

# Digitising accumulated physical records



**A guide to initiating and planning digitisation projects**

© Commonwealth of Australia 2011

Published by the National Archives of Australia  
naa.gov.au

This work is copyright. Apart from any use as permitted under the *Copyright Act 1968* no part may be reproduced by any process without prior written permission from the National Archives.  
Inquiries should be directed to the Government Information Management Branch, National Archives of Australia, PO Box 7425, Canberra Mail Centre, ACT 2610, Australia.

## CONTENTS

INTRODUCTION .....	5
Scope .....	5
Glossary .....	6
International standard for digitisation.....	6
Contact us for further help .....	6
Acknowledgments .....	6
WHY DIGITISE? .....	7
Saving space.....	7
Integrating with current business information systems .....	8
Having better access.....	9
Protecting the records .....	10
PLANNING TO DIGITISE .....	11
Understanding the records.....	11
Types of digitisation .....	12
Large-scale digitisation projects .....	13
Preparing a business case.....	13
Project management.....	14
Documenting the project .....	14
Quality control and quality assurance.....	14
Digitisation equipment.....	15
Digitising in-house or outsourcing.....	16
Intellectual property rights .....	17
SOME DECISIONS ABOUT THE RECORDS .....	18
Status of the records .....	18
Digitising microfilm or microfiche .....	18
‘Packaging’ digitised records .....	18
Managing digitised records .....	19
Storing digitised records.....	19
Digital preservation implications .....	20
Managing source records after digitisation.....	20
Destroying or transferring source records .....	21
Metadata requirements .....	21

THE DIGITISATION PROCESS .....	25
Preparing the source records.....	25
Handling other formats .....	26
Handling digitised records during digitisation .....	27
Partially digitising source records .....	28
Managing masters and derivatives .....	28
Handling security classified material.....	28
TECHNICAL SPECIFICATIONS .....	29
APPENDIX 1: Explanation of technical terms .....	31
APPENDIX 2: Characteristics of scanner technology .....	34
APPENDIX 3: Sources .....	37

## INTRODUCTION

This guide is for information and records managers, information architects and other staff considering scanning or digitising accumulated physical records.

This guide:

- explores reasons for digitising accumulated physical records
- identifies issues commonly encountered during digitisation projects, including records management and handling, and suggests possible solutions
- assists with developing a business case and planning a digitisation project
- ensures the application of appropriate technical standards depending on the value of the records being digitised
- ensures accountable decisions are made about destroying or retaining records after digitisation.

For digitisation projects, project managers will need to refer to the National Archives document, 'General Disposal Authority for Source Records that have been Copied, Converted or Migrated', and identify current technical requirements for digitisation. Further information on technical standards is available from the National Archives.

Appendix 1 explains technical terms, Appendix 2 covers characteristics of scanner technology and Appendix 3 lists sources for more information.

In this guide 'original' or 'pre-digitisation' records are referred to as 'source records'. Records created by the digitisation process are referred to as 'digitised records'.

### Scope

This guide covers the digitisation of accumulated paper or non-digital records for ongoing business use. It also addresses the management of subsequent digitised records.

It does not cover ongoing digitisation of incoming paper documents for incorporation into workflow systems.

**Note:** Digitisation projects are often initiated by business areas in response to a particular business need. In such cases it is important to obtain records management advice, especially during feasibility and planning stages.

## **Glossary**

Terms used within this document are consistent with the [Glossary of records management terms](#) produced by the National Archives.

## **International standard for digitisation**

[ISO/DTR 13028 Information and documentation—Implementation guidelines for digitization of records](#) is currently being developed.

## **Contact us for further help**

For further assistance or for advice on technical specifications please contact the National Archives of Australia at: <http://www.naa.gov.au/records-management/help/index.aspx>

## **Acknowledgments**

This guide was developed for the National Archives of Australia by Barbara Reed, Recordkeeping Innovation Pty Ltd, in conjunction with Rodney Teakle, Mark Semmler, Michaela Olde and other staff of the National Archives of Australia.

A number of Australian Government agencies were consulted or provided comment and their assistance is appreciated.

This guide was developed with reference to national and international professional literature and in particular Archives New Zealand's *Digitisation Standard* and Queensland State Archives' Digitisation Disposal Policy and Implementation Toolkit.

## WHY DIGITISE?

Agencies may have many reasons for considering digitising existing records. These and the expected benefits may influence the processing or subsequent treatment of the records. Some of the common reasons are:

- saving space
- integrating with business information systems
- having better access
- protecting the records.

### **Saving space**

If saving physical space and associated storage costs is the primary reason for the project, the source records will need to be destroyed or transferred to the National Archives after digitisation.

The National Archives of Australia has issued the General Disposal Authority for Source Records that have been Copied, Converted or Migrated (GDA). There are conditions attached to this GDA (see the section on destroying or transferring source records) especially regarding when the source records were created.

If destroying the source records is not authorised under the GDA, you will need to contact the National Archives to discuss whether destroying or transferring them can be authorised. This should occur early in the project or the expected savings in space and storage costs may not be achievable.

#### **Example**

An agency had accumulated 300 shelf metres of records that were identified as 'long-term temporary' records. They were located in expensive office space because of access requirements. However, the space was increasingly needed to accommodate staff. A digitisation project was planned to reduce physical storage needs and to meet access requirements for the records. Following the successful completion of the project, the office was refurbished for other use.

## **Integrating with current business information systems**

If you intend to integrate digitised records into current business information systems, consider these factors:

- Does the current business process incorporate electronic records and documents that are equivalent to those being considered for digitisation? Is the intention to create a seamless process incorporating both the current and older records? Or is a separate process acceptable to retrieve the digitised records?
- Will the current business process need to be redefined? Are there opportunities for change or upgrade due to the records being available digitally?
- What information (metadata) from or about the digitised records will be needed to link the digitised records to the business information system? This may be something simple like a case file number or client name. It may be picked up automatically during scanning if it is always located in the same place on the source records.
- If the information needs to be manipulated, or an enhanced search capacity is needed, then optical character recognition (OCR) processes should be used.<sup>1</sup> Metadata will need to comply with appropriate standards and be constructed to provide links to relevant fields within the current business information system. For further information see the section on metadata requirements (p. 21).

---

<sup>1</sup> Optical character recognition (OCR) is the process of translating images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text.

## Having better access

If the main reason for digitisation is to provide better access to the records, the impact on the users will be a central issue in the planning. These are some of the questions that you will need to consider:

- Are the users within the agency or government, or are they members of the public? This will affect decisions on whether the existing retrieval mechanisms are adequate or whether other mechanisms will be needed. See the section on metadata requirements.
- Is the equipment adequate for achieving the expected benefits? Retrieval rates, the speed of interaction between networks and the display quality and performance of computer terminals all affect the user experience.
- What level of quality will be acceptable? Different levels of quality are available in digitisation processes, particularly those that convert the text into editable format using OCR technology.
  - If detail and precision are a key requirement, additional care will be needed to ensure that the technical quality of the digitised records is high and there should be greater attention paid to quality control and assurance practices.
  - If the aim is to access information quickly, projects should focus on using OCR technology that facilitates free text searching.
- You should also consider accessibility requirements and ensure that the formats are appropriate. For further information, see the [Web Content Accessibility Guidelines \(WCAG\) 2.0](#) which was adopted by the Australian Government in February 2010.

### Example

An agency is conducting a project to convert paper-based seismographic records to digital format. These records are precise instrument recordings and their meaning is conveyed through relative positions on scaled grids. Imprecise, skewed or fuzzy digitised records will render them meaningless. Ensure digitised records are of a high enough quality and that they still convey accurate information.

## **Protecting the records**

Many records targeted for digitisation have long-term business use. These records may have deteriorated and become fragile. Digitisation can be a way of protecting them by reducing ongoing handling. The fragility of the records will need to be considered when planning the digitisation. Digitising also mitigates the risks associated with having single copy physical records.

### **Example**

A land information-based agency has maps and plans dating back to the earliest European settlement. These are still used for checking boundaries and property rights. However, they are old and fragile and are being damaged by physical handling. The agency is planning to digitise the maps and plans to a high quality and then to restrict access to the physical records.

## **PLANNING TO DIGITISE**

### **Understanding the records**

Understanding the source records is vital for making appropriate and cost-effective decisions about digitisation projects.

#### ***Purpose and role of the records***

The purpose and role of the source records will inform decisions on the purpose of digitisation, the likely cost savings, the appropriate disposal (destruction or transfer) and the degree of integration into current systems. Questions that should be answered include:

- Why were the records created?
- Are they being used for the same purpose now?
- If not, how does the business process which created them differ from the one used now?
- Why are they still needed in the current business process?
- Will digitising the records add value to the current business process? How?

#### ***What controls supported the source records?***

Records need to be analysed in the context of their controls, such as registers and indexes, and also in the context of other information about how the records were created and managed.

#### ***Media and formats***

The physical composition and state of the source records will affect issues such as security, physical handling, the scanning equipment needed and possibly the structuring of the project.

You will also need to know the quantity and condition of the source records and their current and anticipated use.

### **Example**

A registration agency has a set of records dating from 1909. The earliest registration sheets are handwritten with ink annotations in different colours on foolscap paper. Some of these early records are frayed and damaged from handling. From about 1920 the registrations are typewritten sheets using quarto-sized paper folded into packets and tied with string. From the 1940s the registrations were attached to correspondence with metal pins and kept in paper files. They were recorded on paper of varying quality, including now fragile paper used for carbon copies.

It might be possible to split this project into three separate jobs, each with particular document preparation techniques and each suited to different scanning equipment. The early registers would be suited to flat bed scanners and more robust papers could be fed through document scanners. Different conservation requirements for different formats might also be a factor. The potential to automatically capture metadata is likely to be affected by the different physical formats and how the records were accumulated.

Note: If the project were split into three separate jobs, it would be important to ensure that file order was maintained when the files were reassembled. This would need to be considered when planning the workflow.

### **Types of digitisation**

There are two different types of digitisation or scanning:

1. Creating a digital photocopy or image of the source record—the digitised record is static and cannot be changed and the content cannot be manipulated.
2. Optical character recognition (OCR)—this ‘translates’ the digitised record into machine-readable text, which can then be changed or manipulated.

While these two processes are often performed at the same time, they are quite different.

OCR technology can achieve high quality results with most kinds of text especially those with a typeface. However, it may not be very accurate with other text such as handwritten text. Quality assurance processes need to be built in to verify the resulting text. The level of quality assurance will depend on the aim of the project and the format used. See the section on quality control and quality assurance for more information (p. 14).

## **Large-scale digitisation projects**

Digitisation projects involving large volumes of records can be extremely costly and resource-intensive. If you break down big projects into smaller components based on clearly defined parameters you will be able to:

- achieve smaller objectives more quickly—it can be frustrating to have very long projects without measurable achievements
- learn from experience—digitisation projects rarely run as smoothly as planned. Identifying and resolving the issues is part of a learning process that can be applied to subsequent parts of the project
- optimise equipment and resources—different formats or ages of records may dictate different methods of capture of the digital image (such as flatbed scanners or automatic feeders). Segmenting projects in this way means resources can be used more efficiently.

Issues to consider include:

- How long can the business afford to have the records unavailable?
- Can source records be digitised on request and out of sequence? If so, do your procedures ensure that the exceptions are managed effectively and integrated into the project?

## **Preparing a business case**

When you have a thorough understanding of the business requirements and the records themselves, you should prepare a business case. The business case should describe the records and their relationship to other records. It should also outline the benefits and costs of digitisation. It will provide the basis for obtaining approval for the project.

The business case should clearly address:

- the purpose of the digitisation project and likely benefits
- project planning and management, including quality controls and assurance procedures
- records management issues associated with both the source records and the digitised records
- decisions about whether to keep, destroy or transfer records, ie sentencing

- requirements for handling records
- appropriate technical specifications.

### **Project management**

Digitisation projects need to be managed like any other project. Most agencies have a preferred project management method and supporting software.

A digitisation project requires more flexibility than many other projects and many project management methods may not provide the flexibility needed. The project management approach used may need to be modified.

### **Documenting the project**

All digitisation projects should be carefully documented. The degree of documentation depends on whether or not the digitised records will replace source records which will then be destroyed. In this case a higher standard of documentation is required. You should assess the risks to determine the documentation requirements.

If the source records are to be destroyed, documentation of decisions about the digitisation process and specifications is an important safeguard against potential challenges to the authenticity of the records. In such cases, the National Archives suggests that you consider the following guidelines and adapt elements from them as appropriate:

- Public Record Office Victoria, PROS 10/02/S1: [Digitisation Requirements v 1](#), Public Record Office Victoria, 2009, pp. 9–11.
- Archives New Zealand, [Digitisation Standard](#), Archives New Zealand, 2007, Appendixes 1, 3 and 7.

### **Quality control and quality assurance**

A quality control statement sets out the level of quality that the digitisation is to attain. Quality assurance is the process of verifying that the level of quality has been achieved.

The levels of quality control and quality assurance depend on whether the digitised records will replace the source records. If so, then higher standards of quality control and quality assurance are needed.

Quality controls need to be devised and documented for each part of the digitisation process and for both external providers and in-house projects.

Quality control should address:

- the accuracy of the digitised records such as:
  - operator training
  - data verification (for OCR processes)
  - extent and frequency of sampling
  - criteria for checking digitised records
  - acceptable variations
- storage reliability
- processes that deal with failures in the quality of digitisation
- logging and analysis
- physical integrity of the records—ensuring the record components are still in order.

If digitisation processes are outsourced, agencies will need to assure their quality to confirm they meet the contractual obligations. Quality assurance should be carried out early in the project to identify any issues before a large volume is processed and then periodically throughout it.

Further information can be found in Appendix 7 of Archives New Zealand's [Digitisation Standard](#), 2007.

### **Digitisation equipment**

Purchasing high-quality scanners may be an expensive option for one-off projects. Engaging specialist service providers may be more cost-effective if digitisation requirements are of limited scope, quantity or duration. Other government agencies may be able to provide advice about specialists and may even be able to provide occasional access to spare scanners.

It is unlikely that a low-end office scanner will meet the demands of quality and quantity.

It is important to consider the occupational health and safety of staff when selecting and installing equipment, especially the way the equipment is set up.

The speeds and capabilities of scanners published by vendors are often not a good indication of the actual time required for processing. To make accurate estimates you should undertake trials on the records to be processed.

Further information on scanning equipment can be found in Appendix 2.

### **Digitising in-house or outsourcing**

Either the agency itself or a commercial provider can undertake a digitisation project. There are issues to consider for both options. In some cases a combination of the two might be best—for example using an external service provider on agency premises.

When outsourcing a project you should be careful to ensure that security handling requirements are met. See the section on handling security classified material for more information.

Some issues to consider when deciding between an in-house or outsourced digitisation process are outlined below.

In-house:

- records are continuously available and under agency control
- requires purchase or leasing of equipment which may be difficult to justify (such as for a one-off project)
- requires dedicated and specifically-skilled staff
- requires initial and ongoing training of specifically-skilled staff
- allows enhanced control over application of metadata
- skills and quality assurance are maintained in-house
- allows greater control over the security of the records
- quality controls can be easily adjusted as issues are identified
- the agency must bear costs associated with technical infrastructure problems such as network downtime or equipment failure.

Outsourcing:

- records are unavailable to the agency for a period
- payments for scanning images are all-inclusive
- high production levels or volumes are available
- trained operators can be expected
- vendor absorbs costs of technology upgrades, failure or downtime

- involves physical transportation and handling protocols and processes for despatch of records to the vendor
- the agency needs to maintain quality controls and assurance processes independent of vendor quality processes
- complex contractual arrangements are needed to specify standards, security controls, quality requirements, communications, variations and how to resolve problems
- project brief and specifications need to be clearly articulated.

### **Intellectual property rights**

The rights to intellectual property are generally the same for digitised records as for source records and no additional consideration is necessary. However, intellectual property rights may need to be considered if the digitisation of the records has increased their use, particularly when records are made publicly available.

In some cases agencies may be dealing with records created by other bodies gained through machinery of government (MoG) changes or administrative re-arrangements. It may be necessary to clarify who has control of the records before any digitisation project starts.

## **SOME DECISIONS ABOUT THE RECORDS**

### **Status of the records**

The intended use of the digitised records will affect:

- the disposal status of the source records
- the quality and standards that should be applied in the digitisation
- the methods that should be employed for managing the source records after digitisation.

For more information see the section on managing source records after digitisation.

If the digitisation is mainly to create a reference copy for ease of access, the source records are unlikely to be destroyed and will still exist as the authoritative version. In this case, a lower-quality process may be acceptable.

However, if the source records are likely to be destroyed the digitised records will be the only official record. You will therefore need to apply more stringent format standards and quality controls.

When the digitised records will be retained permanently by the National Archives, higher standards and quality controls should be applied. You need to ensure that the digitised records are compatible with the National Archives' long-term storage requirements and digitisation standards, so contact the National Archives for specific advice. In the case of some temporary value records there may be other reasons to apply higher standards.

### **Digitising microfilm or microfiche**

In some cases records may have been converted to microfilm or microfiche and the original source records may or may not still exist. Digitising the microfilm or microfiche version may be an effective technique. Or it may be that the initial conversion project was done to lower standards than those of today. If the original source records exist, it may be better to use them as the source. When this option does not exist, you may have to accept a lower-quality result from the digitisation of the microfilm or microfiche.

### **'Packaging' digitised records**

If digitised records are intended to replace source records that will be destroyed, you should consider how closely the physical representation of the source records needs to be reproduced. For example, paper files may contain booklets, pamphlets and other multi-page items, and you may want the digitised record to reflect this. You may choose to use formats that allow images to be packaged as multi-paged images.

Similarly, images should be stored in systems that represent the physical order of documents in files. This will enable metadata to be transferred between formats.

- File and document naming conventions can be used for representing the physical order. (For example, each document might be called 12345-001, 12345-002 et cetera to represent individual documents on file 12345).
- Directory structures can be used to recreate file structures. (For example, 12345 is the parent structure containing documents 001-250).

### **Managing digitised records**

The type of system used for ongoing management of digital records depends on how they will be used. For example a set of images for reference purposes would need at least some form of indexing to assign and maintain links between the image and a retrieval point.

When the digitised records will be incorporated into a business information system the management protocols of that system will be applied to the digitised records. Such systems will offer much greater functionality for ongoing management of the images. Decisions on how to integrate the digitised records into the business system are complex. While such decisions could be separated from the actual digitisation project, they should be considered during the planning stage as they may have an impact on:

- the metadata extracted from the digitised records as part of processing
- the formats and specifications of the digitised records
- the extent to which the digitised records can be changed
- whether the digitised records in the new system should be considered 'new' records. If the records have been substantially changed in the way they are presented and are considered to be 'new' records, any previous decisions about whether or not they can be destroyed should be reviewed.

If the digitised records are regarded as the functional replacements for the source records, the business system should be able to maintain the representation for each record as it was when digitised and any subsequent modifications within the business system.

### **Storing digitised records**

Digitised records can consume considerable storage space. File compression may be an option—contact the National Archives for more advice. Various versions of digitised records will be produced during the process (such as a raw version, an enhanced version, a quality checking version and a final format version). While all but the final can

be discarded soon after the quality assurance process is completed, this still consumes storage space. This space must be available during the production process if it is done in-house.

If the source records are relocated for digitising you should take care to protect them in transit, in temporary storage and during processing. Temporary storage should conform to the [Standard for the physical storage of Commonwealth records](#) produced by the National Archives. The security requirements of the Protective Security Policy Framework should be followed at all times. See the section below on handling security classified material for more information. You will also need to consider security requirements for the digitised records, the equipment and the storage media used.

Options for the ongoing storage of digitised records include:

- a dedicated server
- magnetic tape
- WORM (write once, read many) storage media such as CD or DVD.

If your agency does not have a digital preservation plan, the best approach to ensure that digital records survive is to maintain them in network storage wherever possible. Further information can be found in the section on digital preservation implications.

### **Digital preservation implications**

Once the records are digitised, you will need to ensure that the software, formats and media are all maintained to facilitate access to the digitised records.

A digital preservation management plan will help ensure a systematic approach to accessibility and reliability of all digital records. For more information see [Digital Recordkeeping: Guidelines for Creating, Managing and Preserving Digital Records](#).

### **Managing source records after digitisation**

Decisions on managing source records after digitisation depend on whether destruction of the source records is authorised. See the section on destroying or transferring source records for more information.

All source records will be needed until quality assurance processes are complete. Further detail is provided in the section on quality control and quality assurance. Individual items will need to be easily retrieved in case they need to be re-digitised to meet the quality requirements.

If destruction of the source records is not authorised the records will need to be managed in ways that preserve their integrity. This will usually involve reconstructing

the source records to the way they were before digitisation. If the records have been dismantled as part of preparing the documents and they need to be retained permanently as part of the collection at the National Archives, you should contact the National Archives for advice about repackaging the physical records. See the section on preparing source records for digitising. You should discuss this with the National Archives before starting the digitisation project as it may affect the digitisation process.

As a general safeguard you should consider keeping all source records until the project is finished. This will allow you to revert to the older system in the unlikely event that it becomes necessary.

### **Destroying or transferring source records**

Whether or not source records can be destroyed is fundamental in determining how a project will proceed. This should be established early in the planning stage as it will significantly affect other decisions.

The National Archives of Australia is responsible for authorising the destruction of records. It does this by issuing legally-binding documents that are known as 'records authorities' or 'disposal authorities'. In the case of source records for digitisation projects, the authorities that are most likely to apply are:

- the [General Disposal Authority for Source Records that have been Copied, Converted or Migrated](#)
- a records authority that has been issued specifically for your agency.

All agencies can use the General Disposal Authority for Source Records that have been Copied, Converted or Migrated provided a number of conditions have been met. It is essential that you consult the authority and that you destroy records only if the conditions are met.

If the General Disposal Authority does not apply, source records can only be destroyed with approval from the National Archives in the form of a records authority issued specifically for your agency. The records management staff of your agency can advise you about the disposal provisions relating to your agency's records.

### **Metadata requirements**

The capture of metadata should be as automatic as possible as it is more accurate and less tedious than manual entry. However, this is not always technically possible. Techniques such as bar coding of documents and linking the barcode number to the original control records might be an interim option for linking images with a control system.

OCR technology offers greater possibilities for automatic metadata capture. Automatic capture of key fields might be possible, especially when source records use a standard format or template. However, all data gathered by OCR technology will need to be quality assured to verify the accuracy.

The comprehensiveness of the metadata depends on whether the digitised records are reference copies, or whether they are intended to reproduce the functionality of the source records and therefore act as authoritative records of business action. If the latter, then metadata will need to be more comprehensive. If the source records continue to be the official records of business and the digitised records are for reference purposes only, then simple metadata may be adequate.

Determining metadata for digitised records will depend on three issues:

- the requirement to manage the digitised records
- the extent of integration into current business systems
- the *Australian Government Recordkeeping Metadata Standard (AGRkMS)*.

Each of these issues is outlined further below.

### ***The requirement to manage digitised records***

Metadata will be needed to link the digitised records to some form of control. At a minimum, it needs to link an image to some basic registration information such as a number. Ideally this number should be linked to its previous (non-digital) identification number so that the original control records (such as indexes and registers) can still be used to retrieve the records. If a minimalist registration system is used for the digitised records then the original control records will continue to be the major retrieval tool.

If the digitised records are managed as a separate set of resources you should consider using a content management system to provide enhanced control and retrieval.

### ***The extent of integration into current business systems***

Digitisation is often undertaken to make the information in source records more accessible. Integration with a current business system usually provides this enhanced accessibility. In this case, the business system will provide the controls and retrieval points needed. At a minimum, the metadata from the digitised records will need to provide a link to the business system.

A greater degree of integration could be considered. This will involve extracting more metadata from the digitised records, often using OCR technology. In this case, make sure that the syntax (form) and semantics (meaning) of the metadata from the digitised records match those of the current system. This needs careful analysis and mapping

between the fields in the current system and the data in the digitised records. The metadata extracted from the digitised records will then be imported (usually through an xml schema) into the business system.

If the digitised records are intended to replace the source records that will be destroyed, greater attention is needed to preserve the features of the source records that prove their authenticity.

### ***The Australian Government Recordkeeping Metadata Standard***

Recordkeeping metadata describes information about records and the way they are managed and used. Adding metadata to documents assists in establishing their authenticity by demonstrating how they were managed. The [Australian Government Recordkeeping Metadata Standard](#) (AGRkMS) issued by the National Archives should be applied when digitised records are intended to replace the source records as the authoritative evidence of business actions.

### ***Metadata for digitisation***

The digitisation process itself needs to be recorded in the metadata because it is a records management process. In the AGRkMS, the process of digitising is recorded as a relationship between the item (the digitised record) and the agent (the person or organisation associated with the digitisation process).

The metadata should record details about the equipment used. This metadata is automatically created from the scanning equipment and includes information on:

- the capture device
- calibration settings
- the date of last calibration.

### ***Metadata for each digitised document***

The AGRkMS has seven mandatory properties which should be applied to all digital records. These are:

- the category (in this case 'item')
- a unique identifier
- the name of the document
- the date and time the digital record was created—this is not the same as the original creation date of the source record which may also need to be included

- disposal information for the record if available including the records authority, disposal class number, disposal action and disposal trigger date. This information may be inherited from the item which contains or 'packages' the individual image. See the section on 'packaging' digitised records
- format (drawn from authoritative format registries such as the Global Digital Format Registry (GDFR)<sup>2</sup> or PRONOM<sup>3</sup> including the name and version of the creating application
- extent (file size in bytes).

The AGRkMS also has four conditional properties that must be applied to digitised records under certain circumstances:

- security classification and caveats—if not 'UNCLASSIFIED'
- a rights statement—if policies governing the use and access to the record exist
- the language of the record—if not English
- an integrity check (or checksum)—if the digital record is transferred between systems, including transfer to the National Archives.

---

<sup>2</sup> Global Digital Format Registry <http://www.gdfr.info/>

<sup>3</sup> The National Archives (UK) <http://www.nationalarchives.gov.uk/PRONOM/>

## THE DIGITISATION PROCESS

### Preparing the source records

Document preparation is a major component of every digitisation project. You will need to assess the condition of the source records to determine whether they can withstand the physical handling involved in digitisation.

They need to be adequately managed, tracked, replaced (if necessary) and you need to make sure that the number of records received back is the same as what went out.

You will need to define business rules for document preparation. Commonly encountered areas needing rules are:

- the management of fragile or damaged source records. Not all source records will withstand the physical handling required for automatic document feeds. If records are fragile or damaged consider alternative equipment (see Appendix 2 on characteristics of scanner technology). Some records may require minor remedial treatment to enable them to be passed through a scanner. Alternative methods might be to photocopy the fragile pages and scan the copies or to enclose them in clear protective covering to enable them to withstand the digitisation process. Care should be taken not to damage or compromise source records that will be retained. You should consult a professional conservator if remedial treatment is required
- removing papers from bindings such as staples, file clips or paper clips
- managing loose items in ways that ensure the original order of the records is maintained
- aligning pages to enable automatic feeding into high-speed scanners
- dealing with adhesive notes, white out, blank pages, faded, torn or illegible pages and reverse pages not relevant to the file
- order of digitisation—from front to back of file or back to front.

Testing is recommended before you decide what method and equipment to use. If automatic document feeders (ADFs) are used for original records, check how the pages feed through and whether paper jams can be removed without damaging the originals.

The extent to which the original source records can be dismantled to enable digitisation depends on their format. If the source records are to remain as long-term records, it might be necessary to reassemble or rebind them. Consult the National Archives about

an appropriate form of repackaging. The aim will be to retain the integrity of the source records as evidence of business.

### **Handling other formats**

Not everything in an analogue form is suited to digitising using standard document-style procedures. For example, there might be enclosures on files containing objects such as maps, plans, drawings, audio, video, photographs, samples or specimens. You will need rules for dealing with these.

If a sequence of paper files contains material known to require different handling, a separate process might be needed. For example, x-rays in medical files will need a different process from paper documents.

The reliability and authenticity of records depend not only on the content being digitised but also on evidence of all processes. It is important to capture all the annotations, comments and information that relates to the source records. This could include annotations or notes on the backs of pages, captions on photographs or comments on envelopes that contain photographs. You might need a process to ensure that all the information is digitised. For example, in sets of source records where some pages have annotations on the back and some don't, it would be best to have a check box or other indication of quality assurance to assure users that the backs of the pages not copied were not merely overlooked but excluded from copying because there was no content.

Decisions on how to package the images should be based on how to best reflect the source records. For example, multi-page images could be used to capture both the front and back of pages.

There should be guidelines for document preparation (see the section on documenting the project). In establishing these guidelines, you should be aware that the unexpected will inevitably arise. You should allow for flexibility and the ability to revise document preparation procedures. Having well considered business practices will help with this.

If in doubt, contact the National Archives for advice.

### **Example**

An agency received approval to digitise and OCR microfiched records. The testing was done and the project methodology established. During the project, however, it became clear that the original microfiching had been done on cameras supporting different specifications. Some of the records were 'grid locked', which fixed them onto a particular position with regular spacing between documents on the fiche. Others, however, were manually placed. Inevitably, a proportion of the manually placed documents were skewed or overlapping. This made a significant difference to the digitisation process as only the gridlocked fiche could be automatically processed. The manually-placed documents needed manual intervention. The original project specifications did not identify this problem until production, adding significantly to the resources required.

### **Handling digitised records during digitisation**

During digitisation, versions of the digitised records are created for processing purposes. The initial capture of the digitised records is usually done in a proprietary format used by the scanner. These are known as RAW files and are subsequently converted to the preferred format. During that process various validation checks on the quality of the digitised record occur.

Options also exist in common scanning software for enhancing the digitised record, using techniques such as 'sharpening' or 'clipping' of highlights or shadows, 'blurring' to eliminate scratches, 'deskewing' to straighten pages and 'spotting' or 'despeckling' to touch up specific areas.

Depending on earlier decisions about whether the digitised records will be the continuing record of business, the business rules for enhancement should be clearly documented:

- If the digitised record is to become the record of business, this documentation is critical in proving the authenticity and integrity of the conversion process and therefore the reliability of the record as evidence of business.
- If the digitised record is only for reference and is not intended to replace the source record, this documentation is less critical to maintain. Rules will still be needed for practical implementation purposes.

If digitised records cannot be enhanced because they are intended to replace the source records, copies could be made to allow for enhancement.

## Partially digitising source records

A thorough understanding of the records will help you decide whether the whole set of records or just some should be digitised. Partial digitisation options might include:

- chronological—digitising from a certain date
- digitising only certain formats such as all bound registers
- digitising only key documents.

Chronological or format-based approaches are relatively straightforward. Such projects are not uncommon and some agencies may already have a portion of the source records microfilmed or microfiched from previous projects. You will need to base decisions about disposal on the source records as a whole.

Where a ‘key document’ approach is taken, the digitised version is only a portion of each source record. It removes the key documents from their original context, thereby changing the nature of the record. The partially digitised record will become a ‘new’ record in the context of the ongoing business. Consequently, disposal decisions on the source records may be affected. This approach is also likely to involve greater document identification and preparation as files may have to be dismantled. If the source records need to be kept, the documents will have to be replaced correctly in their original context, again requiring physical resources and a quality assurance process.

### Example

An agency has a series of case files relating to customer service. It has decided to digitise only the initial service contract and not the records of contact or queries that have also accumulated on the paper files.

## Managing masters and derivatives

Digital masters of source records are a sequence of digitised images that most closely represent the original source records. Master files should be held separately and should be unalterable. They should be stored as uncompressed files.

Master files are used to enable further productions of derivative files which may be in a lower-quality resolution or made to enhance their capacity for distribution over networks.

## Handling security classified material

Obligations for dealing with security classified material are set out in the [Protective Security Policy Framework](#) (PSPF), issued by the Attorney-General’s Department.

## TECHNICAL SPECIFICATIONS

The technical specifications that apply to the digitisation process will depend on whether the digitised records are to replace the source records as the record of business, and whether the digitised records need to be retained long-term.

Project managers should contact the National Archives to identify appropriate technical standards for digitisation. This is essential if the records being digitised require long-term retention. Issues to consider are:

- file formats
- resolution
- type of image
- colour resolution or bit depth
- colour management
- compression.

The recommendations for technical specifications are based on the following general principles to help you select formats that will optimise the chances of formats being readable in the future.

- Preferred formats are:
  - open source (non-proprietary) formats
  - those which have publicly available technical specifications
  - those which are widely used within the public sector.
- Formats that contain embedded objects or which link to external objects beyond the specific version of the format should be avoided.
- Formats that are supported by several software applications or operating systems should be used.
- Formats used should be readable by a readily available viewing plug-in if the specific production software is not available to all users.
- Technical specifications should be supported by a body of accessible and product-independent technical expertise.

- Adequate technical support should be available for ongoing maintenance and assurance of migration capability.

If the source records will be retained and the digitised records are for reference or convenience purposes or are not intended for long-term retention, it is not essential to adhere to such high technical standards.

## APPENDIX 1: Explanation of technical terms<sup>4</sup>

Area	Issues
Formats	<p>Categories of file formats are:</p> <p><b>raster:</b> also known as bit-mapped formats, where images take the form of a grid or matrix with each picture element (pixel) having a unique location and independent colour value. Examples include TIFF, JPEG, GIF and PNG</p> <p><b>vector:</b> also known as object oriented, based on a set of mathematical instructions typically used by drawing programs to construct an image—not relevant to digitisation that will use raster formats</p> <p><b>multi-page or encoding:</b> for metafiles which may contain either vector or raster images. Such formats enable the contents to be consistently displayed and used across different computer programs and operating systems. Typically they support internal metadata and multi-page images and enable security management. Examples include Adobe PDF</p>
Resolution	<p>This is a measure of the ability to capture detail in the original work, often measured in pixels per inch (ppi). The optimum resolution depends on the nature of the documents being scanned. Photographs, for example, require much greater resolution.</p> <p><b>ppi:</b> (pixels per inch) is a measurement of resolution for computer display</p> <p><b>dpi:</b> (dots per inch) is often used interchangeably with ppi, but actually refers specifically to measurement of the resolution for computer printers</p> <p>Since raster images have a specific resolution such as a specific number of pixels per inch, scaling a raster image involves the distribution of available pixels across the designated space. Image resolution subsequent to scaling is</p>

<sup>4</sup> This section is based on Appendix 5 of the Archives New Zealand [Digitisation Standard](#), 2007.

Area	Issues
	<p>referred to as effective resolution. If an image is enlarged, then unless additional pixels have been added by means of interpolation (re-sampling), the size of each pixel must be increased accordingly. Consequently, the enlarged image will have fewer pixels per inch (lower resolution). When an image's physical dimensions are decreased, its resolution increases.</p> <p>The effective resolution formula is:  [actual image resolution] / [scaling percentage] = [effective resolution]  Example: 150 ppi / .75 = 200 ppi</p>
Type of image	<p><b>bi-tonal:</b> presents each pixel as either black or white</p> <p><b>greyscale:</b> presents each pixel as black and white and a range of intermediate greys</p> <p><b>colour:</b> presents each pixel by assembling three components corresponding to the colours red, blue and green</p>
Bit depth	<p>This is a measure of the number of colours (or brightness in greyscale images) available to represent the colours (or shades of grey) in the original document.</p> <p><b>1 bit, black and white</b> or line art: only black and white pixels</p> <p><b>8 bit:</b> uses 8 bits to describe each pixel as black and white and a range of intermediate greys (greyscale)</p> <p><b>24 bit colour:</b> a resolution that enables storage of 8 bits of information describing the red, blue and green components of every pixel, thus enabling a much greater palette of colours</p> <p><b>36/48 bit RGB colour:</b> uses an extended colour space, creating a much larger file and requires storage in formats that explicitly support this colour depth (TIFF or PNG)</p>
Colour management	<p>Colour management is required because every different device produces or reacts to light differently. Each device therefore needs to have different colour values to produce similar results. The goal of colour management is to provide a system that guarantees that images look the same</p>

Area	Issues
	<p>wherever and whenever they are viewed.</p> <p>The International Colour Consortium (ICC) colour management system uses a standardised and known 'colour space' based on the human eye and then compares all devices within the workflow to a known standard.</p>
Compression	<p>This involves algorithms designed to reduce the size of the image for storage or transmission. Various options exist but decisions should be made based on the characteristics of the document to be captured. The two major categories of compression are:</p> <p><b>lossy:</b> where information is removed from the stored information during the compression process. Note: there are different levels of lossy that can be selected.</p> <p><b>lossless:</b> where no information is irretrievably lost and where the decompressed object will always appear exactly the same as the original. Examples include LZW or ZIP lossless compression with TIFF files.</p> <p>Newer forms of compression are fractals and wavelets.</p>

## APPENDIX 2: Characteristics of scanner technology

Scanner technology changes rapidly. The characteristics listed below explain critical features to consider when selecting a scanner, although decisions on the appropriate standards depend on the nature of the digitisation being undertaken.

Characteristic	Explanation
<b>Style</b>	<p>Various styles of scanners are:</p> <p><b>Flatbed scanners</b> provide a flat glass area for scanning. Some more specialised flatbed scanners may have bevelled edges especially suited to volumes. They may have adapters to enable sheet feeders or transparent media adapters suited to fragile documents or books to be scanned.</p> <p><b>Sheet feed scanners</b> scan separated pages, typically at a higher rate than flatbed scanners. They are designed for high throughput and are not suited to fragile material or non-standard sized material.</p> <p><b>Slide scanners</b> are designed for digitising transparent material such as slides and negatives.</p> <p><b>Drum scanners</b> are used in high-end design and publication applications producing a high quality image.</p> <p><b>Wide format scanners</b> are used for scanning maps, plans and larger documents that are manually fed.</p> <p><b>Overhead scanners</b> are used for specialised scanning typically involving fragile material or material of archival value that cannot be laid flat such as volumes of newspapers or ledgers.<sup>5</sup></p>

<sup>5</sup> Queensland State Archives, Digitisation Disposal Policy and Implementation Toolkit, September 2010, <http://www.archives.qld.gov.au/government/ddp.asp>

Characteristic	Explanation
<b>Resolution</b>	<p><b>Optical resolution</b> is the actual resolution achieved by the scanner using the in-built sensors.</p> <p><b>Interpolated resolution</b> is software-enhanced resolution that increases the perceived resolution of the image by interpolating extra pixels calculated from the adjacent actually scanned pixels. Most suppliers quote interpolated resolution.</p>
<b>Automatic document feed</b>	<p>This is the ability to process original documents in bulk, sequentially digitising rather than hand feeding individual pages. Varieties of automatic document feed (ADF) are:</p> <p><b>simplex:</b> scanning one side of a document in one pass</p> <p><b>duplexing:</b> scanning both sides of a document by re-feeding the document through and processing the reverse side</p> <p><b>duplex (true duplex):</b> capturing both sides of a document as an image in one pass through the scanner.</p>
<b>Original size</b>	<p>This is the maximum page size a particular scanner can accommodate. Typically scanners are A3 or A4 enabled, but consideration for non-standard paper sizes may be needed.</p>
<b>Speed of scanning</b>	<p>Digitising speeds depend on the resolution required and whether the digitisation is in colour or black and white (greyscale).</p>
<b>Capability</b>	<p>Scanners support multiple requirements. One scale to assess the capability of scanners is:</p> <p><b>low:</b> 250 pages per day</p> <p><b>medium:</b> 750 pages per day</p> <p><b>high:</b> 2000 pages per day</p> <p><b>production level:</b> 15 000 pages per day.</p>

Characteristic	Explanation
<b>Ports and image transfer supported</b>	<p>Interfaces between scanners, printers and computers for storage are required. Typical interface protocols are:</p> <p><b>USB</b> (Universal Serial Bus) specification 2</p> <p><b>SCSI</b> (Small Computer System Interface)—requires a separate card</p> <p><b>parallel:</b> connecting simply through a parallel port on computers</p> <p><b>firewire:</b> a transmission protocol typically found on high-end, high-resolution scanners.</p>

## APPENDIX 3: Sources

### **Sources referenced in this guide**

Archives New Zealand, *Digitisation Standard*, Archives New Zealand, 2007, published at <http://continuum.archives.govt.nz/files/file/standards/s6.pdf>

Attorney-General's Department, *Australian Government Protective Security Manual (PSM)*, Attorney-General's Department, published at <http://www.ag.gov.au/pspf>

Attorney-General's Department, *Australian Government Protective Security Policy Framework (PSPF)*, Attorney-General's Department, published at <http://www.ag.gov.au/pspf>

Australian Government Information Management Office, *Web Content Accessibility Guidance (WCAG) 2.0*, Australian Government Information Management Office, published at <http://www.w3.org/TR/WCAG20/>

National Archives of Australia, *Australian Government Recordkeeping Metadata Standard*, version 2.0, National Archives of Australia, July 2008, published at [http://www.naa.gov.au/Images/AGRkMS\\_Final%20Edit\\_16%2007%2008\\_Revised\\_tcm2-12630.pdf](http://www.naa.gov.au/Images/AGRkMS_Final%20Edit_16%2007%2008_Revised_tcm2-12630.pdf)

National Archives of Australia, *Digital Recordkeeping: Guidelines for Creating, Managing and Preserving Digital Records*, National Archives of Australia, May 2004, published at [http://www.naa.gov.au/Images/Digital-recordkeeping-guidelines\\_tcm2-920.pdf](http://www.naa.gov.au/Images/Digital-recordkeeping-guidelines_tcm2-920.pdf)

National Archives of Australia, *General Disposal Authority for Source Records that have been Copied, Converted or Migrated*, National Archives of Australia, February 2003, published at [http://www.naa.gov.au/Images/GDA\\_copied\\_records\\_tcm2-1124.pdf](http://www.naa.gov.au/Images/GDA_copied_records_tcm2-1124.pdf)

National Archives of Australia, *Glossary of Records Management Terms*, National Archives of Australia, published at <http://www.naa.gov.au/records-management/glossary/index.aspx>

National Archives of Australia, *Standard for the Physical Storage of Commonwealth Records*, National Archives of Australia, December 2002, published at [http://naa.gov.au/Images/standard\\_tcm2-1042.pdf](http://naa.gov.au/Images/standard_tcm2-1042.pdf)

Public Record Office Victoria, *PROS 10/02/S1: Digitisation Requirements v 1*, Public Record Office Victoria, 2009, pp. 9–11, published at <http://www.prov.vic.gov.au/publications/publns/1002s1.pdf>

Queensland State Archives, *Digitisation Disposal Policy and Implementation Toolkit*,  
Queensland State Archives, September 2010, published at  
<http://www.archives.qld.gov.au/government/ddp.asp>

Standards Australia International, AS 5044-2010, *AGLS Metadata Standard*, Parts 1 &  
2, National Archives of Australia, published at <http://www.agls.gov.au/>

### **Other useful sources**

AGLS, <http://www.agls.gov.au/>

Australian Government Information Management Office, *Better Practice Checklist—No. 18 Digitisation of Records*, Australian Government Information Management Office, 2004, updated 2008, published at <http://www.finance.gov.au/e-government/better-practice-and-collaboration/better-practice-checklists/digitisation.html>

Federal Agencies Digitization Guidelines Initiative, <http://www.digitizationguidelines.gov/>

International Organization for Standardization, ISO 12653 – 1: 2000, *Electronic Imaging—Test Targets for the Black and White Scanning of Office Documents, Part 1: Characteristics*, International Organization for Standardization.

International Organization for Standardization, ISO 12653 – 2: 2000, *Electronic Imaging—Test Targets for the Black and White Scanning of Office Documents, Part 2: Methods of Use*, International Organization for Standardization.

Standards Australia, AS ISO 15801-2006, *Electronic Imaging—Information Stored Electronically—Recommendations for trustworthiness and reliability*

JISC Digital Media, <http://www.jiscdigitalmedia.ac.uk/>

National Archives of Australia, *AGLS Metadata Standard, Australian Government Implementation Manual, Version 3.0*, National Archives of Australia, published at [http://www.naa.gov.au/Images/AGLS%20Australian%20Government%20Implementation%20Manual%20v3\\_0\\_tcm2-914.pdf](http://www.naa.gov.au/Images/AGLS%20Australian%20Government%20Implementation%20Manual%20v3_0_tcm2-914.pdf)

National Archives and Records Administration, *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files—Raster Images*, National Archives of Australia, 2004, published at <http://www.archives.gov/preservation/technical/guidelines.html>

Public Record Office Victoria,  
<http://www.prov.vic.gov.au/records/standards.asp#standards>

### **Resource under development**

International Organization for Standardization, ISO/DTR 13028, *Information and documentation—Implementation guidelines for digitization of records*, International Organization for Standardization.