

Novice to Know-How Module Text

Course 5: Ingesting Digital Content

Module 3: Creating a File Manifest

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. Introduction.

In previous modules we have talked about integrity checking and characterization, two processes which generate information about our digital content at a file level. But where should we store that information?

In this module we will introduce verifiable file manifests. An important resource used to store metadata about files which can be used to manage them over time.

2. What is a Verifiable File Manifest?

The term "manifest" is typically used to describe a list of items of cargo being transported from one place to another. It provides a way of listing and identifying each piece of cargo so that the shipment can be audited on receipt. This enables any missing or damaged items to be identified and ultimately replaced.

A file manifest works in exactly the same way. It provides a list that can be checked at any time to ensure that all the files are present and as they were at the point the manifest was created.

"Verifiable" means that along with a list of filenames, a verifiable file manifest includes a checksum for each file. With precise information that captures the fixity and identity of each file on the manifest, it is possible to perform a quick and easy automated verification of a batch of files. This makes the use of manifests a really powerful tool for digital preservation.

3. Why is it Useful for Digital Preservation?

Whilst issues such as file format obsolescence require careful consideration, our digital files are also at the mercy of far more mundane risks. Every time we move, store, process or otherwise interact with our digital files, something might go wrong. For example:

- Whilst copying files over a network, the network goes down and only half the files are copied.
- Files are copied from one storage location to another, but the destination is not big enough to hold all the files.

- Files are copied to a new location, but some files are written over by files with the same filenames
- A human operator presses "delete" at the wrong moment
- And, of course, almost all software is imperfect - it contains bugs. Unexpected behavior can easily lead to lost or damaged files

Having an approach that can be used to verify that a batch of files are still all present and correct is incredibly useful for mitigating these risks. If detected quickly, it's usually possible to go back to the source for an additional copy of the original data.

4. What Form Will a Verifiable File Manifest Take?

A typical verifiable file manifest will consist of a list of information about a batch of files. The manifest will contain at a minimum two key pieces of metadata for each file:

- A file path. This identifies which file the manifest is referring to, and where it is stored
- A checksum. This provides a way of verifying the integrity of the file (see the Integrity Checking module for more information)

It is also possible to use the manifest to store additional basic technical metadata about the files, such as the metadata produced by a characterization tool such as DROID.

The most common formats for storing file manifests are in a spreadsheet, a comma separated value file (.CSV) or database. Each of these is an easily processible format. If the manifest is stored as a spreadsheet or CSV file it is easy to store a copy along with the digital content. Although it is good practice to maintain another copy centrally in case of issues with storage of the digital content.

5. What Tools are Useful for Capturing the Data?

Most software tools used for creating checksums will be helpful in creating a verifiable file manifest. A list of common integrity checking tools is included in the "What is Integrity Checking" module and also in the resources for this course. Likewise, many characterization tools can also be used, these are also included in the course resources.

The Bagger software, from the Library of Congress, takes the concept a little further and packages a batch of files - known as "a bag" - into a container which includes a verifiable file manifest.

A typical Digital Repository application should be able to receive a verifiable file manifest alongside a batch of files on ingest and verify the manifest to provide confidence that all files have been ingested into the repository and none have become damaged.

6. Wrap-Up.

So, a verifiable file manifest is a list of files used to store metadata about a collection of files that will help us to manage those files over time. It is particularly useful in helping to mitigate risks associated with moving or copying data.

At a minimum the manifest should contain a file path providing a location for each file, and a checksum to facilitate integrity checking. Additional basic metadata from characterization can also be added. There are many tools available for capturing this data and it is best if it is stored in a processible format such as a spreadsheet, CSV file, database.

In the next module "Creating a Digital Asset Register" we will look at a complimentary way to capture high-level metadata about collections of digital content. But first, a quick knowledge check.