Digital transfer initiation – how to prepare

1 Introduction

Archives New Zealand (Archives) has developed tools and methods for processing digital transfers. These can also be used by public sector organisations to assess their collections before transferring to Archives. A key component is the preparation of an initial test extract or copy of digital information and records (files) that are eligible (ie, that have acceptable transfer characteristics), along with their accompanying metadata. Creation of a test extract is necessary for Archives to determine the feasibility of a full transfer.

The following activities, which help an organisation assess their files in readiness for transfer, can also be useful for an organisation to understand how well they are managing their files for digital continuity. The activities described underpin a successful transfer as they identify issues which are checked during the Transfer Preparation stage and so enable an organisation to discover potential issues in advance.

2 Understand the files

Analyse your files to identify:

- the file formats held, specifically old or obsolete formats and unusual format modifications
- duplicates and versions, by generating and comparing checksum values for each file
- layers of content, such as embedded objects and
- any system files, missing files and empty folders.

Organisations can use an automated tool like DROID (Digital Record Object Identification)[1] for file format identification. Another tool such as SQLint[2] can be used to understand more details about the file set intended to be transferred. SQLint provides an easily readable overview and statistics of the files in the transfer. It can be used (among other things) to:

- quality check the accuracy and consistency of files and content sentencing (eg, showing timelines based on last modification dates) and
- locate obvious sensitive, non-business related and/or draft material by 'black listing' potentially problematic words or characters in the file and folder names.

3 Identify what metadata is needed

At a minimum, Archives expects organisations to provide the mandatory metadata elements as given in the guidance *Minimum requirements for metadata* (16/G7). Please note the following:

- Archives has no fixed requirements for the schema or structure of this metadata but we prefer CSV, TXT, Excel or XML file formats
- Archives' require the metadata to be **structured** consistently
- the file with metadata uses UTF-8 coding

[1] DROID is a file format identification freeware created by The National Archives in the United Kingdom, and can be downloaded from their website (http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/).

[2] SQLint is a simple command-line linter which reads SQL files and reports any syntax errors or warnings it finds. A linter or lint refers to tools that analyse source code to flag programming errors, bugs, stylistic errors, and suspicious constructs (https://github.com/purcell/sqlint).

- file folder names are free of non-standard characters (only ASCII) and
- importantly, there is someone in the organisation who understands the metadata and can assist Archives to understand it which will facilitate mapping to Archives' systems.

It is recommended that organisations use the export format options of the system(s) in which their files are stored to export both the files and their associated metadata. The easiest option is to export all the metadata fields available in the system and then analyse those in collaboration with Archives to decide which fields provide context and assist with discovery. As some systems do not have metadata export functionality, organisations may need technical knowledge and/or IT support to do this. Archives can provide some advice and support, but organisations may also need to consult the system designers or vendors.

3.1 Checksums

A checksum value is an essential metadata element that is required to ensure the integrity of files. **Checksums must be generated by the organisation before the files are transferred to Archives**, either from within the original storage system or immediately after the files are exported from it. The checksum values must be provided to Archives as part of the metadata describing each file, or in a separate file (ideally both). Providing the checksum for each transferred file allows Archives to validate the file and make sure that all the files have been transferred successfully with no changes or errors introduced during the transfer.

Checksums can be generated by DROID as well as free online tools such as Free Commander (Windows) and SHA1SUM or MD5SUM (Linux). See the guidance *Checksums* (17/F25).

4 Create an initial test extract or copy

So that Archives can determine whether a full digital transfer is feasible, the transferring organisation must identify and assess an initial test set of eligible files and their metadata as explained above. The organisation must then extract or copy these onto a removable hard drive that can be secured with encryption (which Archives can provide if required), or arrange an alternative method for secure transport to Archives. To copy and synchronise the files, Archives recommends using the tool 'rsync'[3] (Remote Sync) which preserves the integrity of the files and the metadata. Archives is happy to advise transferring organisations on its use.

NOTE: Organisations must not delete any files or metadata at this stage of the transfer process as Archives only needs a reliable copy.

5 What happens next?

Once the initial test extract is received by Archives, several analytical processes (both automated and manual) are run over the files and metadata. This analysis identifies any content, technical, metadata and accessibility issues that may affect a full transfer and ingest into the Government Digital Archive. Archives then consolidates the analysis results in a report for discussion with the transferring organisation. The report outlines the transfer readiness of the files (ie, the quality and consistency of sentencing decisions), identifies unique file formats and potential digital preservation issues. This makes it possible to recommend that the organisation undertakes further preparation or proceeds with the digital transfer.

[3] **rsync** is a utility for efficiently transferring and synchronising files across computer systems, by checking the timestamp and size of files (https://en.wikipedia.org/wiki/Rsync).