Recordkeeping for Good Governance Toolkit

GUIDELINE 15: Scanning Paper Records to Digital Records

PARBICA

Pacific Regional Branch International Council on Archives

The original version of this guideline was prepared by the Pacific Regional Branch of the International Council on Archives (PARBICA) for use by countries around the Pacific. This means that the guideline may refer to things that you are not familiar with or do not use in your country. You may find that you need to change some of the advice in this guideline to suit your own government's arrangements. To obtain an editable copy of this guideline, contact the national archives, public record office or other records authority in your country, or contact PARBICA at http://www.parbica.org.

Recordkeeping for Good Governance Toolkit

Guideline 15: Scanning Paper Records to Digital Records

CONTENTS

What is scanning? 2

Plan before you scan! 3

Technical standards 5

Equipment 7

Managing digital files 10

Digitisation vs micrographics 13

What scanning involves: A 10-step process 15

Partnership projects 17

Further information 18

WHAT IS SCANNING?

Scanning is also known as digitising, imaging or digital reformatting. It involves making a digital copy of physical or analogue records such as paper documents or photographs. Records are scanned using digital cameras or various types of scanning equipment including flat-bed (like a photocopier) and sheet-feed (where paper is fed through a system of rollers) scanners.

This guidance does not apply to ‘born-digital’ records: records that were created and are kept in a digital format, such as a Word document stored on a shared computer network.

Why scan?

Organisations scan records for both records management and archival purposes. These include:

- **In the course of business**. More and more paper records are being digitised as public offices and organisations increasingly adopt electronic records management systems. Organisations need to make decisions about what to scan (such as the files that are highly used or those which require long-term retention, rather than all records), and when to scan (usually when the records are received into an organisation). A digital copy should be as close as possible to the original, so that the copy can act in the place of the original if required for evidence or business purposes.
- **To improve access**. While an original record is unique and can only be used in one place at any one time, a digital copy can be accessed from a variety of locations, including computer networks and the internet.
- **To provide back-up copies as security**. If the original copy is destroyed, damaged or stolen, the back-up copies allow your organisation to still use the content of the record. A common practice when scanning for archival purposes is to create a high-quality ‘preservation master’ file (a file that is used only for preservation purposes), with separate access copies being produced for day-to-day use, as well as a number of ‘surrogates’ or lower-quality back-up copies.
- **To preserve the original records**. Having digital copies available means that the original records can be kept in stable storage conditions and not handled too often. When scanning for preservation purposes it is important to create a high-quality digital file so that the records do not need to be re-scanned later on.
- **To save physical storage space.** This depends on your organisation being able to dispose of paper originals that have been scanned.

PLAN BEFORE YOU SCAN!

Before starting a scanning project it is vital to give careful consideration to the following questions:

- *What records are you intending to scan and why?* What is the quantity of copying involved? Are you copying text documents or photographs? Will you scan in black and white or in colour? Are there any copyright barriers to scanning? Having clear answers to these questions will help you to answer other questions below. For instance, if they are old, fragile and/or large format records you may need particular equipment such as an overhead digital camera.

- *How well organised and described are the original records?* If they are not well organised or described you will need to invest resources into organising and describing the records before you scan them. It is a waste of money to scan records that are not described and are in a state of chaos. Are there any special handling requirements for scanning the records? For example, do bound volumes need to be taken apart?

- *What hardware and software will be used* to support the scanning process and does it satisfy the technical requirements for producing acceptable digital copies of records? See the following Technical standards section for more information.

- *What file formats will be generated* through the scanning process? Are the formats compliant with open standards (that is they are not based on inflexible commercial/proprietary dependencies), and are these formats suitable for long-term access? See the Technical standards section for more information.

- Scanned records require metadata that describe the records being copied in order to enable access to, and management of, the copied records. Who will create the metadata, what metadata standard(s) will be used, how will metadata be created, where will the metadata be stored and how will the metadata be managed? See the Managing Digital Files section for more information.

- What *quality assurance* mechanisms will be put in place to check the quality of the scanned copies and their associated metadata?

- *What image, document, content or records management software will be used* to store and manage the scanned records? How good is this software? What are its metadata capabilities?

- *What vendor support is available* for the software? Is the software likely to be able to support long-term access to the scanned records?

- *How/where will the scanned records be stored*? On compact disks, internal hard drives, portable/external hard drives, or central servers? How will back-up copies be made and kept?

- *Can the organisation be confident of its ability to preserve the scanned copies* in digital form for as long as those copies will be required? (This may involve migrating the reborn digital records across new generations of hardware and software).

- Will the scanning process only produce page images of the original records, or will it also permit *optical character recognition* of the content of the original records, so that the scanned content can be machine-readable and searchable?

- Will the organisation only scan an identified set of older paper records, or will it start to scan new/current paper records as and when those paper records are received or created?

TECHNICAL STANDARDS

Before you start to scan, your organisation will need to establish technical standards which will determine the nature of the digital files you create. Technical specifications include format (the file type), image resolution (the amount of detail/number of pixels an image holds, often counted in dpi or dots per inch), compression (the reduction of the size of an image file for storage purposes), and colour management. If you don't identify these requirements at the start, you can't know whether the results of your scanning will meet your needs.

Format: Formats used should be based on open non-proprietary standards. Use of such standards will help ensure that your digital copies will be able to be accessed by different software systems over time. The table below shows some different types of file formats, and their suggested uses.

File format	Advantages	Disadvantages	Used for
RAW	Contains all the information acquired during scanning or capture.	Proprietary format – best to convert to DNG (Digital Negative).	Processing – should be used for storage within the DNG file.
DNG (Digital Negative)	New Adobe open standard for RAW format images.		Archiving master images.
TIFF (Tagged Image Format File)	Meets the main requirements for preserving the attributes of the image – resolution, bit-depth, colour management.	Uncompressed TIFF files can take up large amounts of storage space, be slow to download, and be hard to use onscreen because of size.	Provides high-quality master images. Suitable for printing, not suitable for web browser display.
JPEG (Joint Photographic Experts Group)	Smaller files produced, less storage space required.	Image deterioration may occur with JPEGs. May not be suitable where an accurate copy is needed.	Creating access copies – especially for the web.
PNG (Portable Network Graphic)	Often used to display images online. Suitable for editing pictures.	Large file sizes, not widely used.	High quality lossless copies ('lossless' refers to image compression where the compressed file appears the same as the original).

GIF (Graphics Interchange Format)	Black and white, simple graphics.	Contains no printing or dpi information	Low-resolution access copies.
PDF (Adobe Portable Document Format)	Good for text documents.	Less suitable for photographs and images.	Good for machine-readable copies of the content of text documents. Allows keyword searching of the content of documents.

Resolution: The chosen resolution will depend on why you are creating the digital copy in the first place. If the original record is to be destroyed and the digital copy will take its place, a higher-resolution image may be needed to fulfil any legal or evidence requirements.

A resolution of 400 dpi is normally considered the minimum for master image files. Lower resolution files such as thumbprints or surrogate copies are often 72, 90 or 120 dpi. It's important to remember that the quality of any surrogates taken from the master image will depend on the quality of the master.

Compression: Some digitisation devices such as digital cameras produce images using lossy compression – a compression method that reduces the size of the file and loses data in the process (unlike lossless compression, where the compressed image appears exactly the same as the original). Images taken using lossy compression may not be suitable as copies replacing original records, as the image may not be completely accurate. They may, however, be suitable for online display for access purposes.

Bit depth: The bit depth of the image refers to the maximum layers of brightness available in an image. For example, 1-bit image = 2^1 = 2 levels of brightness; 2-bit image = 2^2 = 4 levels of brightness; 3-bit image = 2^3 = 8 levels of brightness; 4 bit image = 2^4 = 16 levels of brightness, and so forth.

Like colour images, black and white images can be represented in 8-, 16- or 32-bit images.

EQUIPMENT[1]

It's easy to feel overwhelmed when you're trying to choose equipment for a scanning project due to the range of digital imaging hardware and software available. Remember that equipment doesn't always have to be sophisticated – many small-scale digitisation projects have been successful using very basic equipment.

Computers

Operating systems: The most popular operating systems for stand-alone systems are Microsoft Windows and Macintosh OS; UNIX is the most common for networked workstations. Increasingly popular are various types of Linux (Ubuntu is probably the most well-known). Your organisation's current computer platform also needs to be considered.

Processor: Image manipulation software requires substantial computing power. A faster processor means a more efficient image manipulation process. The processor should be a recent model to ensure that current image software can run on it and that it will support intensive image editing.

Memory: This is also referred to as Random Access Memory or RAM. Advanced imaging software applications normally require three times the memory as the image file size (for example 30 MB image files require 90 MB memory). More memory may be required if additional software is used at the same time, or if image operations are complex. At least 2GB of RAM is recommended for a scanning project.

Hard disk space: Allow space not only for the imaging software but also for its working files, work in progress and an accumulation of images. Disk space requirements can be substantial, depending on the process used. At least 1Tb SATA HDD is recommended.[2]

Display monitor: This is a key part of the system for image processing and verification. Monitors should be as large as possible, capable of displaying at least 24-bit colour (16.8 million colours), and have a graphic card with sufficient memory. As high-quality images are often captured at a level that is beyond the capability of the display monitors, the most advanced display technology should be used, consisting of large monitors (21" or better), and 24-bit display with a corresponding graphic card. The cheaper LCD (flat), consumer-class monitors are not the best choice as they have many limitations. Only more expensive LCDs are regarded as suitable for professional digitisation.

[1] Adapted from the National Library of Australia's DOHM – Digitisation of Heritage Materials.

[2] A SATA or Serial Advanced Technology Attachment is a storage interface that can connect mass storage devices such as hard drives. HDD is hard disk drive.

Archival drive/storage: Required for archiving and backups. The drive/storage should be an external Hard Disk Drive (HDD), archival tape or CD/DVD, if the previous options are impractical (though the CD/DVD option is not recommended due to the vulnerability of the medium). Alternatively, files may be written to a network drive, though this requires the workstation to be connected to a network drive.

Scanners

Flatbed scanner: The most popular image capture device for flat objects, the flatbed scanner can be used to capture non-transparent objects and some transparent materials such as 34mm slides.

Desirable features

Resolution: minimum resolution of 600 dpi. Higher resolution (2000 – 3000dpi) is recommended if used to scan film transparencies and negatives.
Bit depth: Minimum of 36 bits, but 48 bits or higher recommended (and commonly available).
USB interface: for plug-and play operation (plug-and-play refers to a device plugged into a computer that loads and plays automatically without the need to manually install software).
Transparent media adapter: for scanning transparent materials, such as slides, if required (generally, most flatbed scanners don't give the best results when scanning transparent material).

Film scanner: if the scanning project involves large quantities of transparent materials, a film scanner may be required. A film scanner is more expensive than a flatbed scanner, but will produce a higher quality image. Some film scanners are capable of capturing larger transparency formats, and so they produce very high quality digital images. An alternative to a film scanner is a flatbed scanner with a transparent media adapter (TMA). This method will not produce something as high quality as the film scanner, but may be enough for your needs.

Desirable features:

Resolution: minimum resolution of 2000 dpi.
Bit depth: minimum of 48 bits or higher recommended (commonly available).
USB or FireWire interface: for plug-and-play operation.

Digital cameras

Digital cameras are used like film cameras, but have the benefit of creating immediate digital images that can be reviewed as soon as they are created. High-end digital cameras are more expensive than flatbed scanners, but digital cameras in general are falling in price, while the quality of images they produce is steadily improving. You still get what you pay for, however – if you spend a small amount of money you will get a camera which produces relatively low quality images. While these images may be suitable for a website, they may not be useful for archival purposes. Essentially, the equipment you buy will link back to the purpose for the project in the first place.

Desirable features

Resolution: only high-end digital cameras with a minimal resolution of 10 megapixels should be considered for creating master images. Much image archiving carried out today takes place on cameras ranging from 21 to 60 megapixels.
Lens: fast (f/3.5) zoom lens with macro capability, equivalent of 35mm to 105mm.
Sensitivity: adjustable ISO sensitivity of 50/100/200/400.
White balance modes: manual.
Other: manual exposure, manual focus, RAW capability.

Scanners and digital cameras can be used to capture many of the same formats such as prints, documents, large maps and even film slides or negatives. For some items such as glass plates a digital camera may be more appropriate than a scanner, though a scanner may be better suited for capturing other objects.

Imaging software

Most professional digitisers use Adobe PhotoShop software for image manipulation after processing through specialist scanning software such as Flextight or Oxygen that does most of the work usually done in Adobe PhotoShop. There are several types of freeware and shareware products available on the internet however, and some products support the production of high-quality images.

During the image capture process, little or no image enhancements should be made to the master images created for archival purposes. This will ensure that the consistency of the image capture process is retained and will match the recorded information (metadata).

Imaging software should be able to handle all the necessary manipulation of images that is required. The following are some features to consider:

- parallel work flow – this allows the user to work on images as they are scanned, instead of having to wait until the end of a batch
- 16-bit work flows
- import and export file formats
- convert image file formats
- operations such as brightness/contrast adjustment, resizing
- multiple images open at the same time
- ability to handle large-size images
- multiple ‘undo’ levels
- batch and macro facilities for repeat operations
- ability to save workspace settings to restart work that was stopped temporarily.

MANAGING DIGITAL FILES

Organisations need to manage all their records so that they are able to be found, retrieved and used over time, and digitised records are no exception. Scanning projects often produce a large number of digital files that require ongoing storage, preservation and management. Digital files should ideally be captured into a system – an image management or records management system – along with the necessary metadata to allow users to find and use the files in the future. You will also need to decide how to store the digital copies – for example on a server, a portable hard drive, or on CDs, and where and how back-up copies will be stored.

The speed at which technology changes means that organisations storing and preserving digital files need to think about strategies to keep files usable over time. These might include migrating data to new software and hardware, refreshing of digital storage media, and adopting standards for image files and metadata. Migration of data files, which is often carried out every five to six years, is costly and organisations planning on keeping digital records longer than this should factor additional costs into management. Migration of a digital file into any new system must also include the migration of any metadata associated with the file. See Guideline 18: Digital Preservation, for more information.

Metadata

Each digital record needs to have key metadata elements associated with it to ensure that the object can be managed and retrieved over time. Much of the metadata will be included as part of the item’s descriptive data. Essential or mandatory metadata elements should include the following:

Metadata element	Description
Record number	A unique number identifying the item.
Title	The official name given to the item.
Series	The group of records the item belongs to.
Date	The date the item was created.
Description	What the item is about – often a limited number of characters.
Subject	The subject and topic that clearly describes the content of the item. Subject descriptions should be approved thesaurus terms to aid searching and retrieval.
Format	The data format of the item, eg PDF, JPEG, MP3 etc.
Extent	The file size (in cm) or duration (if a sound recording) of the original item.
Author or creator	The name of the person or organisation primarily responsible for the item.

Content management software

There are numerous off-the-shelf products that have been created to store and manage metadata such as that created when doing a scanning project. Many of these database management systems can be expensive, due to their substantial storage capacity. Smaller systems like Microsoft Access generally have size limitations, but can be more convenient and cheaper if your organisation is not dealing with vast amounts of data.

Spreadsheets are also often used to manage object metadata, but one of the key problems with using a spreadsheet for this purpose is that it is not secure. Unlike a database, it is very easy to lose or amend great amounts of data in a spreadsheet.

Any system used to store and manage object metadata will need to be able to interface or connect with the repository housing the digital object. For this to happen, IT staff need to write a script to make sure that the metadata database 'speaks' to the repository database.

Disposing of original records

Depending on the records being scanned and the purpose of scanning, your organisation may be able to dispose of original paper records after scanning. Original records that may need to be kept after scanning might include those with established national or cultural significance, or those which are required by law to be retained in their original form.

Your organisation should appraise the records to be scanned before starting work, and manage their disposal in line with existing business and legislative requirements. It is recommended that you seek advice from your national archives before deciding to dispose of original records that have been scanned. See also Toolkit Guidelines 7, 8, 9 and 10 for advice on disposing of public records.

Outsourcing a scanning project

Scanning projects can be carried out either by your organisation (in house), by using the services provided by a commercial vendor (out sourcing), or by using a mix of these two options. For some one-off projects, it may be more economical for an organisation to outsource rather than to purchase an expensive scanner. You may also be able to negotiate the loan of equipment from other public offices.

The table over the page lists some of the issues to consider for the two options:

In house	Outsourced
• Original records are always available and are controlled by the organization. • Requires purchase (or leasing) of equipment which, if the project is a one-off, may be difficult to justify. • Requires dedicated and specifically skilled staff. • Skills and quality assurance maintained in house. • Organisation pays for costs associated with technical infrastructure problems. • Greater controls on the security of the record.	• Original records are unavailable to the organisation for a period of time. • Generally requires payment for the cost of scanning, not separate payment for equipment or staffing. • Trained operators can be expected. • Quality control still needs to be carried out by the organization, independent of vendor quality processes. • Vendor pays for costs associated with technology problems that occur during the scanning process. • Involves physical transportation and handling protocols and processes for moving to vendor premises.

DIGITISATION vs MICROGRAPHICS

As technologies improve and become cheaper, the scanning of paper to digital copies is becoming increasingly common and viable, even for small agencies. Another reformatting option used for long-term preservation purposes involves the copying of paper to microforms, creating images about 25 times smaller than the original. These are generally microfilm (reels) or microfiche (flat strips), and the technology is known as 'micrographics'. The table below lists some of the strengths and weaknesses of using both digital and microform technology:

	Advantages	**Disadvantages**
Digitisation	• Highly accessible – can be viewed in multiple locations and be made available over the web • Software and hardware are becoming cheaper (though there are necessary costs involved with management of digital files – see disadvantages) • Copies can be made for security and disaster recovery purposes quickly and cheaply • Digital images can be re-used for a range of purposes	• The costs required to preserve digital images and ensure they are accessible over time may cancel out any short-term savings made by freeing up physical space • Digital storage media do not have long life spans – files will need to be monitored and copied to new media over time • Maintaining and using digital files is totally dependent on electronic systems • Poor organising can affect access: digital images need to be organised with good metadata so they can be found and retrieved easily
Micrographics	• Long-established and proven technology • Can survive for long periods of time when processed and stored appropriately	• Harder to find information. Unlike a digital file that allows keyword searching, microforms need to be searched by the human eye

	- Not dependent on electricity: if necessary, can be read using a strong magnifying glass	- Like paper (and unlike digital images), microforms can only be used at a certain time in one location - The technology is falling out of use, with equipment and industry support becoming more difficult to source - Each additional set of copies is an additional cost - The cost of appropriate storage facilities can be high

WHAT SCANNING INVOLVES: A 10-STEP PROCESS[3]

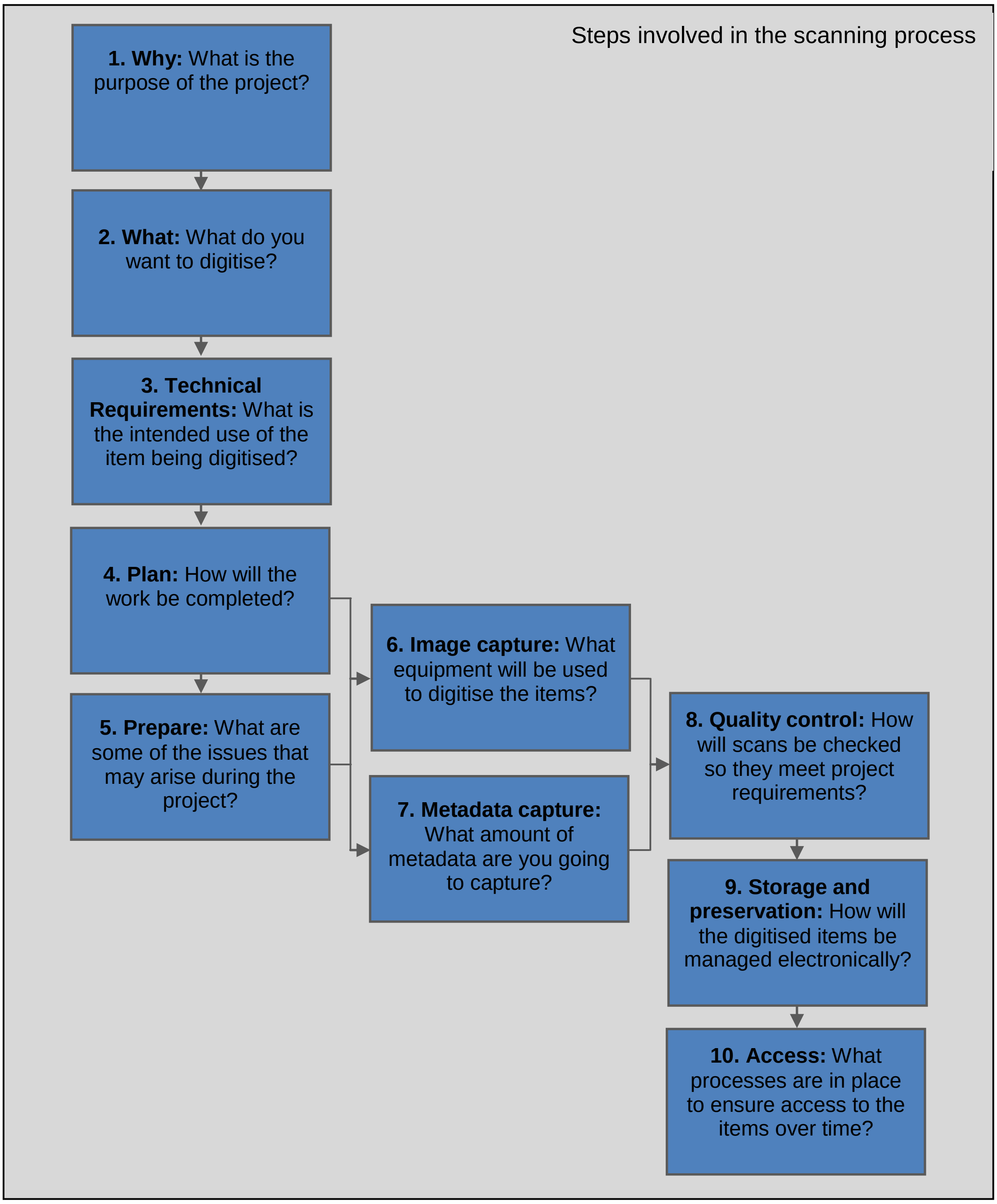
1. **Why?** What is the purpose of the project? What benefits do you hope to achieve?
2. **What?** What do you want to digitise? This will be linked to the purpose – for example, if you are digitising for preservation reasons, you may select records that are physically fragile or in poor condition. The availability of resources will be a factor in deciding what to digitise. Do you have the resources to sustain the work over time?
3. **Technical specifications** – determining technical specifications should take into account the intended use of the document. For example, specifications need to meet expectations about the use of records for evidence or legal purposes.
4. **Plan** – like any project, planning and management decisions need to be made. These include deciding:
    a. who will do the work – in house or outsourced?
    b. whether it will be carried out in association with other activities
    c. defining the cost (staff, equipment and software, impact of overhead costs).
5. **Prepare** – good preparation helps make sure any problems are identified early on, and that the digitization process is smooth. Preparation includes conservation treatments of records, creating instructions for staff, training, moving records.
6. **Image capture** – what equipment are you going to use and why? Do you require additional software to process the images?
7. **Metadata capture** – capturing and maintaining the right metadata will help users find and use the records at a later date. The two types of metadata that are important to capture are metadata specific to the image and the imaging process; and metadata about the record, the business transaction, and the people associated with the transaction. Metadata can be embedded into the object or be in a separate system, but there needs to be a connection between the digital object and the metadata.
8. **Quality control** – like any photographic activity, the quality of scanned images can vary enormously, if the scanning is done for example with inappropriate lighting levels, poor focus, or poor colour contrast. It is vital to build good quality control over scanning processes and end products into the management of the project. This may be as simple as checking a sample of images and metadata at regular intervals during

[3] From ‘What digitising involves’ in *Keeping Archives 3rd ed.*, Australian Society of Archivists, 2008, p. 408.

the project to make sure they meet the requirements decided on at the start of the project.

9. **Storage and preservation** – once captured, digitised images must be able to be stored and retrieved. An image or records management system are recommended, and strategies such as migration and backing up. Storing large master files can be expensive.

10. **Access** – The technical standards used in creating digital images, the technology and software used, and the way a digital file is stored and managed over time, will all impact on how easily the images can be accessed and used in the future.


Steps involved in the scanning process

1. Why: What is the purpose of the project?

2. What: What do you want to digitise?

3. Technical Requirements: What is the intended use of the item being digitised?

4. Plan: How will the work be completed?

5. Prepare: What are some of the issues that may arise during the project?

6. Image capture: What equipment will be used to digitise the items?

7. Metadata capture: What amount of metadata are you going to capture?

8. Quality control: How will scans be checked so they meet project requirements?

9. Storage and preservation: How will the digitised items be managed electronically?

10. Access: What processes are in place to ensure access to the items over time?

PARTNERSHIP PROJECTS

In the Pacific region historically significant records are sometimes copied by organisations in partnership with external partners. For instance, genealogical records are of interest to the Church of the Latter Day Saints based in Utah, USA and also the commercial organisation Ancestry.com. Partnering with these kinds of organisations may be a cost-effective means of copying valuable and at-risk original records for access and preservation. Care must be taken, however, before entering into partnership arrangements of this kind, as they often have 'strings attached' whereby the organisation paying for the scanning requires exclusive use of the copied material for a set number of years. Even without such conditions attached, your organisation will need to consider whether or not there is information being copied that you would rather not share with the world – as usually the aim of the partner organisation is to provide global access to the records in question.

Other possible not-for-profit partner organisations where you will be able to exercise greater levels of control over the copying process, access to and use of the information include:

- *Pacific Manuscripts Bureau*

The aim of the Pacific Manuscripts Bureau is to copy at-risk historical research material in the Pacific Islands for use by academic researchers. Copies made by the Pacific Manuscript Bureau are available to the original custodian organisation and to member research libraries around the world. Original custodians are able to specify conditions for the access and use of material that is copied.

http://rspas.anu.edu.au/pambu/

- *The British Library's Endangered Archives Program*

The Endangered Archives Program is a funding scheme that provides resources for scanning projects which produce preservation copies of archival material at risk of loss or deterioration. The scanned copies produced using the British Library funding are lodged with the original custodian organisation and with the British Library in London.

http://www.bl.uk/about/policies/endangeredarch/homepage.html

FURTHER INFORMATION

Fulton, Wayne ‘A few scanning tips’, http://www.scantips.com/

National Library of Australia, *Digitisation of Heritage Materials*, http://www.nla.gov.au/preserve/dohm/

The Recordkeeping for Good Governance Toolkit is produced by the Pacific Regional Branch of the International Council on Archives with assistance from the National Archives of Australia and AusAID.