

Novice to Know-How Module Text

Course 2: Introduction to Bitstream Preservation

Module 5: Integrity Checking

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. What is Integrity Checking?

Integrity checking (also known as fixity checking) is a core element of digital preservation. It allows us to make sure that digital content has not been lost, changed or damaged.

It is a process that involves using software to generate a checksum (more on this on the next slide) that represents the structure of a digital file. We can then compare checksums generated at different points in time to see if any changes have occurred.

In this module we will examine the process in detail and look at a way it could be implemented.

2. What is a Checksum?

A checksum is a (close to) unique string of numbers and letters (for example **02ace44afd49e9a522c9f14c7d89c3e9**) that can act as a 'digital fingerprint' for a file. Even the smallest change to the file will cause the checksum to change completely.

Checksums can be generated using a range of readily available and free-to-use tools. It is important to note that whilst checksums can be used to detect if the contents of a file have changed, they do not tell you where in the file that the change has occurred.

Common types of checksum include MD5, SHA-1 and SHA-256.

3. What Checksum Algorithm Should We Use?

Different checksum algorithms have different strengths and weaknesses. For example, MD5 checksums are quicker to calculate but less complex, while SHA256 take more processing but are stronger cryptographically.

The first question to answer when selecting a checksum algorithm is: what are we using the checksums for? If it is just to detect data corruption or loss, then MD5 checksums will be

sufficient. If, however, they are to be used to identify duplicate files, or to check for malicious damage, then you will want to use a stronger algorithm such as SHA256.

Other factors that might affect your choice include:

- What tools are available? Can you use open-source software? Do you need a graphical interface? Most tools will produce a limited selection of checksum types.
- Do the checksums need to be interoperable with a number of tools? You will then need to use an algorithm common to all of the tool.
- How powerful are the computer(s) you will be using to generate the checksums? You might not have the processing power to support SHA256 generation.

4. Why Should We Use Integrity Checking?

There are three main uses of integrity checking:

When Moving Data.

Digital content is always at higher risk whenever it is being moved or transferred. Therefore it is wise to use integrity checking to ensure no damage has occurred.

- Examples of this use may include:
- After receiving digital content from an external depositor.
- After moving digital content between storage media, e.g. when replacing storage media.

To Check Storage.

Integrity checking is a great way to monitor the status of your files over time.

Regular integrity checking can detect any irregularities, such as damaged or missing files. Creating multiple copies of each digital object allows you to replace any damaged or missing files if you should discover any problems. Keeping these copies in different locations on different types of storage media helps to mitigate the risk of loss. More on this later!

To Prove Authenticity.

Well-documented integrity checking can help you prove to users the authenticity of digital content.

Integrity checking can show users that a transfer of digital content has been successful. It can also provide evidence that no changes have happened over time.

Now we know why we do integrity checking, what tools can we use for the process?

5. Tools for Integrity Checking.

There are many free and easy to access tools that can carry out integrity checks.

With tools like DROID and FITS, generating checksums is one of a number of functions that they can carry out. For tools like Fixity and Checksum by Corv, integrity checking is their sole purpose, this can make them quicker when checking large quantities of digital content.

Which tool you use will depend on the amount of digital content you have to check and the resources you have available. Links to the four tools mentioned are included in the resources for this course.

Now we have had a high-level look at what integrity checking is, and why we use it, we will take a step by step walkthrough of how it works.

6. A Walkthrough of Integrity Checking.

1. We begin by using a software tool to create a checksum
2. We can think of this checksum as a fingerprint for the file
3. At some point in the future, we will 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. To do this 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. Comparing the checksums, we can see that they are different. This confirms that our file is no longer identical to how it was. Perhaps it has been damaged by media failure or "bit rot".
9. Digital preservation can help us try to avoid this damage and also gives us options for actions to take when and if it does. We will take a look at a possible process for dealing with this on the next few slides

7. Saving Multiple Copies.

1. Digital preservation can help us try to avoid this damage and also gives us options for actions to take when and if it does. We will take a look at a possible process for dealing with this on the next few slides
2. We generate checksums from each file, and we can see that they are all the same, so 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... ...and replace it with a copy of one of the others.

7. So, by maintaining multiple copies of the file we can have confidence that even if files are damaged we can recover undamaged copies. It may even be possible to automate this checking and replacement process.

8. Wrap-Up.

So, integrity checking is an important process for monitoring digital content over time. Allowing us to establish authenticity and check for loss, changes or damage to the files in our care.

In the next course we will be introducing a free tool that can be used for integrity checking, DROID. This will include how to download and install it, set it up, and the basics of how to use it.