

Novice to Know-How Module Text

Course 2: Introduction to Bitstream Preservation

Module 2: File Formats and Digital Preservation

The development of this course was funded by The National Archives (UK) as part of the "Plugged In, Powered Up" digital capacity building strategy.

1. File Formats and Digital Preservation.

In the previous module we looked at what files and file formats are. In this module we will be delving into what digital preservation issues they raise.

Much of digital preservation is about addressing risks to the digital content we want to preserve. Therefore, we will be framing the file and file format issues in relation to related risks.

Finally, towards the end of the module we will look at how we might identify file formats that we might "prefer" for preservation as they have lower risk profiles.

2. Digital Preservation File Format Issues.

There are a number of important issues to consider in relation to digital preservation and file formats, they include the following.

Obsolescence.

Formats evolve as new functionality is identified and added. File format obsolescence may become an issue as new formats, or versions of formats, emerge or as new software phases out support for old formats. Lack of backwards compatibility may render digital content unusable. This can happen as vendors try to entice people to buy new versions of software or with open-source solutions as formats fall out of common use.

File format obsolescence was originally thought to be a major digital preservation risk, but the problem may not be as severe as originally perceived. Many established file formats in common use (e.g. image formats) are quite stable, and have remained supported and useable. It is quite likely that the majority of file formats you deal with will fall into this category.

Lossless vs Lossy.

As part of their encoding some formats use a process called compression to reduce the size of files. Compression comes in two forms: lossless and lossy. Lossless reduces the size of the file without any loss of information or quality, lossy compression generally involves "throwing away" some of the information to make a file that is very similar in content but much smaller and often lower quality.

One rule of thumb could be to choose lossless formats for the creation and storage of "archival masters", and lossy formats for delivery/access purposes. This allows high quality copies to be stored for preservation but smaller files to be created for users to access.

An example of how this can be implemented is choosing the TIFF image format, which has lossless compression, for high-quality archival images, and the lossy format JPEG to make much smaller access copies.

Proliferation of File Formats.

Proliferation of file formats can be a major challenge and there are two main issues. The first is usually caused by rapidly evolving bespoke data file formats, often linked to proprietary software and/or related instruments. Each vendor's may have its own format, this can lead to a large number of incompatible proprietary, closed formats.

The second is when, often due to collecting from external depositors, an organization ends up with many different formats for a particular type of content. For example, you may have text documents in Word, PDF, Text, Rich Text, and Open Office formats (to name just a few!)

Managing all these formats (which are at risk and which tools can be used for each) can be a challenge. You may wish to identify "preferred formats" and use these to inform content creators and help decide which formats you will use for archival copies of digital content.

Support for Metadata.

As discussed in the module "Files and File Formats", many file formats record metadata about the file and its contents within the file itself. This can be a very useful for digital preservation, particularly if you are hoping to automate processes due to a lack of resources or the need to manage large numbers of files.

There are many readily available tools to help with extracting metadata from files, including the UK National Archive's DROID. An explanation of the process can be found in the course "Using DROID for Integrity Checking and Characterization".

Support for Content and Functionality.

When investigating file formats for preservation we need to think about how the format supports the information contained within the digital content and what functionality is important to a user's experience. These are often referred to as the "significant properties".

As an example, the formulae that sit behind cells in a spreadsheet give context to the numbers we see on screen. Therefore, it is best to preserve these files in a format that will retain the formulae. If we save just the information seen on screen to, say, a PDF, we are losing a lot of the essential meaning from the file.

Availability of Specifications.

In the module “Understanding Files and File Formats” we looked at different types of file formats in relation to the availability of their specifications: open-source, proprietary but open, and proprietary but closed. This is important as being able to access a file format’s specification can help with digital preservation decisions and processes.

It can allow us to check how closely a file matches the specification its format (a process called validation), it can allow us to understand the metadata and functionality a format supports, and it can enable the reverse engineering of software for defunct formats.

Therefore, many favor open-source, or at the very least open proprietary formats for preservation.

3. Selecting Preferred File Formats.

In the previous section we examined some of the key issues relating to file formats and digital preservation, and mentioned that identifying preferred file formats for preservation can be a helpful exercise.

Having a list of preferred file formats can allow us to offer guidance to content creators and depositors, and to plan for preservation. The preferred list should cover the types of content you expect to receive regularly, and represent the file formats you have most confidence in your ability to preserve.

In selecting formats for the list you should consider the issues already mentioned as well as which formats are in common use and are likely to be well supported long-term. It is also worth working through the steps on the next two slides.

4. Know What You (Will) Have.

When developing a list of preferred file formats it is essential that you consider types of digital content you already have for preservation. Also, consider what content you expect to receive.

If you already (or are expecting to) have large amounts of content in a particular format that is commonly used, you may wish to add that straight to the list. Likewise, if you expect to be receiving a particular type of content you may wish to pick a format that you can recommend to depositors and content creators.

We will be looking at how to audit your current collections in the “Ingesting Digital Content” course.

5. Get Some Advice from Others.

While a preferred file format list is most useful if it reflects your own preservation context, you do not need to start from scratch. Consider contacting similar organizations to discuss if they have done work in this area. A number of organizations have also made their list available online. You can find links to some of these in the Resources section of this course.

You may also wish to consult with colleagues from your IT department who can offer advice on formats' compatibility with existing or proposed systems. It is also something to discuss with vendors if you decide to procure a digital repository system.

6. Drafting and Using Your List.

Once you have completed your research and analysis you are ready to draft your preferred files formats list. During the drafting phase remember to be realistic about your organization's preservation capabilities and how much strictly you can enforce requirements on content creators and depositors. You may need to consider listing "acceptable" formats as well as preferred.

Remember to review your list periodically to make sure it remains fit for purpose reflecting current technology and good practice. Also, consider sharing your list to help others working through the same process!

7. What's Next.

Now we have established the digital preservation issues around files and file formats, we need to start thinking about how we address them.

The remaining modules in this course will begin to explore approaches to the preservation of digital content. We will begin with an "Introduction to Bitstream" before providing an "Introduction to Workflows", and then describing a key process called "Integrity Checking".