Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 4. CREATION AND MANAGEMENT OF DIGITAL DOCUMENTS

LESSON 4. FROM HARDCOPY TO ELECTRONIC DOCUMENTS

NOTE

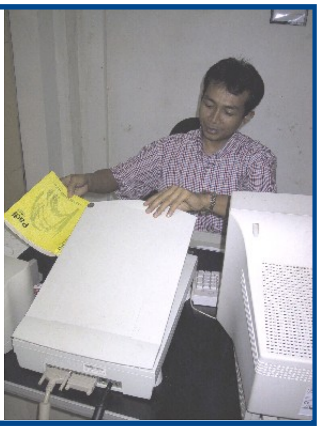
Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

Learning Objectives

At the end of this lesson you will be able to:

- identify the **different phases of the digitizing process**.



The process

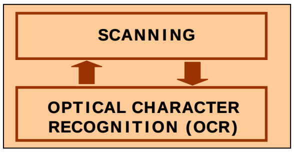
The process of converting a stack of books and papers into a set of electronic documents includes the following phases:

OBTAINING DOCUMENTS	**Collecting** documents to be digitized.
REGISTERING DOCUMENTS	Registering documents to **keep track** of them.
SCANNING	Transforming hardcopy into electronic files (**image format**).
OPTICAL CHARACTER RECOGNITION (OCR)	Converting documents from image to **text format**.
PROOFREADING AND REFORMATTING	Making **corrections** to the document text and layout.
PRODUCING THE FINAL VERSION	Adding **metadata** and other elements to complete the product.

At the end of this process documents will have the requirements needed to be included in a digital library.

The process

Before starting, consider the following options:



SCANNING

OPTICAL CHARACTER RECOGNITION (OCR)

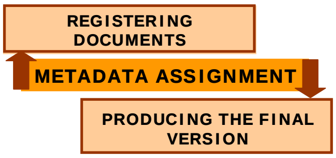
It is possible to **scan and OCR in a single operation**, but it may be better to do these tasks separately: scan using the software that came with your scanner, then OCR the resulting files in a dedicated OCR program.

This is because OCR is more **time-consuming** than scanning. Rather than tying up the computer attached to the scanner, it may be better to have someone else (or several people) do the OCR separately.

The dedicated software that comes with the scanner is designed for that scanner, so it produces the best-quality output. But it may not be able to do OCR, or it may lack all the features of a specialist OCR program.

A disadvantage of scanning and performing the OCR separately is that scanning alone produces image files, which can be very large. A solution is to store them on rewritable CDs, and delete the ones you have finished with.

The process



REGISTERING DOCUMENTS

METADATA ASSIGNMENT

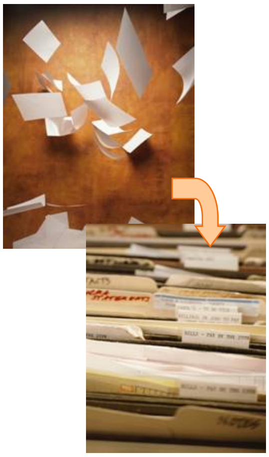
PRODUCING THE FINAL VERSION

You can assign metadata at the same time as registering the documents. Or you can do it later, when producing the final version of the electronic document.

If you want to include only simple metadata (the title, author and publication date), this can be done by secretarial staff.

If you want more detailed metadata (such as keywords and abstracts), you will need a specialist (such as a librarian) to do the job.

Managing documents



If you have to scan a large number of documents, you should first register them and use a **filing system** to keep track of them.

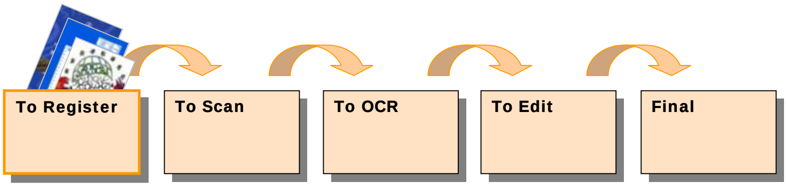
If not, you risk misplacing hardcopies (embarrassing if they must be returned to their owners), losing files, skipping steps in the process, or duplicating work – perhaps without realizing it. You also risk losing electronic versions of files because they have been misnamed or saved in the wrong subdirectory.

Moreover, a good filing system is vital so **everyone** in the digitizing team knows what they are supposed to do and can fill in for one another in case of absence.

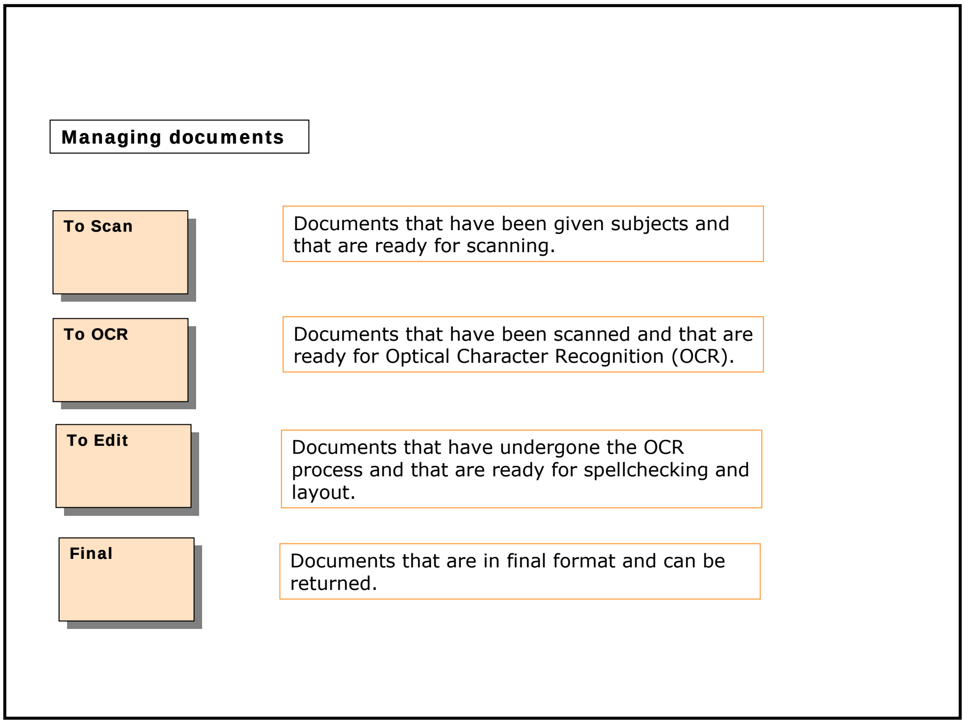
Managing documents

Keep the **hardcopies** of documents at each stage of the process separate from those at earlier and later stages. As each document is processed, take it out of one folder, process it, and put it in the next folder.

Click on any folder to view which type of documents it contains.



Documents that you have received but which have not yet been registered.



Managing documents

To Scan

Documents that have been given subjects and that are ready for scanning.

To OCR

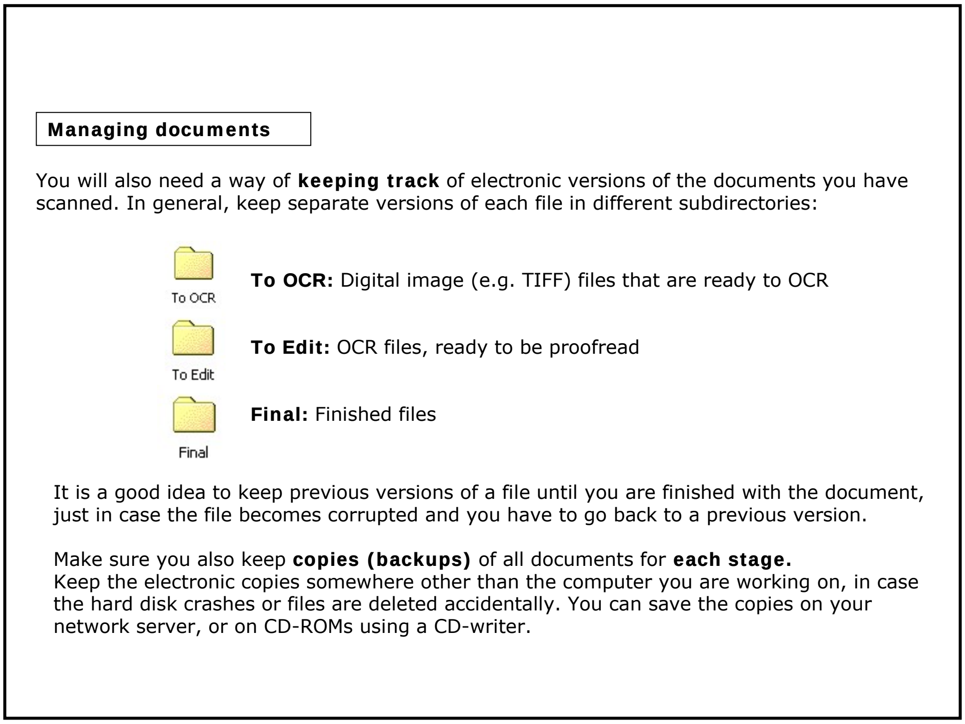
Documents that have been scanned and that are ready for Optical Character Recognition (OCR).

To Edit

Documents that have undergone the OCR process and that are ready for spellchecking and layout.

Final

Documents that are in final format and can be returned.



Managing documents

You will also need a way of **keeping track** of electronic versions of the documents you have scanned. In general, keep separate versions of each file in different subdirectories:

To OCR

To OCR: Digital image (e.g. TIFF) files that are ready to OCR

To Edit

To Edit: OCR files, ready to be proofread

Final

Final: Finished files

It is a good idea to keep previous versions of a file until you are finished with the document, just in case the file becomes corrupted and you have to go back to a previous version.

Make sure you also keep **copies (backups)** of all documents for **each stage.**
Keep the electronic copies somewhere other than the computer you are working on, in case the hard disk crashes or files are deleted accidentally. You can save the copies on your network server, or on CD-ROMs using a CD-writer.

Registering documents

This is the first book I have to scan, but first of all, I have to register it...

REGISTERING DOCUMENTS

As soon as a document arrives you should **register it** so you can keep track of it.

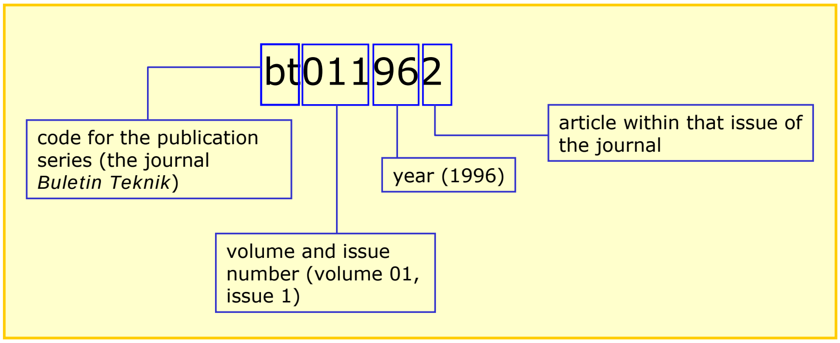
You first have to assign a **filename** to each document. The filename is the basis of a good filing system.

Give each document a filename so you can identify it easily.

On the hardcopy of each document, **write the filename** somewhere unobtrusive (such as inside the front cover or on the back) so you can identify it easily. If you have to return a book to its owner, do not write on the book itself; use an adhesive label instead.

Registering documents

Example of a filename:



bt011962

code for the publication series (the journal *Buletin Teknik*)

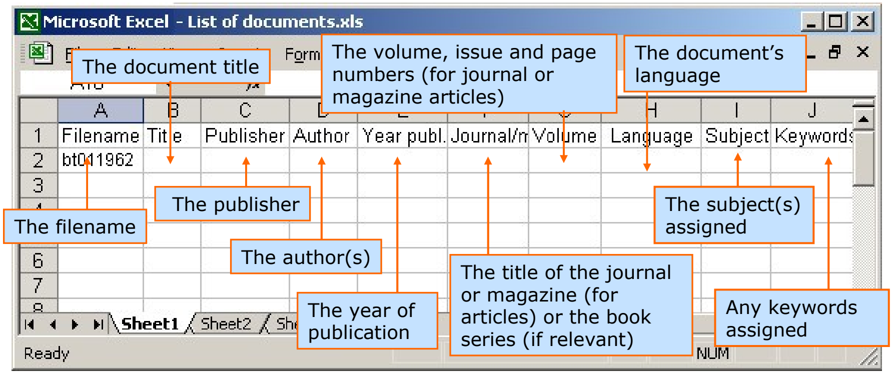
volume and issue number (volume 01, issue 1)

year (1996)

article within that issue of the journal

Registering documents

You can use a spreadsheet to keep track of the documents you are registering and to store the **metadata** for each document:



Microsoft Excel - List of documents.xls

	A	B	C	D	E	F	G	H	I	J
1	Filename	Title	Publisher	Author	Year publ.	Journal/m	Volume	Language	Subject	Keywords
2	bt011962									

The document title

The volume, issue and page numbers (for journal or magazine articles)

The document's language

The filename

The publisher

The author(s)

The year of publication

The title of the journal or magazine (for articles) or the book series (if relevant)

The subject(s) assigned

Any keywords assigned

If you work in a library, you may be able to download this information from the catalogue database.

Registering documents

More information on the spreadsheet:

You may need to add **extra columns** if you also want to record other items, such as the title in English or another language or the publication city.

You can also add columns to this spreadsheet so you can note the following:

- where the document **came from** (e.g. from which library or personal collection), and where and when to return it (if it must be returned);
- **date scanned**, by whom;
- **date of OCR**, by whom;
- **date proofread**, by whom;
- whether the file is in **final format** (ready for use); and
- notes on the **status** of the document.

You can **print out the spreadsheet** file so staff can refer to it and make notes by hand. Update the spreadsheet file regularly so it stays accurate.

Scanning documents

Before scanning, clean any dust off the documents to be scanned, and make sure that all the pages are present and in the right order.

If the document is in poor condition (as with well-used library books), try to find a fresh copy.

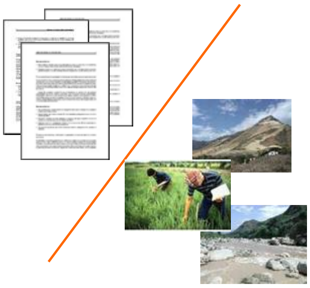
SCANNING DOCUMENTS

If you have a **sheet-fed scanner**, cut the book open (easy and neat if you use a printer's cutting machine) to get **individual sheets** you can feed through the scanner. If necessary, you can rebind the books later.

If you don't want to damage the books, you can photocopy each page and feed the **photocopy** through the scanner - though this uses a lot of paper and reduces the quality of the scan. If the book contains **photographs**, you should scan them separately by hand: photos do not photocopy well.

Scanning documents

To scan a document, place it face down on the scanner platen, or put the pages into the sheet feeder. Then, in the software, choose a setting: **resolution** and **colour**.



There is a trade-off between image **resolution and quality**: the better the quality, the more disk space the image takes up (and the slower it downloads over the internet).

For a textual document you can choose a setting with low resolution (72 dpi) and black and white.

If your document contains both **text and graphics**, it may be best to scan twice: once to scan the text in black and white, and again to scan the pictures in colour.

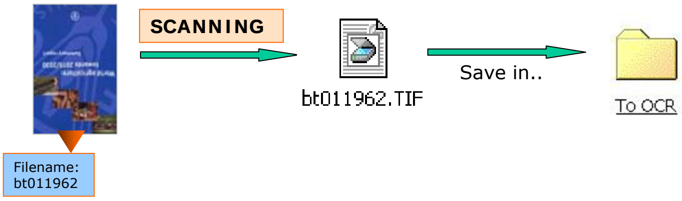
Then save the text and each graphic as **separate files**. You will reincorporate them into the document later.

How to scan pictures and diagrams **How to scan tables**

Scanning documents

The software may produce **a separate image file for each page** of the document, in TIF format or in its own proprietary format for you to convert later.

Follow the file-naming convention you have chosen: e.g.: **bt011962.tif** for the document with the filename **bt011962**. Then, save these files in the 'To OCR' subdirectory.



If you are combining scanning and OCR, you can save the resulting OCR file in a format that can be read by your word processor (e.g.: DOC, RTF or SXW) or your web editor (e.g: HTML).

If you have chosen to produce your document in **HTML** format, put the HTML document in its own subdirectory, along with the pictures that go with it. Save the images with the **same name** as the document, but numbered consecutively (e.g: 01, 02, 03, etc...).

Scanning documents

When you have finished scanning, you will probably have to **return** the documents to where they came from.

It is a good idea to **keep the hardcopies** of documents until you have finished the **whole process**, in case you need to refer back to them (for example, you may need to rescan a page if the file has been corrupted).

If you have cut the bindings in order to scan them, you may have to rebind them.

Optical character recognition

OPTICAL CHARACTER RECOGNITION

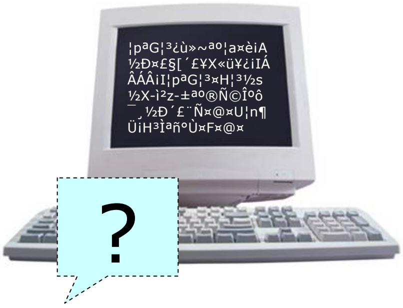
Now you can **OCR** the file that is in the 'To OCR' subdirectory.

OCR software converts a scanned image into a text file that a word processor can read. To do this, it must first **recognize where the text is on the page** (it may be able to detect blocks of text automatically, or you may have to do it manually).

The software then breaks the text blocks down into lines and individual characters. It tries to match the image of each letter against patterns it recognizes as an 'a', 'b', etc... If it does not recognize a particular character it may ask the user for help.

Optical character recognition

You may encounter problems with languages that use Latin scripts with a lot of accented characters (such as á, å, æ, ẫ, etc.), and non-Latin scripts (Amharic, Arabic, Burmese, Chinese, Cyrillic, Hindi, Japanese, Khmer, Korean, Thai…).



Digitizing documents in these languages has three main problems:

- **Recognizing characters** – the software may not be able to distinguish two similar characters (such as à and â).
- **Correcting characters** – the software may not be able to correct errors automatically.
- **Representing characters** – some characters require complicated coding to be displayed correctly on the screen or in printouts. They may appear as hollow boxes (□) or garbage characters (ÄÇù♠).

Optical character recognition

As a solution, you should use OCR software that is **specific for your language**.

Moreover…

- Use a language-specific **dictionary** in your spellchecking or word processing program. See the Help screens in your word processor for how to add a custom dictionary.

- Make sure you use **Unicode** to represent characters. You may be able to find programs that **convert** from other encoding systems to Unicode.

If the OCR software fails to recognise a large number of characters, it may be better to retype all or parts of the document, or to scan the text as an **image** (but remember: in this case users won't be able to search the full text).

How to set up a custom dictionary in Microsoft Word

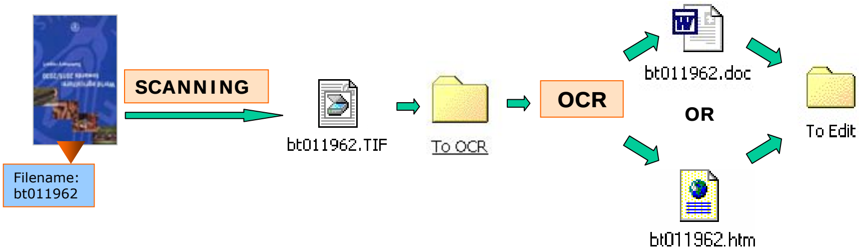
Optical character recognition

When saving your file to the 'To Edit' subdirectory:

- **choose the format** (DOC, RTF or SXW if you want to produce PDF documents - or HTML - if you want to produce HTML documents); and

- **name the file** following your file-naming convention.

See the example below:



Optical character recognition

You have registered a document as **bt021973**. After scanning, you will save the file as __________ in the folder __________. You then OCR this file and create a file named ______________. You save this file in the folder _____________.

1 **bt021973.doc**

2 **bt021973.tif**

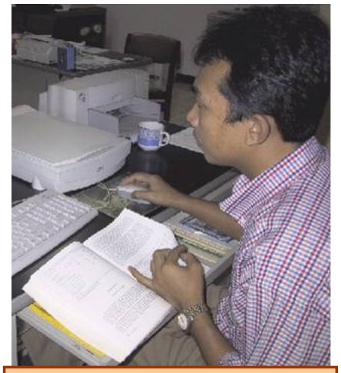
3 **To Edit**

4 **To OCR**

Click on each option and drag it in the appropriate space.

When you have finished, click on the Confirm button.

Proofreading



PROOFREADING AND REFORMATTING

Now you have to do **proofreading**, that is making **corrections** to the document text and layout. You can do this in two ways.

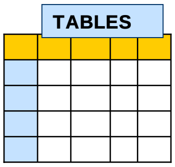
1) Comparing the scanned text **on screen** with the hardcopy, and entering the corrections directly into the computer. You can use your word processor's **spellchecker** to help you find spelling errors quickly.

2) **Printing out** the scanned text and comparing it with the original copy. Mark any corrections on the **printout**, then enter them into the computer. This is a slower method, but may be the best option if you do not have enough computers for each proofreader.

You can **combine** these **two methods**: first correct any obvious mistakes (such as major layout problems and spelling errors) on screen. Then print out the file and check for hard-to-spot errors by hand.

Proofreading

Make sure you proofread **tables and graphics** carefully. Make sure that they have all been scanned, and that the filenames are correct.



If you have OCRed or retyped the tables (rather than scanning them as graphics), you have to proofread them carefully. You should check:

- the **layout:** are the cell contents in the right columns and rows? Are headings, lines and footnotes correct?
- the **contents**: are the numbers correct? Are the commas and decimal points in the right places?

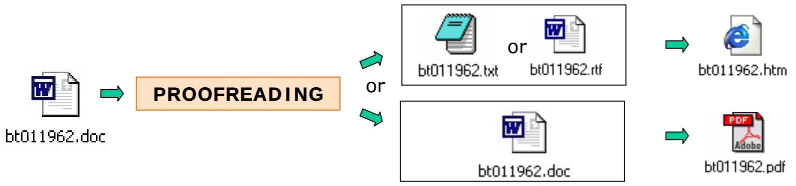
Check the **file size** and **format** of the graphics. If necessary, reduce the file size in an image manager. You may have to rotate the graphic to make sure it is oriented the right way in the document.

Check that the **colours** are accurate. The scanner may have been set by mistake to black-and-white, when you want a colour picture. Sometimes scanners can produce images that have a colour cast. In both instances, you may have to rescan the pictures. Check that the **captions** of figures have been included in the graphic (or are in the text file).

Proofreading

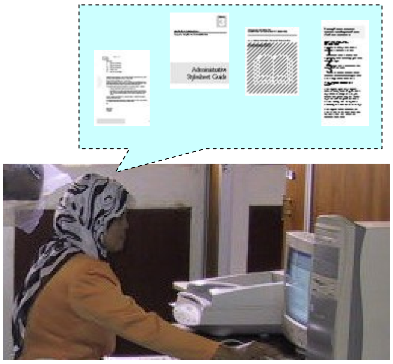
You can do proofreading using either your web editing program (for a HTML file), or your word processor (if the file is destined to become PDF). Word processors are generally **easier to use** for editing and may have a more powerful spellchecker, so you may still decide to use a word processor for these tasks, **then save** the document in **HTML** format.

However, such files are generally large because the word processor inserts many unnecessary formatting codes. So, after editing the document in your word processor, try saving it in an **intermediate format**, such as TXT (plain ASCII text) or RTF.
Then, open this in your web editor and save it as HTML.



Layout

Your OCR software may produce a document that consists of straight text: no columns, no pictures, no headers and footers.



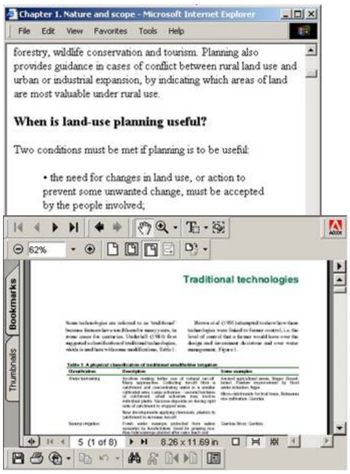
You may have to **reinsert** these by hand, or **correct** where they appear on the page. You may also want to change the typeface, heading styles, and so on, to make the document **more attractive and readable**.

Alternatively, you may be able to adjust the settings of your OCR program so it preserves the layout of the page. This can be helpful, but it is rarely totally satisfactory.

It may be best to correct **major layout problems** before doing the proofreading.

You can correct **more detailed** layout problems at the same time as proofreading. But it is probably better to do it afterwards in a separate operation to avoid proofreading errors.

Layout



forestry, wildlife conservation and tourism. Planning also provides guidance in cases of conflict between rural land use and urban or industrial expansion, by indicating which areas of land are most valuable under rural use.

When is land-use planning useful?

Two conditions must be met if planning is to be useful:

- the need for changes in land use, or action to prevent some unwanted change, must be accepted by the people involved;

Traditional technologies

For **HTML** documents, you should probably use a **simple layout**: a single column of text, and so on. Use your web editing program to insert the pictures and captions. Make sure that all the links to the images are correct, or the images will not display in the document.

For documents destined to become **PDF**s, you can use your word processor to create a **suitable layout**.

If you want to create **both HTML and PDF** versions of the document, do all the proofreading and layout in your word processor, then convert the finished result into PDF and HTML formats.

Do not try to recreate the original layout exactly: it can be very difficult and time-consuming.

Producing the final version

For many documents, you may have to **add some information** to the text so that readers can **identify** it easily.

For a book, make sure the book title, author or editor, publisher and publication date are all included.

For chapters in a book, also include the title and author of that chapter and the original page numbers in the printed version of the book.

For journal articles, include journal title, date, volume and issue number, the article title and authors, and the page numbers in the original printed journal.

You can include this information on the first page or in a footnote. You can also put the book or journal title in a header or footer on each page. This information is especially important for scientific articles, where the reader needs to be able to cite the original source accurately.

Producing the final version

In HTML and PDF files, you can add '**bookmarks**' and hyperlinks to a document. You can, for example, build a 'live' **table of contents for that document**, so the user can click on a chapter title in the Table of Contents, and jump directly to that chapter in the text.

If you have not already **assigned metadata** to the document, you should do so now.

When you have finished, you can put your documents in the 'Final' folder: they are now ready to be included in a digital library.

How to assign metadata (see how_to_assign_metadata.pdf)

All the documents to be included in the digital library are ready!

PRODUCING THE FINAL VERSION

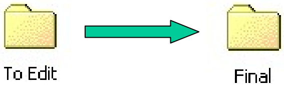


Image files

Because OCR is time-consuming, you may decide to **include images** of the document pages in your digital library, **rather than converting them to text**.

Image pages are much faster to make, but they take up a lot more disk space, so take longer to download than text files. They contain no text, so they cannot be searched, edited or indexed. That means more detailed metadata (such as keywords) may be needed so users can locate the document in your digital library.

You can use any suitable image format for your images: TIFF, JPG, or GIF. You can also use an **image PDF**, which will allow users to see the document page by page using Acrobat Reader.

Image files

Depending on the design of your digital library, it may be possible for it to contain both image-only and text-based documents.

You can also create an **image PDF** that also contains the **full text**, hidden from view (**Searchable Image PDFs**). That displays an exact replica of the original manuscript, but allows a digital library program to index the text and users to search it.

Because an exact image of the text appears on the screen, a few spelling errors in the hidden OCRed text don’t matter much. That means there is no need to:

- proofread the text as carefully as for full OCR.
- reformat the OCRed text so it looks pleasing.

On the next screen, you will find documentation on how to make image and searchable image PDFs, including a list of advantages and disadvantages of the different types of PDF files.

Guidelines and procedures

Here you can download and print the documents provided in this lesson.

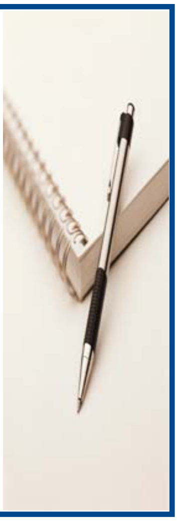
You may use them as tools for your job.

- **How to scan pictures and diagrams**
- **How to scan tables**
- **How to set up a custom dictionary in Microsoft Word**
- **How to assign metadata**
- **Using Image and searchable Image PDF**

Summary

- There are six stages in digitizing documents for a digital library: **registering**, **scanning**, **optical character recognition**, **proofreading** and **reformatting**, and **producing the final version**.
- Before scanning a large number of documents, first catalogue them, and use a **filing system** to keep track of them.
- To scan a document, place it face down on the scanner platen, choose a suitable **setting** (resolution and colours) and scan each page of the document at the settings you have chosen.
- OCR software converts a scanned image into a text file that a **word processor** can read.
- To obtain the final version of a file, you have to **proofread** it to correct the spelling.
- You may also have to adjust the layout. For many documents, you should **add some information** to the text so that readers can **identify** it easily.
- You have to add **metadata** to describe each document.

The proofreading and reformatting processes are very time-consuming. You can avoid them by creating a digital library of **images** rather than of text.



Exercises

The following seven exercises will test your understanding of the concepts covered in the lesson and provide you with feedback.

Good luck!

Exercise 1

You may wish to keep five file folders to hold the documents as they undergo the scanning process. Put these five folders in the correct order.

□ To OCR

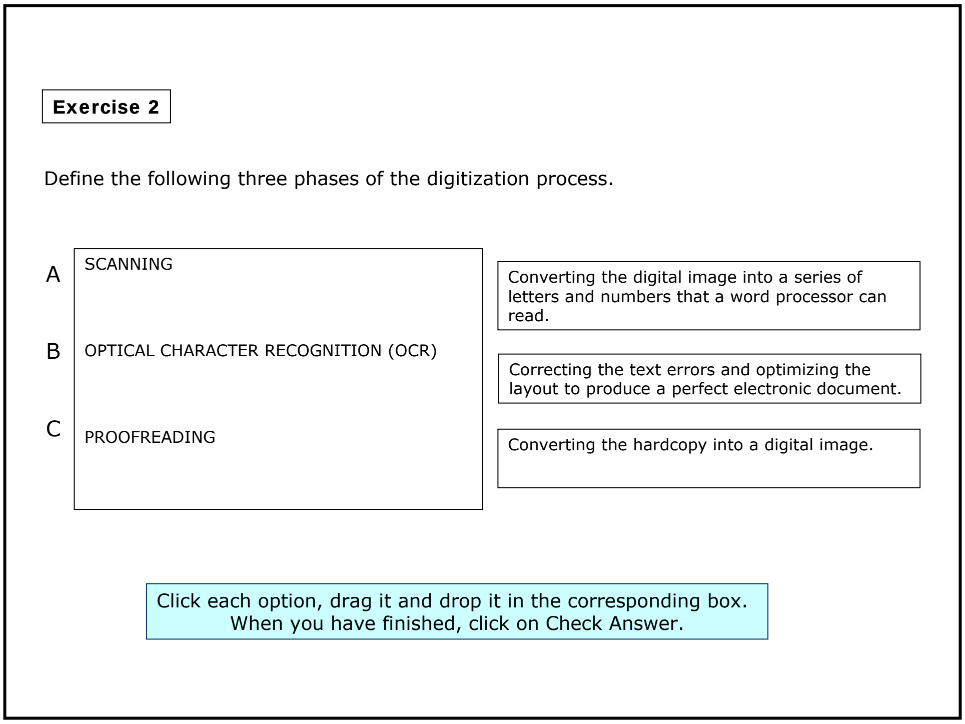
□ To Scan

□ To Edit

□ To Register

□ Final

Please order these items using the dropdown boxes and press Check Answer



Exercise 2

Define the following three phases of the digitization process.

A SCANNING

B OPTICAL CHARACTER RECOGNITION (OCR)

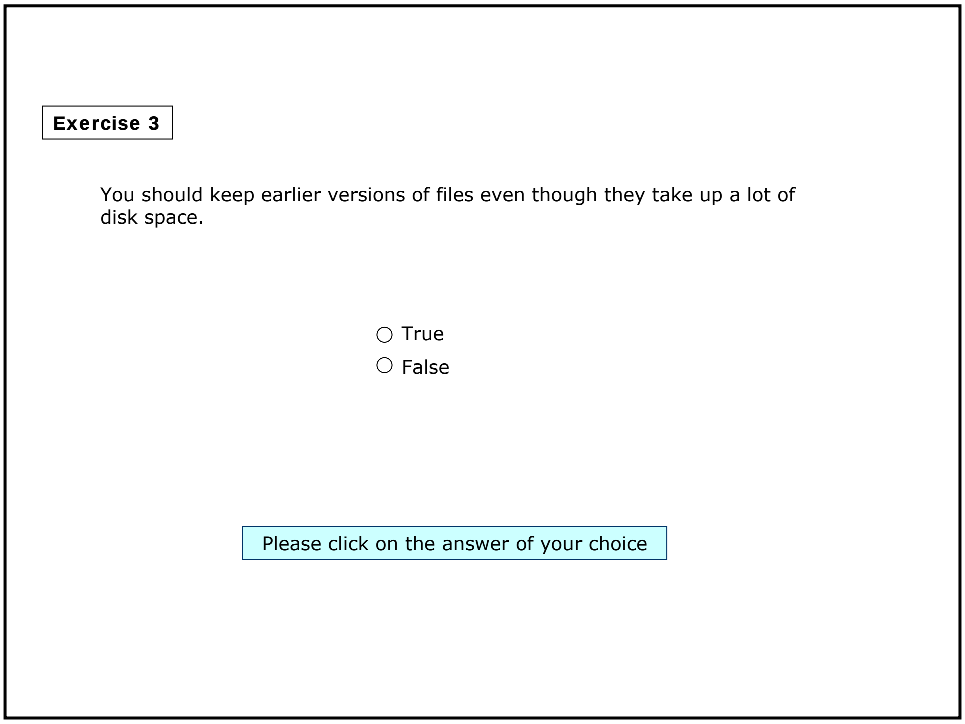
C PROOFREADING

Converting the digital image into a series of letters and numbers that a word processor can read.

Correcting the text errors and optimizing the layout to produce a perfect electronic document.

Converting the hardcopy into a digital image.

Click each option, drag it and drop it in the corresponding box.
When you have finished, click on Check Answer.



Exercise 3

You should keep earlier versions of files even though they take up a lot of disk space.

○ True
○ False

Please click on the answer of your choice

Exercise 4

Scanning is more time-consuming than OCR.

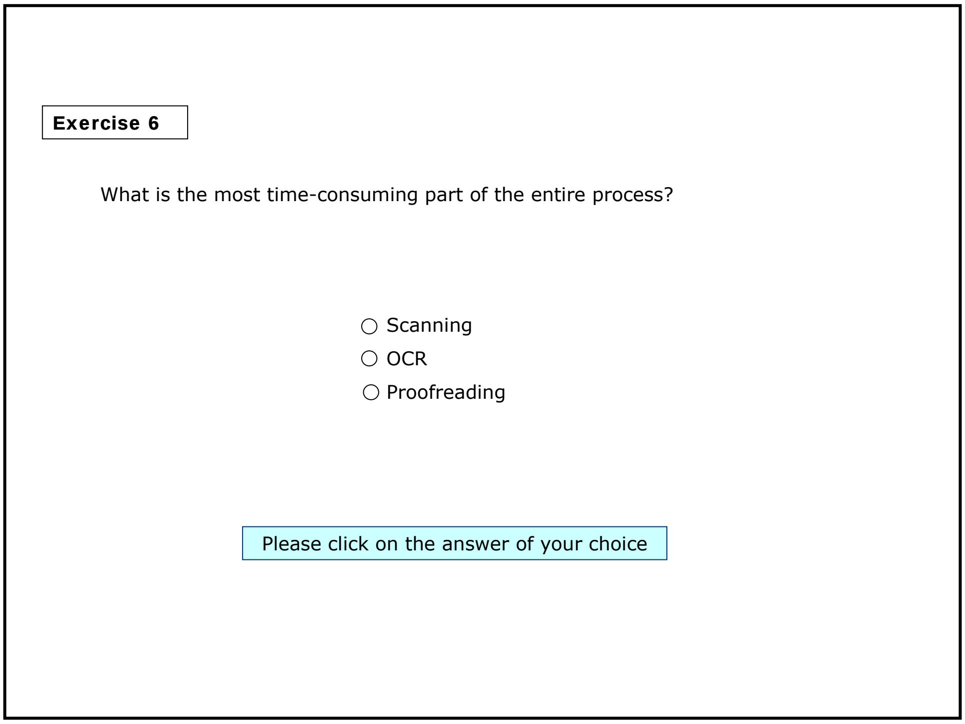
○ True
○ False

Please click on the answer of your choice

Exercise 5

If your OCR program makes many errors trying to read a document, what can you do?

Please write your answer in the input box and press Check Answer



Exercise 6

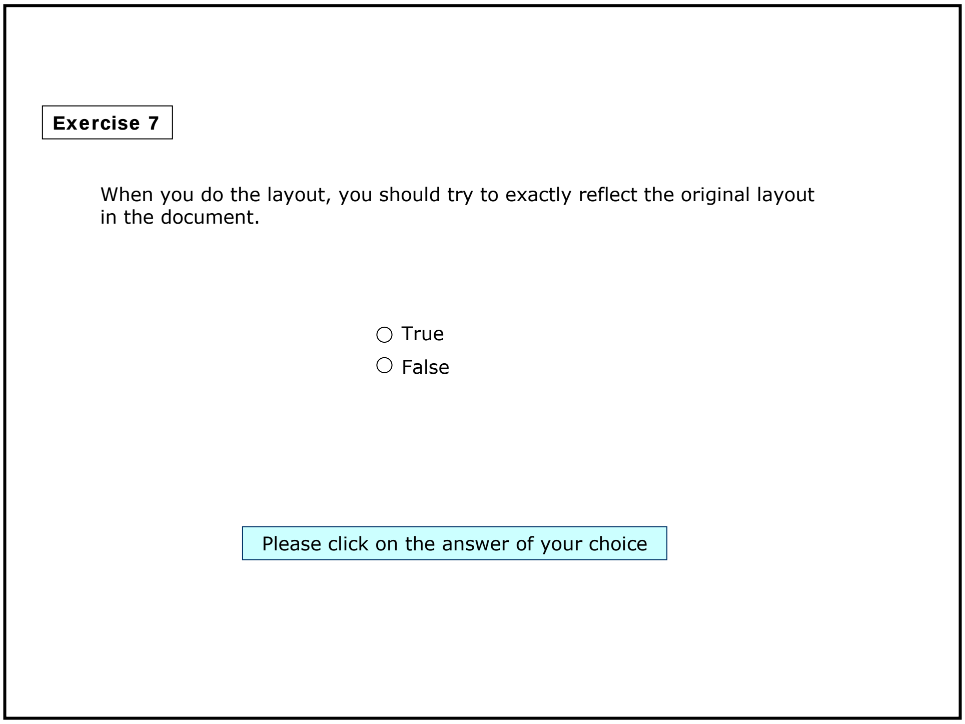
What is the most time-consuming part of the entire process?

○ Scanning

○ OCR

○ Proofreading

Please click on the answer of your choice



Exercise 7

When you do the layout, you should try to exactly reflect the original layout in the document.

○ True

○ False

Please click on the answer of your choice

If you want to know more...

Online Resources:

ReadIris website: example of scanning and OCR software: (http://www.readiris.com)

OmniPage website: example of scanning and OCR software: (http://www.omnipage.com)

FineReader website: example of scanning and OCR software: (http://www.finereader.com)

Guide to Digital Scientific Artwork: (http://www.mlab.nl/GtoDSA/Start.htm)

Additional Reading:

Witten, I.H. & Bainbridge, D. 2002. How to build a digital library. The Morgan Kaufmann Series in Multimedia Information and Systems, Edward Fox, Series Editor. ISBN: 1-55860-790-0