digitized surrogates effectively over time. It does not aim to cover the process of digitization. For useful starting points in preparing and managing a digitization project see:

Advice for community archives from Hampshire Record Office:
http://www3.hants.gov.uk/archives/community-archives/how-to/digitisation.htm

Advice from The British Library:
http://blogs.bl.uk/collectioncare/2013/11/fail-to-prepare-for-digitisation-prepare-to-fail-at-digitising.html

For more in-depth advice on digitization which also covers digitizing audio and video material, see the guide produced by the Canadian Information Network:
http://canada.pch.gc.ca/eng/1445531744547

[1] Adrian Brown, *Practical Digital Preservation: A How to Guide for Organizations of Any Size*, 2013, Facet Publishing, p.3.

Digital preservation or digital curation?

In relation to managing digital objects, some people prefer the term digital curation (or even digital stewardship) to digital preservation, because 'curation' and 'stewardship' implies active and ongoing intervention by those responsible for looking after the material. Whatever term is used, the key is to view the management of digital content as an active, regular and continued process of intervention. Due to their unique qualities, digital objects cannot be 'stored and ignored'.[2]

What are the characteristics of a digital object?

In this guide, 'digital object' is used as an overarching term to describe both born digital or digitized objects. Digital objects come in a wide variety of forms including text, images, social media outputs, spreadsheets, emails, websites, games, audio files and films to name a few. They can range in size and complexity, from an entire web site to an individual tweet. They can derive from a variety of sources, from laptops to desktops to smart phones, and they can be stored on a variety of media, including hard drives, CDs and memory sticks.

There are five core characteristics of digital objects:

1. Digital objects depend on hardware and software
2. When thinking about digital objects there is a distinction between data objects and their rendered form as information objects
3. The end focus of digital preservation is on the information object and the rendering of its significant properties
4. Multiple data objects can represent the same information object
5. The same data object can exist in multiple places

Digital objects depend on hardware and software

A core characteristic and unifying feature of all digital objects is their machine-dependency. A digital object can only be accessed and rendered in a decipherable and useable form through a combination of inter-dependent hardware and software.

In technical terms, a digital object is comprised of a series of zeros and ones commonly known as binary digits, bits or the bitstream. This data (the bitstream) is not directly decipherable by humans. In fact, it is meaningless until it is rendered into recognizable information by software capable of correctly interpreting the bitstream. This interpretation by software is governed by the object's file format. Particular pieces of software are designed to be able to recognize particular file formats interpret and render the underlying bitstream in accordance with the format specification. The software that renders the digital object is itself dependent on having a compatible operating system, hardware platform and possibly other software in order to run. This cycle of dependency can be widened when you

[2] Mike Ashenfelder, 'The Library of Congress and Personal Digital Archiving'. Available at: http://digitalpreservation.gov/documents/lc-digital-preservation.pdf?loclr=blogsig

consider what technical equipment might be necessary to access the bitstream in the first place from the removable storage media where it is encoded.

This machine-dependency has important implications for digital preservation. It is necessary to keep a handle on the chain of inter-dependencies between the hardware and software necessary for rendering a particular digital object, and to consider how accessible the digital object will be over time as software and hardware platforms evolve.

When thinking about digital objects there is a distinction between data objects and their rendered form as information objects

It is useful to make some conceptual distinctions when we are talking about rendering digital objects into a meaningful form. It is useful to think of the digital object's bitstream as a data object that has to be rendered through a combination of hardware and software into a meaningful information object (something that humans can use and understand). The combination of hardware and software needed to render the data object into an information object is commonly referred to as its representation network.

The end focus of digital preservation is on the information object, and the rendering of its significant properties

Within digital preservation, it is the preservation of the information object (its rendered form) that is the end focus. It is possible to change the processes used (the combination of hardware and software) to render the data object (the bitstream) into an information object (something that humans can use and understand), provided that the significant properties of the information object remain intact. Defining and understanding the significant properties of the information object (what it is that we do not want to alter or lose) is important in guiding the process of digital preservation. However, deciding on what those significant properties are can be problematic, as the significance of the properties will always be dependent on the context of use.

Multiple digital objects can represent the same information object

Leading on from this distinction between the data object and the information object, a core characteristic of digital information is that the same information object can be represented by more than one data object.
This can be illustrated by thinking about a digital photograph as an information object that might exist in different formats, for example as a TIFF and as a JPEG. In both cases, the information object, the digital photograph, is the same but it is represented by two different data objects each with entirely different bitstreams. The interpretation of the bitstreams in the TIFF and JPEG files by appropriate hardware and software leads to the same array of colour pixels on the screen (i.e. they represent the same information object) despite being entirely unique and separate data objects. It is common for the different data objects representing an information object to be described as manifestations or representations of an information object.[3]

[3] This example has been taken from Brown, p.196.

The same digital object can exist in multiple places

Each time we copy a digital object, through dissemination or back-up, we are making an exact copy of the data object. This means that exactly the same data object can exist in multiple places. This ease of replication is very useful for digital preservation because if one version of the data object degrades or is lost, we can have others that can take its place without concern that we have in some way lost an 'original'. The ease of replication also presents a challenge when digital objects are in active use, and are being multiplied, edited and changed by multiple people in multiple places, it is sometimes difficult to unravel the relationships between these versions when making decisions on what to preserve.

What are the preservation risks for digital objects?

Digital preservation risks can be broken down into:

- Risks to the preservation of the data object (the bits)
- Risks to the rendering of the data object into a meaningful information object

Risks to the preservation of the data object (the bits)

In order to successfully keep digital objects over time the bitstream needs to be preserved. Terms like 'bit rot' are sometimes used to describe the process through which one or more of the zeros and ones in the bitstream 'flip' or lose their value leading to the possibility that the entire bitstream becomes unreadable. There are various threats to maintaining the bitstream intact which need to be taken into account when planning digital preservation activities.

Storage media decay or damage
Removable storage media such as CDs, DVDs, hard drives and memory sticks can decay over time and result in corrupted files. Although the physical medium on which the bitstream is encoded may deteriorate gradually, in most cases any accompanying deterioration of the data is a more extreme move from 'readable' to 'unreadable' without an in-between stage. Other forms of storage hardware such as servers are also susceptible to failure, which may result in data loss.

Storage media obsolescence
Computer technology changes quickly, and commonly used storage media can become obsolete. Internal floppy disc drives, for example, have not been a feature on new computers for some time, now making access to the objects on them more difficult. As computing moves on, there is a risk that there will no longer be the necessary hardware or software readily available to provide access to the bitstream on any given storage media.

Network/Service/Hardware/software failures
When the digital object is transmitted over networks and services in hardware and software interactions, any failures may give rise to data loss. For example, if software that is carrying a virus is used to access the bitstream, then the result can be data loss.

Human error or malicious intent
The bitstream can easily be changed, deleted accidently, or without realization due to human error. Data can also be deliberately altered and corrupted through malicious intent.

Disasters
As with physical objects, natural disasters such as fire, flood, or earthquake can have a serious impact on data survival.

Risks to the rendering of the data object into a meaningful information object

File format/software obsolescence
Rendering the bitstream into an understandable and useable information object requires software designed to interpret the bitstream in accordance with an underlying file format. Over time, file formats tend to evolve and change, and therefore so do the software applications that interact with them. Although unusual for well-known file formats, less well used file formats may become obsolete over time as the software that renders them is no longer supported. Therefore, it is possible to have successfully preserved the data object (the bitstream) but lack the accompanying means to render it as an information object.

Maintaining the ability to trust and understand the information object
Over time, digital objects will be subject to preservation actions that need to be taken in order to ensure that the bitstream is preserved and can be rendered as an information object. For example, managing the digital object might involve migration to a new file format (see next section) to ensure it can still be successfully rendered as an information object. If these actions are not tracked and recorded it is difficult to trust the integrity or authenticity of the rendered information object. Documentation helps future users to understand the changes that have been made to the digital object and make sense of what they are seeing and using.

The meaning of a digital object may be dependent on additional information that may have been easily and readily accessible when it was originally created and used, but is less clear when revisited at a later date away from the originating context. Capturing relevant contextual information alongside a digital object can be vital to ensuring its ongoing usefulness. This might be 'as simple as capturing the units of measurement used within a spreadsheet, the scale of a map, or the point of origin within a CAD drawing'.[4]

Creating adequate metadata (data about data) for digital objects mitigates a lot of the threats to its ability to be understood and future use. This data can be broken down into administrative, technical and descriptive data which when combined provides the key to being able to make sense of and use the digital object in the future.

[4] Digital Preservation Coalition, *Digital Preservation Handbook*, 2nd Edition, 2015. Available from: http://handbook.dpconline.org/

What are the principle strategies for digital preservation?

The first requirement of digital preservation is to be able to have processes for what is often called 'bitstream preservation' – that is to have the ability to maintain an intact copy of a digital object over time. The advice given in this guidance is orientated around achieving this. This focuses on processes for maintaining the object's zeros and ones, including:

- Introducing redundancy by keeping several copies of each digital object
- Ensuring that copies are stored on different storage technology in different locations
- Regularly monitoring and refreshing storage media
- Introducing technical processes that enable you to check that a digital object has not been altered
- Performing virus checks
- Having mechanisms for documenting all of these processes

Taking the time to develop processes that ensure bitstream preservation should be the priority. These processes are not expensive to implement, and they do not require a high level of technical expertise. Whilst it requires time to think through these processes and come up with a workflow that fits your community archive, once established, these processes should not feel overwhelmingly time consuming. Whilst bitstream preservation does not in itself solve the ongoing issues around successful rendering of data objects into information objects, it does ensure survival of your data, and is the foundation on which ongoing access and use can then be built.

Beyond bit stream preservation of data objects is the ability to render the information object to users in a manner which keeps its significant properties intact. This is often referred to as 'logical preservation'. It relies on having systems in place for bitstream preservation, but goes much further in developing processes for ensuring authentic rendering of information objects. Strategies for doing this are often described under three categories:

- Computer museum
- Migration
- Emulation

Computer Museum

One way of ensuring that a data object can always be rendered as an information object is to keep and maintain its original hardware and software environment. This has the obvious advantage that the information object will be rendered in the manner of its original context of use. However, this method is impractical and mostly out of reach of all but the most specialist computer museums. It requires large amounts of space for legacy equipment and high degrees of technical skill in maintaining the equipment and sourcing spare parts.

Migration

The second option for managing the threat of technology obsolescence relates to mitigation against file format and software obsolescence and involves intervening at a given point to migrate the data object to a new file format. This relies on having the software tools that can convert digital objects from one format to another, creating an entirely new digital object with a unique bitstream. Issues with migration lie in the transfer of significant

properties in the new rendering of the information object. The migration process may not be capable of transferring all of the unique properties of the original object which can result in rendering errors. The look and feel of the information object may be compromised, and sometimes the process of migration may require multiple transformations into intermediate formats.

There are three points at which migration commonly occurs:

- Immediate migration on entry to the repository. This is often called normalization, and involves identifying a range of preferred file formats perceived to have longevity, and converting all files to these preferred file formats from the outset.
- Migration on obsolescence involves leaving the process of migration until it is needed, the difficulty being keeping a watch on technology obsolescence and not leaving it too late.
- Migration on access is the middle ground, where the migration process is performed when the file is needed.

Many archive institutions initially developed logical preservation strategies that included proactive normalization. In the early days of digital preservation, the threat of file format obsolescence was perceived to be greater than it has turned out to be in practice. Now, even institutions with very sophisticated digital preservation strategies will often leave migration until it is needed, so as not to waste up-front time and effort.

Emulation
Emulation involves keeping the data object in its original form and developing software that can recreate the functionality of obsolete technology on a contemporary platform. Emulators are expensive and complicated to create and it is beyond the reach of most institutions to develop these in-house.

Of the three options, migration therefore offers the most doable method of ongoing logical preservation for most archive repositories.

What is a digital repository? What are its basic components? How are they modelled?

Developing a digital repository does not have to be a complex or expensive endeavour. A digital repository is simply a term to describe the 'combination of people, processes and technologies which together provide the means to capture, preserve and provide access to digital objects'.[5]

There is a model known as the Open Archival Information System (OAIS) reference model which is widely used as a means to visualize what a digital repository is and what it does.[6] Although the model itself is relatively complex and somewhat difficult to follow, the concepts underneath are actually very straightforward. An organization developing a digital

[5] Brown, p.15.
[6] For a fuller description of OAIS see: http://www.oclc.org/research/publications/library/2000/lavoie-oais.html

repository must have the means to acquire, ingest and control new digital content. It must be able to perform preservation and management activities on that content in a way that ensures the content is reliable and useable, and it must be able to make that content available again to its designated community. The underlying processes that the digital repository supports can therefore be summarized in the following terms:

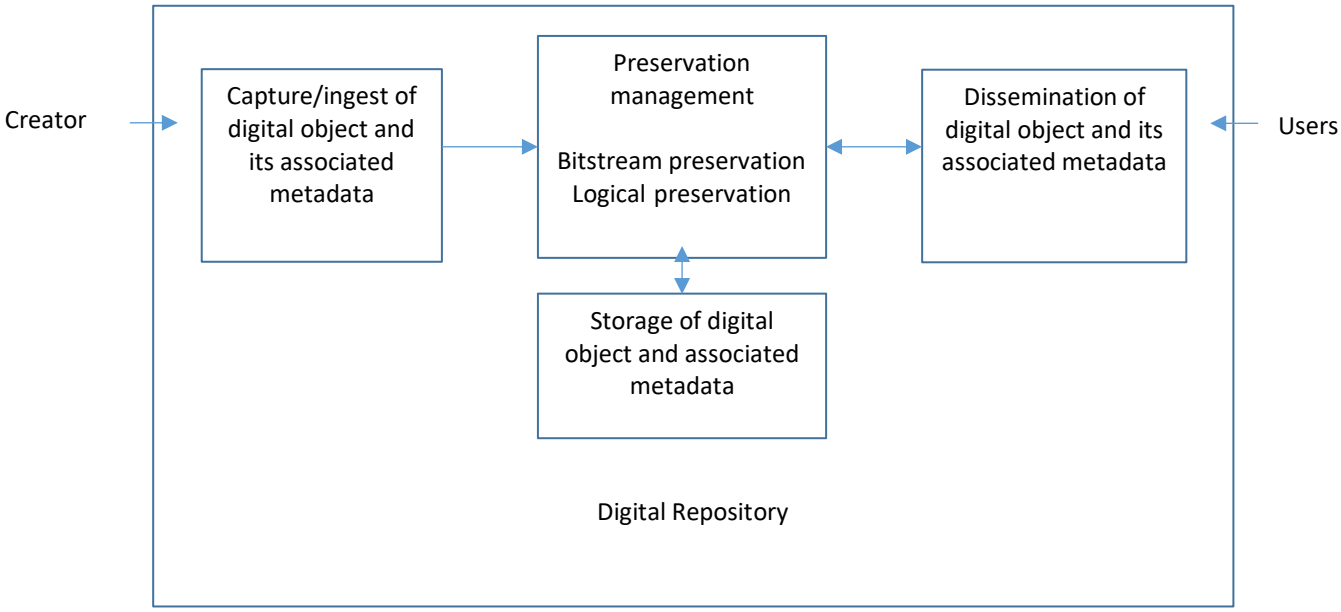


Figure 1: Functions of a digital repository.[7]

In a basic digital repository, the functions that must be provided are therefore:

- Capture (ingest) of new digital objects – this can be achieved with a basic workstation set-up, adopting a relatively straightforward workflow that utilizes a variety of free tools
- Metadata management – descriptive metadata about the digital objects can be incorporated into the existing system that you use for cataloguing your collections, and technical metadata can be generated using free tools and stored as text files in simple spreadsheets alongside the content or in another designated area
- Preservation management – this can be manually instigated as and when it is required using a range of free tools (this guidance focuses on processes associated with bitstream preservation)
- Storage – this relates to storing multiple copies of each digital object using a combination of different storage technologies depending on what is available
- Access – this can be provided via a designated terminal or remotely via removable media

Building a digital repository along these lines enables a flexible and customizable approach. It is suitable for dealing with small volumes of digital objects. It can provide all the elements necessary to enable bit stream preservation and can enable you to incorporate the development of logical preservation strategies, such as file format characterization and format migration in time as expertise grows.

[7] Based on diagram in Brown, p.16.

Should we look to develop our capacities in-house or partner with others?

Collaborating with other organisations can be a useful means of pooling expertise and resources. The decision on how far to explore partnership opportunities will be dependent on your ethos as a community archive.

Some local authority archive services, such as Hampshire Record Office, are willing to offer free advice and guidance to community archives in their area (see http://www3.hants.gov.uk/archives/community-archives/how-to.htm).

If you are interested in exploring partnership opportunities, utilizing the CAHG network to find other archives in your area is a useful place to start.

Practical steps towards Digital Preservation

'Bitstream' Digital Preservation Overview

This section aims to give a basic overview of good digital preservation practice which will help you to achieve bitstream preservation of your digital objects.

Know your digital holdings:
Aim to develop and maintain an overview of your digital holdings (e.g a digital asset register) so that you can keep track of what digital content you have got, where it is, and what preservation issues arise from the material

Aim for prompt check-in on receipt:

- When you receive incoming digital, material aim to deal with it promptly and do not let removable storage media (CDs, floppy discs, memory sticks etc.) creep into your collections undocumented and unprocessed. Acting early with digital content is key to its survival.
- When taking digital material off from the storage media it came in on to bring it into your digital holdings, scan it for viruses and malware to make sure there are no unwanted surprises with the material. If you can develop the set up and capacity to do so, consider keeping the material 'in quarantine' until you have conducted two virus scans on it (one month apart)
- When you take in new material, open the files (or if lots of files coming in at once open a random selection) to check that they are readable.
- Check with the donor that what you have got tallies with what they expected to give you, and having reviewed the material (including for possible sensitive and personal data), make a final decision on whether you consider the material worthy of inclusion in your community archive and therefore worthy of capture and ongoing preservation.

Create file lists of your digital holdings:
Knowing what you have got is key to being able to manage and preserve your digital holdings. You can automatically generate file lists for incoming material using free tools (see below).

Make copies of your digital objects:

- Always create more than one copy of your digital objects. Take advantage of the ease at which digital objects can be replicated and aim to keep two or three copies of each digital object that you want to include in your holdings.
- Aim to document and keep track of the relationships between your digital objects (i.e. which ones came from a particular donor or relate to a particular collection).

Plan out where and how you will store your digital copies:
Think about the different types and combination of storage technology you will use to keep your copies. Your choice will depend on your technical infrastructure and what is available to

you but the principle is to avoid 'keeping all of your digital eggs in one basket'.[8] Therefore, in general, you should aim to:

- Keep (at least) one of your copies easily accessible (i.e. on a non-removable disk). This enables you to revisit your material easily to perform integrity checks (see below).
- Spread the risk (of theft and disaster) by having (at least) one of your copies stored in a different geographical location.
- Spread the risk of technology failure by having your copies on different types of storage technology (if outsourcing storage do not rely on a single vendor to store all your data).
- Avoid storing copies (whether on hard drive, server, or removable storage media) in locations with fluctuating temperature and humidity, and keep the storage technology away from sun, salt, and water etc.
- Make sure your copies are uncompressed and unencrypted. Compressed back-ups are not exact copies of the digital object and are therefore not suitable as digital preservation copies.

Create checksums on your digital copies so you can perform integrity checks over time:
A checksum is the generation of a value that relates to a block of digital data (such as a file) which can be used as its 'digital finger print'. Creating a checksum enables you to undertake integrity checks on your digital holdings over time (see below).

Refresh your storage media:
Storage media has a relatively short shelf-life and needs to be regularly refreshed. Aim to refresh the storage media you are using every five years to avoid data loss. Refreshing means copying all files across onto new media (e.g. from an old CD onto a new CD). Once copied, carry out integrity checks on at least a proportion of the files.

Keep track of file accessibility:
When you get new hardware/software, check that your existing files are still accessible.

Always work from copies and document your actions:
If you perform any actions on your digital objects always work from a copy and aim to document your actions so that others can learn what has been done.

Further details on creating file lists

There are various free tools available that can automatically generate a file list (sometimes called a file manifestation) for incoming digital content, which can generate information on things like file name, file format and size as well as additional useful metadata such as the date of last access.

[8] Relying on a cloud-based services for storage of digital objects is becoming increasingly common. The National Archives (UK) have produced guidance on using cloud-based storage which explores the main benefits and risks of doing so as part of a digital preservation strategy: http://www.nationalarchives.gov.uk/documents/archives/Preserving-Digital-CloudStorage-Guidance_March-2015.pdf; *Digital Preservation Coalition Handbook*.

You will need to think about how you integrate this information with your existing descriptions of collections, and where this information will be stored. A simple solution is to hold this information in a spreadsheet or text file alongside your digital copies, making reference to its existence in your main catalogue.

Karen's Directory Printer and DROID

Karen's Directory Printer is a free, downloadable software tool which enables you to generate or print file lists of your digital files. These lists comprise information about the properties and characteristics of your digital files, such as the file name, file size, file format and version, and date of last modification. The file lists therefore provide you with a directory of information about the files in your digital repository. Having a clear idea of the types of digital objects in your repository will enable you to better manage them in the short-term and plan for their long-term preservation.

DROID stands for Digital Record Object Identification, a file profiling tool developed by The National Archives (UK), and it has a similar function to Karen's Directory Printer. Running your digital files through DROID will result in the generation of a report which includes information such as the file name, file format and version, file size, and date of last modification. Like Karen's Directory Printer, the report generated by DROID will provide you with a directory of information about the files in your digital repository which can be used to plan effective short and long-term preservation.

Both Karen's Directory Printer and DROID have the option to generate a checksum for each digital object which is run through their software. A checksum is a string of characters which are unique to each individual digital object, and when generated, act as its unique signature or fingerprint. Checksums will remain the same for each digital object unless it degrades over time or is altered in any way. They are therefore a useful tool for checking the integrity of your digital objects over time. Every time you re-run your digital objects through the same software you will be provided with a checksum; if the checksum remains the same, you know that your digital objects have not degraded or been altered and therefore retain their integrity. If the checksum has changed, you know that their integrity has broken down. Neither Karen's Directory Printer or DROID will identify what it is about a digital object that has changed, but rather, they alert you to the fact that something has changed, allowing you the opportunity to revise your digital materials for any preservation issues.

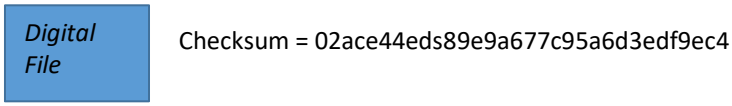
The usefulness of these tools lies in their ability to generate the information to provide you with a clear at-a-glance understanding of the numbers and formats of digital objects in your repository. By having this understanding, you can better manage and preserve their current state, and effectively plan future preservation actions.

Karen's Directory Printer is available to download from:
http://www.softpedia.com/get/System/File-Management/Karen-Directory-Printer.shtml

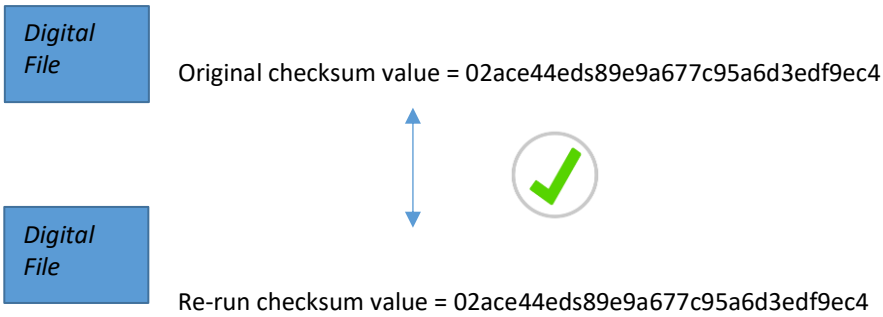
DROID is available to download from:
http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

Further details on generating checksums and performing integrity (or fixity checks) over time

As outlined above a checksum is a unique string of characters that relates to a digital object. A checksum can act as the object's 'digital finger print'. Checksums can be easily generated using a simple software tool.

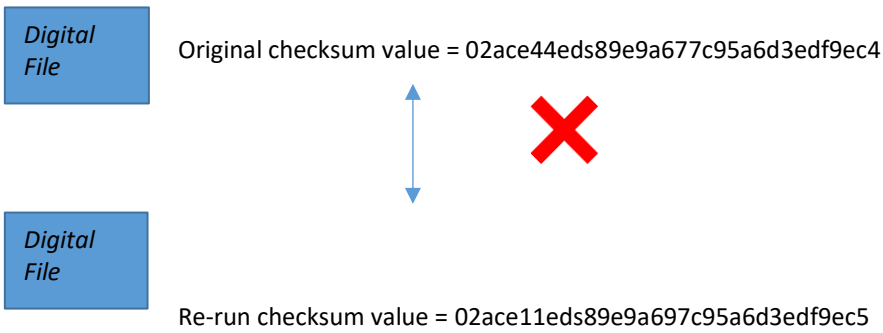


To verify that a digital object has not been altered over time (sometimes described as checking its integrity or fixity) it is possible to re-run the checksum generation at a given point in time and compare the checksum values. If they are the same before and after the test, you know the bits in your file remain intact.



The frequency at which you should re-run checksum generation as a means of integrity/fixity checking will depend on your capacity for doing so. The Digital Preservation Coalition recommends doing this every 6 months.

If you re-run checksum generation and find that the checksums for your digital object do not match, you know your digital file has lost its integrity.



If you have followed this preservation guidance you will have other copies that you can rely on. You can perform checksum tests on one of your other copies, and when you (hopefully)

find that this is unaltered, you can replace the damaged digital object with a copy of the unaltered back-up.[9]

File manifestation and format identification tools such as Karen's Directory Printer and DROID have the capability to generate checksums. There are also free standalone tools for checksum generation tools which are widely available. For example Fixity (https://www.weareavp.com/products/fixity/)

Whenever you generate a checksum you need a means of recording it and linking it to the digital object. It is also a good idea to develop mechanisms for documenting your integrity checking to help you keep track of what is being done.

Detailed Preservation Guidance

This section aims to go through steps towards digital preservation in greater detail. It aims to describe in practical terms how to go about achieving bitstream preservation in a minimal digital repository set-up, and introduces a few extra processes and tools that take you beyond the advice outlined in the previous section.

Knowing your digital holdings

Starting out in digital preservation involves developing processes for effectively managing the digital content that you already have, as well as beginning to develop processes for dealing with incoming digital objects in the future.

The very first step towards digital preservation is therefore assessing what digital objects you have already got, including those languishing on removable storage media (floppy discs, CDs, memory sticks etc.) that have been left untouched amongst your collections. It is useful to see this as part of a broader review of your collections and your constitution as a community archive. Key broad questions you should seek to ask at the beginning are:

- How much of our collection is digital? How rapidly are our digital holdings growing?
- Who has responsibility for managing our digital objects at the moment?
- How much of our digital material is currently unprocessed (on storage media that we haven't properly accessed, documented etc.)?
- What level of resource and expertise do we have available in relation to managing our digital holdings?

[9] Diagrams used adapted from: Digital Preservation Coalition 'Just Keep the Bits' video: https://vimeo.com/169241520

- What infrastructure do we have or are likely to be able to obtain to help us manage our digital objects?

In order to get a handle on your digital holdings it is useful to start to compile an initial register that can act as an inventory of what you have. This can be in the form of basic tabulated information (held in a document or database). The table below provides an example which you may like to adapt and follow:

Level (e.g. collection/ stand-alone object)	Donor/Source	Description	Storage media/location	Object category (digitized image(s), office doc(s)etc.)	File Formats & Extensions	Size (No. of GB)	Additional info

Where you have material on removable storage media (e.g. on memory sticks, floppy discs, CDs, removable hard discs etc.) that has remained unprocessed, you might consider waiting before you try to access and read the files from the storage media so that you can develop and plan out the best workstation environment for doing so. Instead, note down the details of the storage media and its location. You can do some more detailed investigative work later when you have established the basic elements of your digital repository and have a workstation set-up to capture incoming objects (see next section). At this point it is also best to leave the digital material in situ.

Much useful information about what you have will come from looking through the collection itself, but it may also come from any records you have on donations, along with any catalogues and other resources that describe your holdings. Use your register to record anything that is known about what is on the storage media (including any information on the hardware, operating systems, or software used to create the files).
The register may seem light on details to begin with, but you can view it as an active document that you up-date and add to as your digital repository and digital preservation activities develop. Your register, if kept up to date, will be a vital management tool in helping you keep track of your digital holdings and any related preservation issues that need to be addressed.

Assessing priorities for action in existing holdings

You can use the information you have compiled in the register to help guide you in some initial thinking around what to do about unprocessed storage media, and what the priorities for retrospective preservation actions should be. You may discover, for example, that you have more material than you thought on storage media that is now obsolete (e.g. floppy discs), or material on aging CDs that are in danger of becoming unreadable. In this case you will have to start thinking about what you will tackle first. Here you are trying to weigh up the value of the material against the ease at which you will probably be able to take action,

how pressing it is that you act quickly, and the relative costs of doing so. Therefore, key factors for consideration are:

- The relative value of the collection the digital material sits within, and anticipated levels of use
- The relative danger of loss of content due to threat of obsolescence, poor condition or age
- Whether the digital content is unique within the collection or replicated in the analogue materials (you may not know this at this stage if you have not accessed the digital files on the storage media)
- Whether or not it is likely that you will be able to access and process the digital content in-house or whether you will have to use a specialist service to retrieve the files. In the case of the latter, the larger cost implications will need to be factored into deciding whether it is worth trying to retrieve the material. (This is a particularly difficult judgment call to make in cases where information about what is on the storage media is non-existent or sparse).

It is useful to document your initial assessment of priorities. However, being in a position to take any action on dealing with legacy storage media in your collections depends on progressing with the establishment of an adequate set up for capturing digital objects into your digital repository (see next section).

Removing legacy digital storage media from existing collections

When dealing with legacy storage media in existing collections, you may want to develop a process to document the removal of the media from the physical collection. Taking note of the physical characteristics of the storage media is useful if you intend to destroy the carrier once the digital objects residing on it have been captured. Therefore, you may find the following system developed by the Hull History Centre useful:

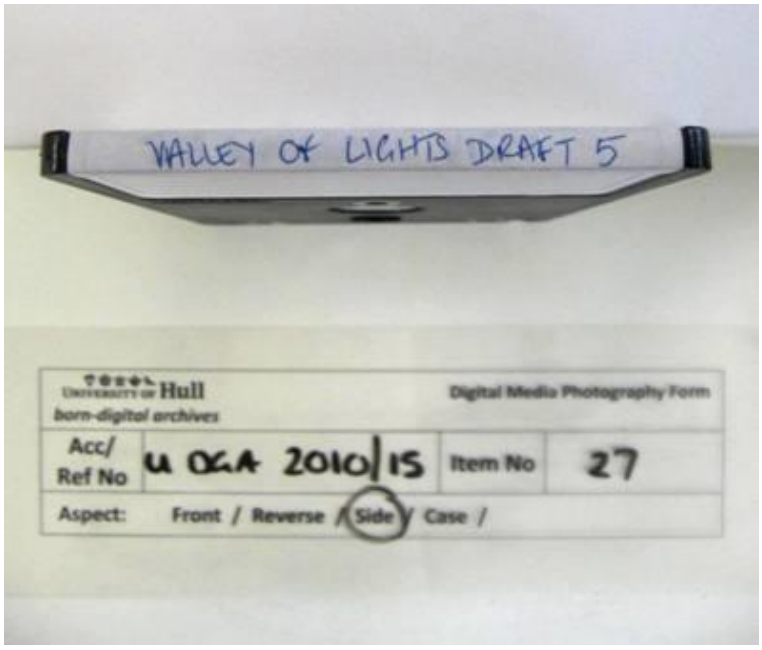
1. Create a generic template form which you can use as a backdrop for photographing legacy storage media, along the lines of the following:

[Insert name of Archive]	Digital Media Photography Form		
Collection Ref No:		Unique identifier for item:	
Aspect photograph taken from: Front / Reverse / Side / Case			

In each instance:

2. Remove the storage media (i.e. floppy disk, CD, memory stick etc.) from its current location in the collection
3. Fill in the details on the form for the media
4. Photograph the storage media placed on the form (from various aspects if information appears on back, front and sides)
5. Store a printout of the photos in the collection at the place where the digital media was originally located

6. Store all your digital photos for future reference, as a form of collection documentation



Example of photographic form reproduced with kind permission from Hull History Centre

Having documented the physicality of the storage media, you may then decide to destroy the media once the files on it are safely captured and held within your digital repository.

Creating a dedicated workstation for check in (capture)

The type of workstation you can create for checking-in digital objects and preparing them for capture and storage in your digital repository will depend on levels of resource available, level of need (i.e. how much digital material you take in), and levels of in-house expertise for creating a more complex and specialist set up.

It is useful to have a dedicated 'clean' computer for accessing and preparing digital media prior to capture. In order to perform quarantine and virus checking, the computer should have up-to-date virus software installed to do regular scans. It should be offline so as not to introduce viruses, and it should not be used for other work or be connected to a network that might be affected by viruses introduced when accessing media.

The table below gives details of a workstation set-up that will enable you to perform the digital preservation processes outlined in the overview in the previous section (in light orange) with further additional options (in dark orange) that will enable you to undertake a wider range of useful preservation actions:

Hardware/Software Components	Preservation processes	Additional notes
Dedicated/isolated off-line computer installed with (for example) Windows, MS Office, Quickview Plus (file viewer) and up to date virus software.	Enables: Quarantine of incoming objects to prevent spread of viruses The capture of a good range of file formats	Quickview Plus (File viewer) costs £40 and is worth the investment as it

For details of Hull History Centre's hardware/software specifications for their workstations (which may act as a useful guide) see: http://www.hullhistorycentre.org.uk/discover/pdf/Idiot%27s%20Guide%201%20-%20Forensic%20Workstations.pdf	The range of legacy storage media that the PC will enable capture from will depend on what internal drives it has or what drives can be purchased and connected to it (see below).	enables you to read most file types without the need to purchase numerous software programs. http://www.avantstar.com/quick-view-plus-standard-edition#fndtn-overview
Checksum generation tool	A checksum is a value derived from a block of digital data that acts as a 'digital fingerprint' which enables you to compare whether that digital data has altered as a result of transmission. Generating checksums on captured digital objects enables you to authenticate them over time by re-running the checksum generation process and checking the values are the same. Generating checksums is therefore a useful process to do at capture.	The ability to generate checksums is often incorporated into free tools that are used to support other pre-capture processes. Karen's Directory printer described below enables checksum generation. A good standalone free checksum generation tool (Fixity) can be downloaded from: https://www.weareavp.com/products/fixity/
File manifestation tool	A tool for enabling you to create a list of the contents of a directory with metadata such as date last created, date last modified, file size, file name, folder name, attributes, extension, compressed size and file version.	Karen's Directory Printer can be used for this purpose and is available to download from: http://www.softpedia.com/get/System/File-Management/Karen-Directory-Printer.shtml
Write blockers. The exact type required may depend on the media you want it to read. For example, you may need one type of write-blocker that will work in conjunction with USB devices and one for handling hard drives on laptops. For further details on possible options see: http://www.hullhistorycentre.org.uk/discover/pdf/Idiot%27s%20Guide%202%20-%20Tableau%20Write%20Blockers.pdf	Enables access to files without making changes to the content. The write blocker does this by enabling 'read' commands to pass while blocking 'write' commands. It means you can open a file without making changes to data stored in the file such as 'date last accessed'. Archive institutions use write blockers	Write blockers are not cheap, and cost in the region of £200-£450 each (see http://shop.avatu.co.uk/shop-by-brand/tableau-

	when capturing incoming digital objects because preventing those types of changes is perceived to be important to maintaining the integrity of the digital objects.	products as a pricing guide). The decision on whether this is necessary will depend on budget and attitudes within the community you serve, i.e. whether having accurate data around things like when a digital object was created/last accessed is considered to be vital.
File characterization tool	You may receive digital files where it is difficult to tell the file format. File characterization tools read the file format from the file signature embedded in the binary digits not just the file suffix. Determining the file format can be important in informing future preservation actions.	DROID is a free file characterization tool developed by the National Archives (UK). To download DROID: http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/ DROID user guide: http://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf
Disk imaging tool	Disk images are single files representing the complete content and structure of a given storage media device. Creating disk images is seen as preferable to just copying and pasting folders and files from storage media as the disk image ensures that essential metadata and technical dependencies are retained. You can create either forensic disk images (which include deleted files and slack space) or logical disk images (which capture files as they appear when using the device – donators may be happier with the creation of logical disk images)	FTK Imager is a free tool that can be used for this purpose. Available to download from: http://accessdata.com/product-download/?/support/adownloads#FTKImager

Developing a workflow for preparing digital objects for check in (capture)

Having set up your workstation, you will need to spend time experimenting and developing a workflow for managing the tasks that you want to perform as part of checking-in (capturing) digital objects prior to storage. The following steps lay out a basic workflow for this, working from the assumption that you are wanting to process digital objects that you have received on physical storage media (hard drive, CD, floppy disk, memory stick etc.). Steps relevant to following the advice given in the digital preservation overview are highlighted in light orange, additional or alternative steps if you have the tools installed are highlighted in dark orange. The aim of this is to provide a generic starting point for developing your workflow, it does not provide hard and fast rules. It assumes that the digital objects on the storage media you are seeking to access will be readable within your hardware/software set up. Your workflow will undoubtedly need adjustment in practice, depending on your unique set-up, and the type of content you are processing and experimenting with the tools is crucial in helping you to work out how the stages fit together and how they will work with your existing processes for documenting collections.

1.	Before you start you will need to make sure that you have new updates loaded for your virus software. If you have decided to keep your dedicated workstation offline (to minimize the chance of infection), updating the software will need to be performed manually by downloading the latest signatures to a dedicated USB stick for transfer to your dedicated workstation
2.	If using a write blocker, make sure this is connected and appropriately set up prior to accessing the media
3.	Insert disk into the appropriate drive (or other media into appropriate reader or attach other storage device to appropriate port)
4.	Run an initial virus check
5.	Copy data from physical media to the workstation. You can directly copy directories and files from the original medium to the workstation, but you should note that various forms of associated data and metadata may not be transferred so alternatively;
6.	Use a tool such as FTK Imager to copy the data from the physical media to your workstation as a disk image, which is a single file that contains an exact bitstream copy of the disk's content, ensuring that various forms of technical dependencies and essential metadata are retained. Think about how you name the disk image: using an ID associated with the collection the physical media came from, along with a unique reference number assigned to the physical media item, will help you keep track of items and manage their inter-relations. For more detailed guidance see: http://www.hullhistorycentre.org.uk/discover/pdf/Idiot's%20Guide%203%20-%20FTK%20Imager.pdf
7.	Use a file manifestation tool such as Karen's Directory Printer to generate a copy of the disk directory information (a file list) and consider how you will keep this information. You can save the information as a text file which can be stored alongside the disk image and/or imported into an Excel spreadsheet for further manipulation in various ways. Consider how you will make the information accessible alongside other cataloguing information for the collection. For more detailed guidance see: http://www.hullhistorycentre.org.uk/discover/pdf/Idiots%20Guide%204%20-%20Karens%20Directory%20Printer.pdf
8.	You may also want to use a file characterization tool such as PRONOM to determine the exact file formats you are taking on. This is useful for moving beyond bitstream preservation and towards logical preservation.

9. Generate checksums for your incoming objects. Karen's Directory Printer or DROID enables you to generate and record checksums, so if you do not use these tools you will need to think of alternative methods for generating and recording checksums, for example using a tool such as Fixity https://www.weareavp.com/products/fixity/
10. Quarantine the objects on the workstation for a month. Update the virus software again and then perform a second virus check 31 days after the first scan.
11. You can now generate multiple copies of your digital objects and transfer these to your dedicated storage
12. Consider how you might create or update an associated catalogue or descriptive tool with information about the steps that were taken to capture the digital objects with information on the location of the files.

Developing procedures for gathering relevant information on the handover of digital material from donators

Adapting your acquisition procedures to make sure they adequately take account of the possibility of material coming to you in digital form is important.

You can save a lot of head scratching further down the line if you have developed procedures which enable you to adequately gather and document relevant information from the donor at point of transfer.

When taking in physical digital media from a donor such as a floppy disc, CD, memory stick etc., try to establish the following key information:

- Do they have a list of what they think is on the media that you can verify the contents against?
- Can they give you details of the file formats, operating system, and computer hardware/software environment that the files were created in?
- Have they tried to access the contents recently, and if so what happened?

The aim is to gather as much information as possible to make your job in assessing the material as easy as possible.

Selecting what to preserve

When offered born-digital content from your donors, you have to decide whether it is worth you taking it in to preserve it. This is a case of weighing up the relative costs that might be involved in preserving the material, against the relative value of the material offered. The questions you will have to ask on an ongoing basis in relation to new incoming material are

similar to those already outlined for weighing-up what to tackle first when dealing with legacy storage media in your collections. You will need to consider:

- Whether the digital content is unique, or whether it is already replicated in your collections and are easily accessible elsewhere
- Whether or not it is likely that you will be able to access and process the digital content in-house or whether you will have to use a specialist service to retrieve the files. In the case of the latter, the larger cost implications will need to be factored into deciding whether it is worth taking the material in
- Whether the file formats of the material are well supported, and therefore more straightforward to preserve in an accessible form over time

You might consider developing a policy that sets out which file formats you can readily accept or prefer to receive material in, and how you make decisions over whether to take material in which can be shown to donors.

Using third party data recovery services

For material that sits on obsolete physical storage media, which you think will be valuable to preserve, you may consider using a specialist data transfer service to retrieve the files from the physical media.

There are lots of companies that advertise these services and so it is worth phoning around to get quotations. When making contact, be clear about the physical storage media the data resides on, and pass on everything you know about the likely contents. After getting an idea of the cost you will then need to decide whether it is worth proceeding.

More Advanced Guidance on Preservation

Links to further preservation advice

Further advice on tools

Community Owned Digital Preservation Tool Registry, COPTR
http://coptr.digipres.org/Main_Page

When looking for free tools to perform a preservation action, the Community Owned Digital Preservation Tool Registry (COPTR) can be a good place to look at options. COPTR describes tools useful for long-term digital preservation and acts primarily as a finding and evaluation tool to help practitioners find the tools they need to preserve digital data. COPTR aims to collate the knowledge of the digital preservation community on preservation tools in one place. COPTR captures basic, factual details about a tool, what it does, how to find more information (relevant URLs), and references to user experiences with the tool. The scope is a broad interpretation of the term "digital preservation". In other words, if a tool is useful in performing a digital preservation function such as those described in the OAIS model, then it is within scope of this registry.

Catalogue of Tools & Services, DCC
http://www.dcc.ac.uk/resources/external/tools-services

The Digital Curation Centre (DCC) is an internationally renowned centre of expertise in digital curation, who provide access to a range of resources which aim to equip those who work with digital data with the skills to manage their data effectively. The DCC have compiled a comprehensive catalogue of tools and services for engaging with data. Of particular interest for digital preservation are the tools for Depositing and Ingesting Digital Objects, Archiving and Preserving Information Packages, and Managing and Administering Repositories. These sections identify a range of tools and services for normalization and migration, web archiving, validation of digital files and preservation planning.

Migration and Emulation Tools
http://blog.case.edu/digitalpreservation/2010/11/29/week_5_migration_and_emulation_tools

This blog post by Virginia Dressler, Digital Projects Librarian and Assistant Professor at Kent State University (USA), provides helpful information about selecting the right preservation strategies, addressing the advantages and disadvantages of both migration and emulation. Dressler also provides links to further resources and tools.

PRONOM
http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

PRONOM is a technical registry created by The National Archives (UK), providing technical information and characteristics about a range of digital file formats. It holds information about software products and the file formats that can be read by those products. It also provides information about file formats, such as their current version, related file formats, format risk, and whether the format is still supported. The registry is extensive, and be searched using a number of entry routes. PRONOM is the technical registry which works behind DROID, the Digital Record and Object Identification tool also created by The National Archives.

Advice on content specific preservation

For advice on preserving moving pictures and sound:

http://dpconline.org/handbook/content-specific-preservation/moving-pictures-and-sound

https://www.clir.org/wp-content/uploads/sites/6/pub164.pdf

http://326gtd123dbk1xdkdm489u1q.wpengine.netdna-cdn.com/wp-content/uploads/2017/02/Screen-Heritage-UK-Moving-Image-Collections-Handbook.pdf

http://natlib.govt.nz/collections/caring-for-your-collections/sound-recordings

http://www.communityarchives.org.uk/content/resource/film-and-videotape-and-the-community-archive-2

For advice on web archiving:

http://dpconline.org/handbook/content-specific-preservation/web-archiving

https://nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf

For advice on digitization:

http://collectionstrust.org.uk/resource/creating-your-digitisation-strategy/

http://326gtd123dbk1xdkdm489u1q.wpengine.netdna-cdn.com/wp-content/uploads/2016/11/JDM-DECIDING_TO_DIGITISE.pdf

Other guides to digital preservation

The Digital Preservation Coalition website (http://www.dpconline.org/) provides up-to-date information on digital preservation training, events, new technology and guidance, including the Digital Preservation Handbook. Available at: http://www.dpconline.org/handbook/

The National Archives (UK) provides a handbook on 'Selecting file Formats for Long-Term Preservation'. Available at: http://326gtd123dbk1xdkdm489u1q.wpengine.netdna-cdn.com/wp-content/uploads/2017/02/The-National-Archives-Selecting-File-Formats-for-Long-Term-Preservation-Aug-2008.pdf

The Library of Congress (LOC)'s section targeted at private individuals on how to ensure the long-term survival of personal digital materials, such as family photographs and videos, is also a useful starting point for very small community archive set-ups. Available at: http://digitalpreservation.gov/personalarchiving/

The State Library of New South Wales (New Zealand)'s Public Library Services have put together a handy guide of strategies for digital preservation. Although the guide is aimed at digitally preserving a library collection, the information is equally relevant and important to archival preservation. Available at: http://www.sl.nsw.gov.au/public-library-services/digital-practice-guidelines-public-libraries/digital-preservation

This presentation, 'Digital Presentation: An Introduction to the Basic Concepts', by the American Association for Library Collections, covers the basic concepts within digital preservation, together with risks to digital material, and some of the tools and strategies for managing them. Available at: https://www.youtube.com/watch?v=RqacRC51CRI

The National Archives (UK) have created guidance for 'How to get started' with digital preservation. It provides some basic information for starting out and a set of basic principles, their 'golden rules'.
Available at: https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/preserving-digital-collections/how-to-get-started/

The National Archives (UK) have also created a handy guide for developing a digital preservation strategy and policy. Available at: https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/preserving-digital-collections/developing-a-digital-preservation-strategy-and-policy/

This report from the Online Computer Library Centre (OCLC), 'You've Got to Walk Before You Can Run: first Steps for Managing Born-Digital Content Received on Physical Media', is intended for anyone who does not know where to begin with managing born-digital materials. It is short and simple, providing guidance on surveying and creating an inventory of digital holdings, together with a set of technical steps for ingesting readable media. Available at: https://www.oclc.org/content/dam/research/publications/library/2012/2012-06.pdf

Glossary of Terms

Bit
A bit is the most basic unit of information in a digital object, and is represented in binary form (ones and zeros).

Bit rot
Also known as 'bit decay', bit rot is the slow deterioration and degradation of digital data (the bits), its integrity and performance over time.

Bitstream
Also known as 'bit stream' and 'binary sequence', a bitstream is a sequence of data (the bits) in binary form (ones and zeros) E.g. 001011010.

Born-digital
Born-digital is an all-encompassing term used to describe digital objects which have been created in a digital form. Born-digital items can include, but are not limited to, digital photographs, spreadsheets and e-mails.

Checksum
A checksum is a string of characters that relate to a digital object, and which act as the object's unique signature or digital finger print. Checksums can be used for checking the integrity of a digital object through comparison of the checksum over time (see **fixity**).

Digital curation
Digital curation is the actions and activities implemented over time for the management, maintenance, and preservation of digital objects.

Digital Curation Centre
The Digital Curation Centre (DCC) is an internationally-reknowned centre of expertise in digital curation. Their primary focus is on research data management, however they provide access to a range of resources which aim to equip those who work with digital data with the skills to manage their data effectively.

Digital preservation
Digital preservation is the actions taken to ensure the continued accessibility and use of digital materials, from a single digital object, to software or hardware.

Digital repository
A digital repository is an electronic location where collections of digital objects are stored and managed.

Digital surrogate
A digital surrogate is a digital copy of a paper-based or physical object. Digital surrogates are often created in order to enable a form of access to an original object, whilst promoting its preservation.

Digitization
Digitization is the process of creating a digital version of a paper-based or physical materials into a digital format. Digitization may take place for the purposes of both access and preservation of original material.

Disk image
A disk image is a file containing an exact copy of the entire contents of an electronic storage device.

DROID
DROID stands for **D**igital **R**ecord **O**bject **Id**entification. It is a free file characterization tool developed by The National Archives (UK) which can automatically create a profile for digital objects, including information such as file size, file format, and date last modified (see **file characterization**),. DROID can also generate checksums (see **checksums**).

Emulation
Emulation is the process of using software and hardware to imitate a digital object in its original form and with its original characteristics.

File characterization
File characterization is the process of identifying the particular characteristics of a digital file. This is achieved by running a digital object through file characterization software, such as Karen's Directory Printer or DROID (see **Karen's Directory Printer** and **DROID**).

Fixity
Fixity is the quality of remaining unchanged. It is important to do periodic fixity checks on digital objects held within a repository to check that the digital objects have not degraded or been altered. Fixity checks can be processed by running digital objects which have previously been given a checksum through the same checksum-generating software, thereby generating a new checksum. This new checksum can then be compared against the previous one to identify whether the checksum, and therefore the digital object, has degraded or been altered in any way.

FTK Imager
FTK Imager is a free disk imaging tool which can be used for the purpose of creating a disk image (see **disk image**).

Ingest
Ingest is the process of capturing material into archival storage for the process of preservation and access.

Karen's Directory Printer
Karen's Directory Printer is a file manifestation tool, which enables you to create a list of the contents of a collection of digital objects with metadata, such as the file size, file name, file format, and date last modified.

Metadata
Metadata is, literally, data about data.
Migration
Migration is the process of copying or converting digital objects from one format to another, such as converting a Microsoft Word document into a PDF. In digital preservation, migration can be utilized for managing the threat of technology obsolescence.

Normalization
Some digital repositories will place a limit on the number of formats which they will support, and as such may only support the formats which most best overall promote functionality, longevity and preservability. Normalization, in this instance, is the process of converting a digital object from its original format to an accepted format, so that a repository can ingest and support the object.

Write blocker
A write blocker is an electronic device which prevents the ability for digital objects to be changed or altered during the process of transfer from one storage device to another. Write blockers were developed as a digital forensics tool, but can be used for digital preservation purposes during the ingest of digital objects into a repository.