Checksums

1 What is a checksum?

A checksum is a computer-generated string of numbers and letters that act as a digital fingerprint for a digital object. Even the smallest change to a digital object will cause its checksum to change completely.[1]

Why are they used?

Checksums are a tool for ensuring the integrity of digital objects. A change in a digital object's checksum indicates a change to the object's data (i.e. changes in content, data loss or corruption). An unchanged checksum indicates that no change has occurred to the object's data since the checksum was created.

What can they be used for?

Checksums are a tool which allows a 'chain of custody' to be established between those who create, preserve, and access digital information and records. In data management there are generally at least three uses for checksums:

1. to confirm that a digital object has successfully been transferred and received from a source without change
2. to confirm that the integrity of a digital object has been maintained while in storage (i.e. it has not been changed or corrupted)
3. to provide users confirmation that the digital object they are accessing has been retrieved, stored and delivered to them without any changes occurring to the data.

2 When are checksums required?

Checksums are required from public offices when transferring digital public archives to Archives New Zealand and to assure the continued unaltered state of digitised information.

The ability for public offices to produce checksums for each digital file is an important readiness characteristic for digital transfer to Archives New Zealand. Checksums can also be a useful tool for public sector organisations to ensure the continuity and integrity of born digital and digitised information and records in their control.

3 Generating and validating checksums

The generation (creation) and validation of checksums can be performed in multiple ways by a number of software tools[2], many of which are open source. Checksums are generated using standard algorithms known as hash functions and there are many different types. Archives New Zealand has the capability to work with any; most often organisations provide either MD5 or SHA1.

[1] For example here is a checksum for a file: f75d91cdd36b85cc4a8dfeca4f24fa14 and here is the checksum for the same file when one letter was changed in the content: 7aca5ec618f7317328dcd7014cf9bdcf

[2] An example is Double Commander which includes the options of generating and validating checksums as part in its basic functions.

It is important to remember that while checksums can detect changes to digital objects, they do not document where the change occurred or what the change is.

4 Checksums at Archives New Zealand

Archives New Zealand uses checksums supplied by the transferring organisation to ensure that files are not altered or corrupted from the time they are exported from the organisation's system until received by Archives New Zealand.

Checksums are also used to monitor integrity in Archives New Zealand's Government Digital Archive (GDA). Rosetta, the long-term preservation system used for managing the GDA, creates checksums using three different hash functions (algorithms) for every single file during upload to the GDA. Then the GDA system monitors and validates those checksums continuously to make sure the integrity of the files is never compromised.