

Novice to Know-How Module Text

Course 5: Ingesting Digital Content

Module 1: Setting Up an Ingest Workstation

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 various workflows and processes required in digital preservation, but what hardware and software do we need to carry these out?

In this module we will be examining what you will need to set up a workstation for the initial processing of digital content when it is received.

This will include a recap on the processes covered, what hardware you may need, and types of software you should consider installing on the workstation.

2. What is an Ingest Workstation?

An ingest workstation is the set-up of hardware and software required to process digital content we have received so that it is ready for long-term preservation. The combined processes and procedures being known as "Ingest".

At a basic level this includes:

- Copying digital content from original storage media
- Virus checking
- Integrity checking
- Creating/processing necessary metadata and documentation

As processes are developed further this may also include appraisal, validation and analysis of file formats, redacting sensitive data, migrating file formats, and 'packaging' content for preservation.

Once the ingest processes have been completed the digital content and its metadata and documentation will be moved from the workstation to long-term storage.

3. Ingest Checklists.

Developing a checklist for the ingest process is an excellent way to ensure good practice and consistency of approach. The steps required will depend on your own workstation set-up, the types of content you are processing, and your digital preservation plans, but as a minimum they should cover the processes described on the previous slide.

Breaking the processes down into simple steps is useful if others will need to use your checklist. Also, it can be useful to mark if steps are required or optional, as there may be variations in how different content types are processed.

4. Recommended Hardware.

When setting up an ingest workstation, hardware requirements are likely to be our first consideration. What physical equipment do we need to facilitate ingest processes. There are three main types of hardware that you will need: a computer, readers for storage media, and write blockers.

5. Hardware: A Computer.

It perhaps seems self-evident that we will need a computer for our ingest workstation, but there are some specific requirements we need to consider:

- We will want to use the highest spec PC or laptop resources will allow as it will likely need to process a large amount of data
- Likewise, we should ensure it has as much storage as you think you will need to hold digital content while it is waiting to be fully processed.
- Finally, if you think the risk of the digital content containing viruses is high, the computer should only be connected to our organization's network for updates of virus software and transfer of processed content to long-term storage.

6. Hardware: Media Readers.

Current PCs and Laptops generally no longer come with media readers other than USB ports installed as standard. So, we will need to consider purchasing external media readers that can be plugged-in to our ingest workstation. These might include:

- CD/DVD drive
- Floppy disk drives (e.g. 3.5" or 5 1/4")
- Zip drive
- Tape drives
- A caddy for external hard drives

An audit of existing and expected digital collections will help decide which you may need. If you do not have the resources to buy everything, prioritize based on the importance of the content or the most common media, e.g. if 80% of the content is on CDs, a CD/DVD drive is the priority.

Sometimes hardware, in particular readers for early storage media can be hard to source. Organizations have found auction sites such as eBay to be useful in finding items for sale.

7. Hardware: Write Blockers.

A write blocker is a tool that stops changes being made to digital content on a piece of storage media. Both hardware and software write blockers are available.

They are particularly useful for digital preservation as they guarantee the integrity of the digital content on the original media. Ensuring no changes, accidental or otherwise, are made.

In their hardware form they are inserted between the computer and storage media to ensure the digital content on the media cannot be altered. The example shown on the right is a write blocker for USB drives. Here, the write blocker would be plugged in to the USB port on the ingest workstation and the USB drive is plugged into the write blocker. We would then be able to access the files to make a copy, but we would not be able to make changes or write any new data to the drive.

8. Software When Starting Out.

There is a huge selection of software available to help with digital preservation. In this section we will start by introducing four key pieces of software you will need when starting out in digital preservation, before describing potential next steps, and some advanced options. Links to all software mentioned in the following sections are in the course resources.

9. Software for Copying.

Like using write blockers to stop any changes happening to the digital content on the original storage media, it is also useful to employ a piece of software to handle copying data on to the ingest system.

Copying software will reduce the risks of changes occurring due to human error, and will also check the copy of the digital content to ensure it is identical to the original. This is often done by utilizing integrity checking, so mirrors digital preservation good practice.

On a Windows computer you can use an inbuilt command called "robocopy", or you can use software such as DataAccessioner or Teracopy, which is perhaps the most popular with digital preservationists.

10. Software for Virus Checking.

Virus checking is particularly important if you are receiving digital content from external depositors to avoid the risk of importing a virus into your preservation system. The chances of the content containing a virus are generally quite low, but it is always worth checking to avoid the risk.

If your organization already has a license for virus checking software, you will be able to use that. Otherwise, there are many free or low-cost options to choose from. AVG and ClamAV are two popular free options.

Remember that virus checking software needs to be regularly updated with new virus definitions to make sure it catches any emerging problems. If you keep the ingest workstation offline from the rest of your systems most of the time, it is important to set and carry out a schedule where you do connect the workstation to the Internet for updates.

11. Software for Integrity Checking.

We have already introduced integrity checking in the module “What is Integrity Checking?”, but to recap: it is process that carries out checks on files over time to ensure that no changes have occurred. Thus, allowing us to identify and fix any errors as well as maintaining authenticity.

Checksums are an essential part of the ingest process, and you will need software that can create and compare them. Ideally checksums should be compared whenever digital content is moved or copied. This includes moving from storage media onto the ingest workstation, and when moving from the workstation to long-term storage.

Popular software tools that can be used to generate and/or compare checksums include: DROID with CSV Validator, Fixity, HashMyFiles, and Checksum by Corv.

12. Software for Characterization.

We briefly looked at the characterization process as part of the “What is DROID?” module. It is the process of using a software tool to understand the make-up of a collection of digital content whilst capturing essential technical metadata.

Having characterization tools installed on the ingest workstation allows us to analyze the content when accessioning and make plans for its preservation from the outset.

Options for characterization tools include:

- DROID
- Apache Tika
- ExifTool (for AV content)
- FITS
- JHOVE

13. Software – Intermediate Level.

As you develop your digital preservation work and your confidence grows you can build further on the basic workflows (see module “Introduction to Workflows” for more) with increasingly advanced types of software. At an intermediate level you may consider software for:

- Taking a complete “image” of a hard drive or piece of storage media. This is called disk imaging and potential software includes FTK Imager Lite or BitCurator.
- Encryption of sensitive or private data using a tool like VeraCrypt or Bitlocker.
- Tools that can help find and delete duplicates (deduplication), like CSV Validator and TreeSize Free.
- Playing back content to help with appraisal decisions. There are many free options, such as VLC for video files.

14. Advanced Software.

Moving on further there are more software tools and processes that might be considered to be of an “advanced” level. These might include software for:

- Packaging digital content with its metadata and documentation (e.g. Bagger)
- Validating file formats (e.g. JHOVE or VeraPDF)
- Carrying out file format analysis (e.g. HxD Hex Editor)
- Redacting sensitive data
- Processing particular types of content (e.g. ePADD for preservation email)
- Converting or migrating content to new file formats.

There are also many larger-scale repository systems, available both as open source or from vendors, that will provide functionality to carry out many of the digital preservation actions of the tools mentioned.

15. Wrap-Up.

In this module we have introduced the concept of an ingest workstation, and the combination of hardware and software needed to process received digital content ready for preservation. Also noting that having an ingest checklist helps ensure consistency in the process.

The workstation itself requires a computer, devices for reading various storage media types, and potentially physical write blockers. For software, to begin with we will need tools to manage copying, virus checking, integrity checking, and characterization. There are also numerous other software tools available to help with more advanced processes. All of the software mentioned in this module can be implemented as part of the workflows covered in the “Introduction to Workflows” module.

Now, before we move on to look in a bit more detail at the metadata we might create during the ingest process, let us do a quick knowledge check.