

Novice to Know-How Module Text

Course 2: Introduction to Bitstream Preservation

Module 1: Files and File Formats

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. Understanding File Formats.

In the first course of Novice to Know-How we started to explore the basic elements of digital preservation. But before we delve deeper into practical first steps, it is important to understand exactly what we are aiming to preserve.

So, our next step is to understand digital files: how they are structured and stored, what file formats are, and what information they contain that can help with preservation. Being able to understand and correctly identify files and their formats is fundamental to successful digital preservation.

2. It is All in the Bits....

In the first course of Novice to Know-How we started to explore the basic elements of digital preservation. But before we delve deeper into practical first steps, it is important to understand exactly what we are aiming to preserve.

So, our next step is to understand digital files: how they are structured and stored, what file formats are, and what information they contain that can help with preservation. Being able to understand and correctly identify files and their formats is fundamental to successful digital preservation.

3. Bits Make Up Bytes.

A byte is made up of 8 bits such as 01101100. A byte can represent 256 different combinations. One byte can store one character of text, e.g. A or £, but one character may be made up of more than one byte.

A byte is the basic unit we use for describing storage. For example:

1024 Bytes = 1 Kilobyte (1KB)

1024 Kilobytes = 1 Megabyte (1MB)

1024 Megabytes = 1 Gigabyte (1GB)

1024 Gigabytes = 1 Terabyte (1TB)

And so on....

4. Character Encoding.

To change characters in text files to and from binary, computers use character encoding. Character encoding defines the contents of a byte (or combination of bytes) that represents a character.

There are several different standards for character encoding. If you try opening a file using an incorrect encoding standard you may get information returned with anything from minor mistakes to a document displayed as nonsensical text. It is therefore important to know the correct character encoding.

Common character encoding standards include ASCII, Unicode, and Big-5. UTF-8 is another common standard that aims to represent the complete character set across different languages and includes over 1 million characters.

5. What is a File?

A file is the unit used by computers to record discrete pieces of information. By recording information in files it allows the computer to organize and store the information in a way that will facilitate retrieval, display, and use.

Typically, files are organized in a file system, which keeps track of where the files are located. The bytes that make up a file might be stored in several different places on the storage used. The file system will record where the pieces are so it can retrieve and reassemble them when the file is needed for use.

Storing information in files also makes sharing and transfer possible, for example via the Internet or as attachments with emails.

6. Files vs Digital Objects.

Files contain information but they require software to be able to render them on screen. A single file can contain everything needed by the software for processing into useable content, e.g. an image or video file, but sometimes multiple files are needed, e.g. a text file which contains an embedded link to a separate spreadsheet.

So, a bitstream is how information is stored on a computer, a file is the unit of storage used to contain a discrete piece of information, and digital content is made up of one or more files that are interpreted by software to render useable information on screen. Therefore, for preservation it is useful to intellectually separate files and the content we use on screen, and we differentiate between bitstream and content preservation.

7. What is a File Format?

A file format is a standard way that information is encoded for storage in a computer file and knowing the file format is critical to rendering the bitstream of a file correctly onscreen.

Some file formats are designed for very particular types of data. For example, JPEG and TIFF are digital image file formats. Other file formats, however, are more like containers and are designed for storage of several different types of data: the Ogg format can act as a container for different types of multimedia including any combination of audio and video, with or without text (such as subtitles), and metadata.

Other formats, such as HTML and the source code of software are text files a computer knows to process in a particular way to allow them to be used for specific purposes.

8. Types of File Format.

There are thousands of different file formats and each is defined by its specification. This sets out the encoding method used and the intended functionality when opened with the correct software. There are differing levels of access to file format specifications. Click on the images below to learn about the different levels of access.

Open Source

Open source file formats are developed through an open community-driven process. The specification is publically shared intellectual property and is often maintained by a standards organization.

Examples of open source file formats include:

- JPEG
- Tagged Image File Format (TIFF)
- Free Lossless Audio Codec (FLAC)
- Portable Document Format (PDF)

Proprietary but Open

Some file formats are developed by commercial companies, often linked to particular software they develop. With "proprietary but open" formats, the company controls the format and owns all intellectual property but makes the specification available (sometimes with restrictions)

Examples of proprietary but open file formats include the Microsoft Open Office XML format such as DOCX and XLSX.

Proprietary and Closed

These file formats are also developed by commercial companies, but the specifications are more closely guarded as a way to control the technology and ensure their market share. With "proprietary but closed" formats, the company does not make the specification accessible and the format can normally only be used with licensed software developed by the same company.

Examples of proprietary but closed file formats include the original Microsoft formats (e.g. DOC, XLS, PPT), RAW image formats developed by camera manufacturers, and Adobe Photoshop's PSD.

9. Useful Metadata Within a File.

It is usually possible to identify the file format of an individual file by its extension, e.g. .doc, .mp3, or .jpg. This and the file name are perhaps the most basic metadata we can capture about a file.

Many files, however, also include useful metadata within the file itself, often in a section referred to as the file header. As an example, image file headers may include metadata such as image format, size, resolution and color space. In the course "Using DROID" you will learn how to use the tool DROID to extract this metadata.

Now we understand files and file formats, in the next module we will examine the related digital preservation issues.

10. Wrap-Up.

In this module we have established that all digital information is stored as a series of 1s and 0s known as binary. The information is translated into binary using encoding standards.

Files are discrete units used to store and organize chunks of information, and a file format controls the encoding of information into the file. A digital object for preservation can be made up of one or more files.

Also, an awareness of file formats and the metadata they can contain is useful for digital preservation. Before we examine this in more detail in the next module, let us do a quick knowledge check.