Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 4. PRODUCTION AND MANAGEMENT OF ELECTRONIC DOCUMENTS

LESSON 1. DIGITIZING PRINTED DOCUMENTS: OPTIONS AND CHOICES

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.

Objectives

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

- understand **whether you should convert hardcopy documents** to electronic documents;
- **select the documents to scan**;and
- **assess the resources required** for the scanning process.

Introduction

To digitize a hardcopy document means to convert it to electronic format.

This process consists of three main phases:

1) converting the hardcopy image to a digital image (**scanning**);
2) converting the digital image into text, using optical character recognition (**OCR**); and
3) correcting text errors and optimizing page layout (**proofreading**).

The hardcopy documents might be books, magazines, journals, extension leaflets, training handouts, photographs, line drawings and even handwritten manuscripts.
You may have a few of these, several shelves full, or you may want to convert your library to a digital library…

Why digitize?

Mr. Touré, manager of a library, is evaluating the advantages of digitizing his library's hardcopy documents.

Hmm... converting hardcopy documents to electronic format would allow us to disseminate them via e-mail or the Internet, saving time and money!

Electronic documents are more **versatile** than printed documents: they can be displayed on a computer screen, edited and printed out.

Electronic documents can be **shared easily**: they can be duplicated easily and cheaply, sent by email or put on a website. They can be added to a digital library and made available to users on CD-ROM, or through an Intranet or the Internet.

Why digitize?

Here is another important advantage: electronic documents are **easy to store and retrieve**. Thousands of documents can be stored on a single **CD-ROM** or **hard drive**.
The user can **find** a document **easily** and **quickly** using the computer's search capabilities.

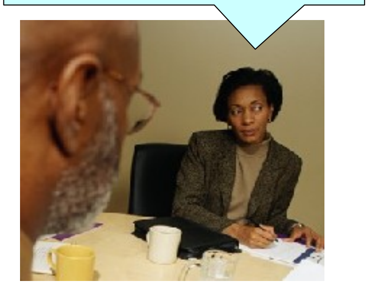
Transforming documents into digital formats also avoids physical deterioration and mishandling of cultural heritage materials such as handwritten manuscripts or books.

Retaining physical reliability is one of the issues related to the **digital preservation** of electronic files, which also include maintaining availability and security of the file collection over time.

Before starting

Scanning is a **time-intensive process**, so it needs careful planning.
Before you start the process, ask yourself these questions:

Yes, the idea is interesting... but before starting the scanning process we must be sure that it is worth it.



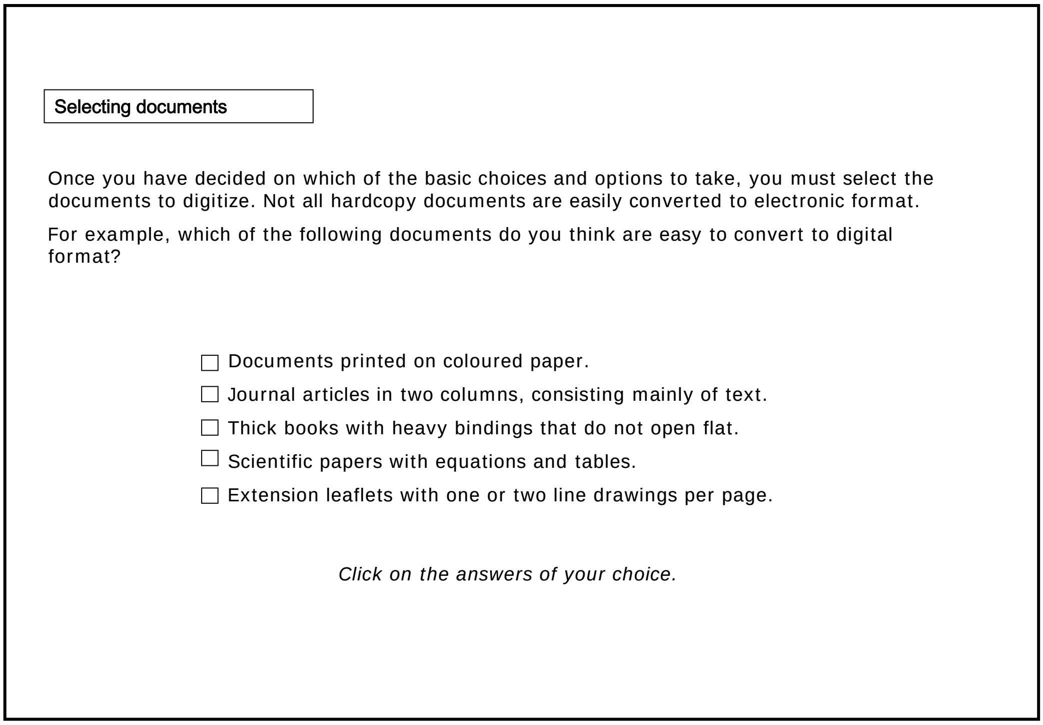
- Who needs the documents and how will they access them? Over the Web, on CD-ROM, etc.?
- What is the main reason for digitizing the documents? Do you want to create a digital library, preserve existing documents, etc.?
- Which documents should be digitized?
- How many documents are there?
- How many languages are we dealing with?
- Who is going to digitize the documents?
- Is this an one-off job or an ongoing commitment?

Before starting

First, decide the output format of the electronic document that you want to create. The basic choice is between image and text formats:

- **Image formats** (TIF, GIF, JPG, image PDF): suitable for **pictures** or **handwritten manuscripts**, and for documents where it is not necessary to search the full text. These are easy to produce, as they are the direct result of the scanning process, but are less useful than text formats.
- **Text formats** (HTML, XML, Microsoft Word DOC, text PDF): they can be obtained by applying OCR to scanned documents.
They are harder to produce, but more useful and easier to use because they **allow full-text searching** and most can be edited using a word processor.

Notice that it is useful to keep the TIF version of a document, resulting from the scanning, for preservation purposes.



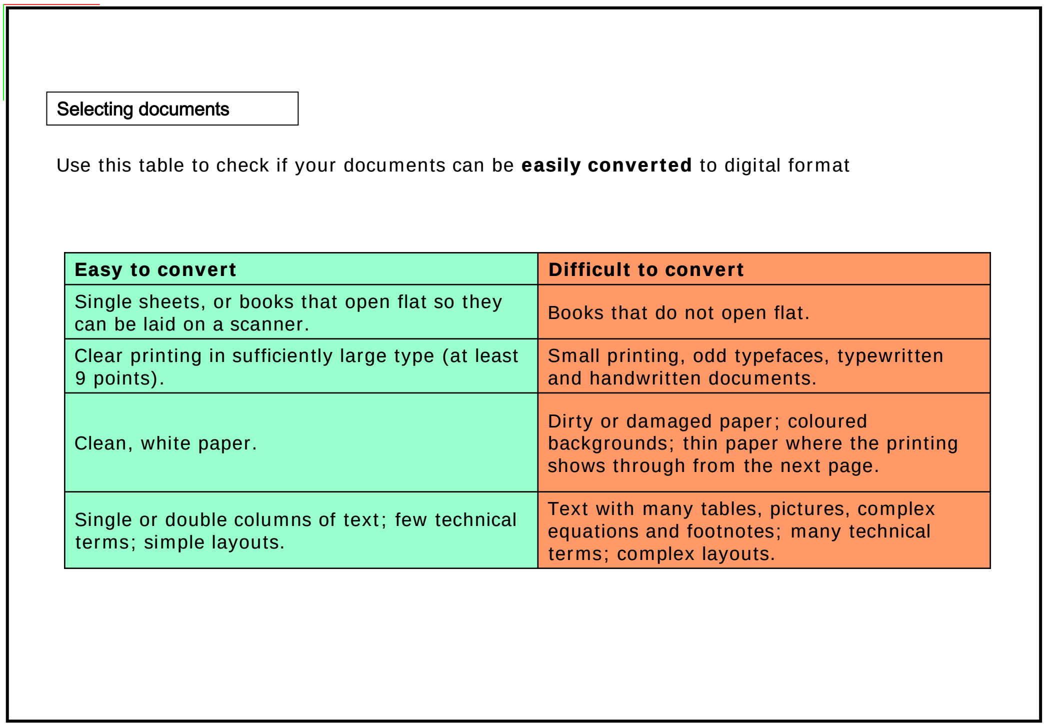
Selecting documents

Once you have decided on which of the basic choices and options to take, you must select the documents to digitize. Not all hardcopy documents are easily converted to electronic format.

For example, which of the following documents do you think are easy to convert to digital format?

- □ Documents printed on coloured paper.
- □ Journal articles in two columns, consisting mainly of text.
- □ Thick books with heavy bindings that do not open flat.
- □ Scientific papers with equations and tables.
- □ Extension leaflets with one or two line drawings per page.

Click on the answers of your choice.



Selecting documents

Use this table to check if your documents can be **easily converted** to digital format

Easy to convert	**Difficult to convert**
Single sheets, or books that open flat so they can be laid on a scanner.	Books that do not open flat.
Clear printing in sufficiently large type (at least 9 points).	Small printing, odd typefaces, typewritten and handwritten documents.
Clean, white paper.	Dirty or damaged paper; coloured backgrounds; thin paper where the printing shows through from the next page.
Single or double columns of text; few technical terms; simple layouts.	Text with many tables, pictures, complex equations and footnotes; many technical terms; complex layouts.

Selecting documents

Make sure you can obtain all the documents you need, and also make sure that documents are not already available in digital format.

You may have to search to find a reasonably **complete set**. Try your institution's library, publication unit, and senior staff (who may have the only copy of certain documents). You may have to borrow documents if your library copy is missing or damaged!

Make sure it is **worthwhile** scanning each document.
For example, you may choose not to include a document that contains information that is clearly **out of date** – for example, instructions to use a pesticide if that chemical has been banned.

Selecting documents

Be careful about **copyright**.
Government documents are increasingly being copyrighted; before reproducing them – check first!
Commercially published **documents** are almost always **copyrighted**, and you must obtain permission from the copyright holder before including them in the collection.
If in doubt, ask the author or publisher.

Be careful also about **security**.
digitazing documents makes them more accessible and easier to copy.
Some types of documents, such as policy discussions, budgets, personnel files and evaluation reports, may be **confidential**.
You can restrict access to such documents by requiring the user to enter a **password** in order to open them, but this is an extra step.

Requirements

Consider the requirements for scanning documents and the relative costs.

Now, let's list what we need to digitize all our documents …

Therefore, you have to consider:

1. the **equipment**: scanners, computers and storage devices;
2. the **software**: scanning, optical character recognition, word processing, spellchecking, image management;
3. the **human resources**: personnel and skills;
4. how much it will **cost**.

Let's analyse each of these items in detail…

Equipment

The first thing you need, is, obviously, the scanner. Scanners come in **three broad price ranges**:

Low-cost flatbed scanners | Low-end scanners with a sheet feeder | High-end professional scanners

Click on each scanner category for details.

PRICE	ADVANTAGES	DISADVANTAGES	WHEN TO USE
From $**100** to $**300**.	Low-cost flatbed scanners can scan both **black-and-white** and **colour** images. Because the price is low, each computer can be equipped with **its own scanner**.	Each page has to be placed carefully by hand on the scanner's glass platen, and the **scanning process** itself is **slow** (only about a dozen pages can be scanned each hour).	Suitable for **small jobs** with a limited number of pages – up to about 400 pages per month on a regular basis, or one-time jobs of up to 2,000 pages.

If you want to scan special types of materials, such as microfiche, slides or oversized materials, you will need special equipment. In this case, but also in other cases, one solution could be to pool resources and purchase one scanner or PC equipment amongst 5 or 10 local organizations.

Low-cost flatbed scanners

Low-end scanners with a sheet feeder

High-end professional scanners

PRICE	ADVANTAGES	DISADVANTAGES	
From **$500** to **$1,200**.	These can handle 10–50 pages at the same time, or about 200 pages per day.	• It is necessary to **cut the binding of books** to make sheets that can be fed into the scanner (photocopying is one option, but this is time-consuming and expensive). • The scanner can scan **only one side** of the page **at a time**, so the stack of pages must be reversed and fed through the machine again in order to scan the other side. • The sheet feeder can become **jammed**.	These scanners are useful for **up to 3,000 pages a month**.

Low-cost flatbed scanners

Low-end scanners with a sheet feeder

High-end professional scanners

PRICE	ADVANTAGES	DISADVANTAGES	WHEN TO USE
From $**5,000** to $**50,000**.	Professional scanners are heavy-duty machines with a **sheet-feeder tray system**, like a photocopier. The best ones can scan both sides of the page at once. Various firms produce dedicated scanning and archiving systems, e.g. high-end scanner that automatically creates **a file for each document**, and allows you to assign **subjects** and **keywords** in a single process.	These systems are **expensive**, and some use proprietary archiving systems that tie you to that firm's software.	These systems are of interest to **large institutions** that wish **to create large digital libraries**.

Equipment

Scanning and optical character recognition require a lot of **computer processing power**.

It is possible to scan several hundred pages, using one computer with a scanner attached. For larger jobs consisting of thousands of pages, however, more computers and operators are needed. Make sure you have **enough disk capacity** (**20 or 30 GB**) to handle the volumes of data you will generate.

Proofreading is very time-consuming but requires less computing power; therefore, several less powerful computers could be used for this task.

If you plan to create a digital library, you will need a reasonably powerful computer to handle the **large amounts of data processing**.

Equipment

You will need a **CD-writer**, for two reasons:

1. to **copy** and store (back up) the large amounts of **data** you produce (using rewritable CDs);

2. to create the **master copy** of the final CD-ROM for distribution (if you plan to distribute your electronic documents on CD-ROM).

A **computer network** is also very useful because it enables you to **back up files** easily, for preservation purposes, and to **share files** among the different people working on the production.

If you do not have a network, you will have to rely on CD-ROMs to transfer data.

Anyway, retaining the 'TIF' versions on CD-ROMs will be very useful as a back-up, and for content refreshing.

Software

You will need the following types of **software**:

- **Scanning software**, to convert the hardcopy image to a digital image and **OCR**, to convert the digital image into text that a word processor can understand (e.g. ReadIris, OmniPage, FineReader).
- **Word processor** and **spellchecker**, to correct text errors and to optimize page layout (e.g. Microsoft Word, Corel WordPerfect).
- **File conversion programs**, to convert files from one format to another.
- **Image management software**, to view, modify and manage images (e.g. CompuPic, Kudo, ACDSee).
- **Image editing software**, e.g Adobe PhotoShop, Corel PhotoPaint, Microsoft PhotoDraw.
- **Adobe Acrobat** Distiller and Reader, if you choose to have documents in PDF format.

When you choose programs, operating systems, etc., remember to consider possible changes due to technology evolution, in order to maintain the ability to display, retrieve, and use your electronic documents.

Personnel

The following types of staff are needed for the digitization process:

- A **manager** to coordinate the team and manage documents.

- People skilled in using computers who are highly motivated and quality-oriented for **scanning.**

- People skilled in using computers (especially word processing) to do the **OCR, proofreading** and **layout**. As best results and productivity are achieved during a limited number of hours each day, this work should either be organized on a part-time basis, or on a full-time basis employing only experienced, highly motivated and quality-conscious people.

A **training course** or **workshop** will be necessary to teach the team members the extra skills they need, and to develop a work flow that suits your organization.

Costs

When budgeting for scanning, you need to include the following items:

But how much will the entire process cost? It's time to have a look at the budget!

- **Equipment**: scanner, computers, office furniture.

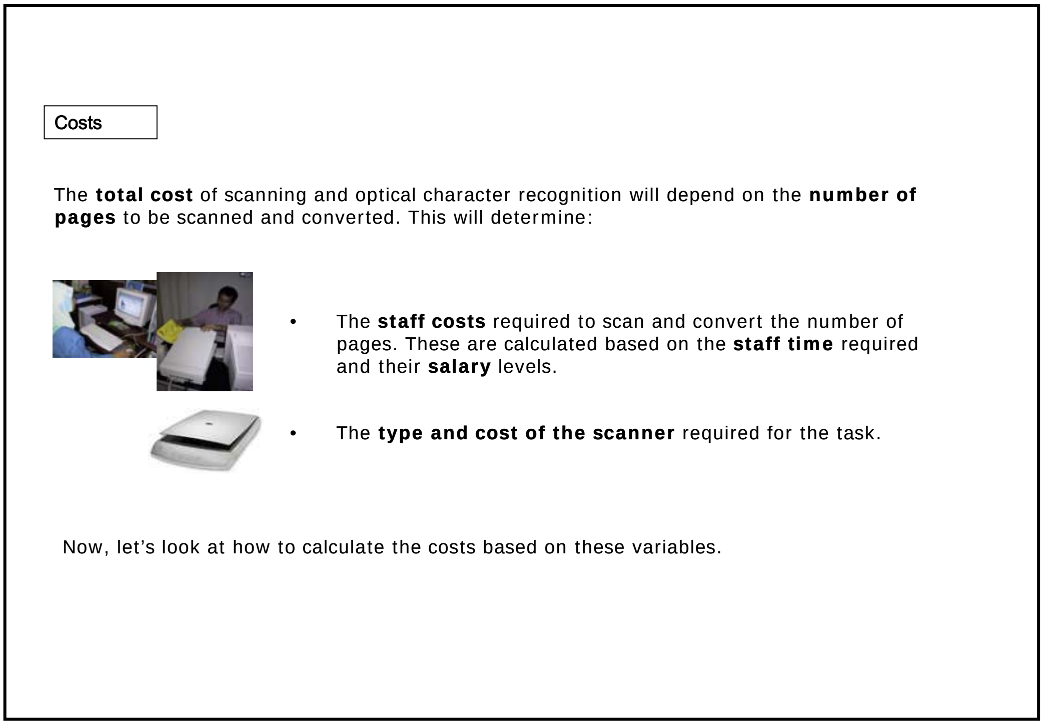
- **Document acquisition, registration, categorisation and return**: mailing and transport costs, staff time.

- **Scanning**: staff time.

- **OCR, proofreading and layout:** staff time, consumables (disks, paper).

- **Management and overhead**: staff training, management staff time, overhead.

If you want to create and distribute a digital library you must also add in **duplication, marketing and distribution costs**.

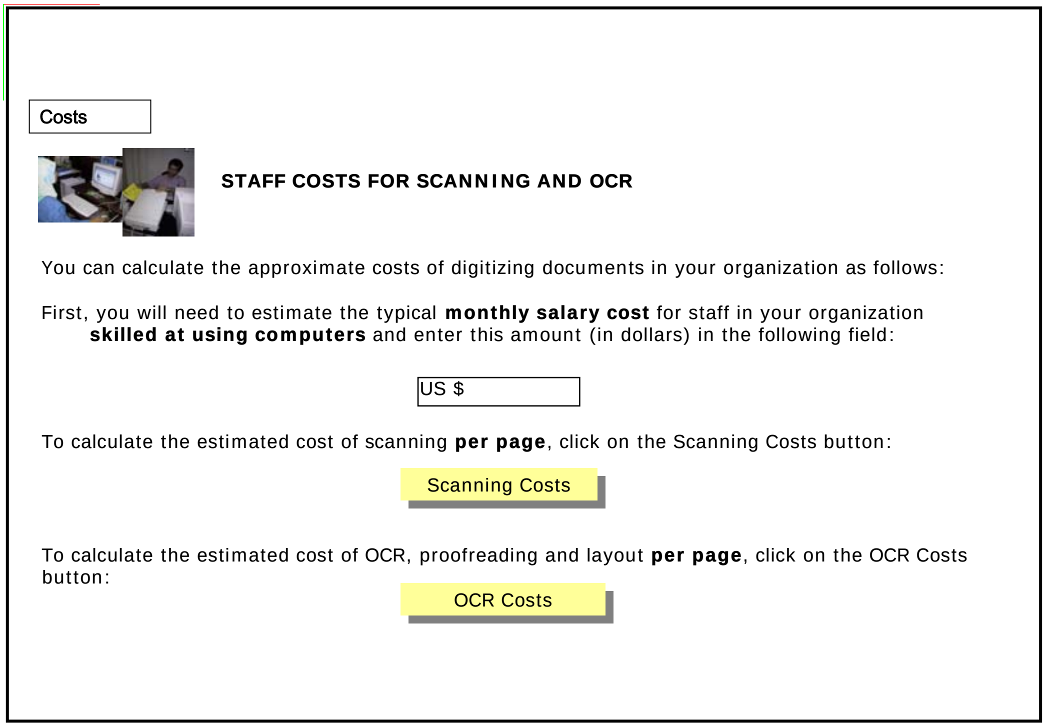


Costs

The **total cost** of scanning and optical character recognition will depend on the **number of pages** to be scanned and converted. This will determine:

- The **staff costs** required to scan and convert the number of pages. These are calculated based on the **staff time** required and their **salary** levels.
- The **type and cost of the scanner** required for the task.

Now, let's look at how to calculate the costs based on these variables.



Costs

STAFF COSTS FOR SCANNING AND OCR

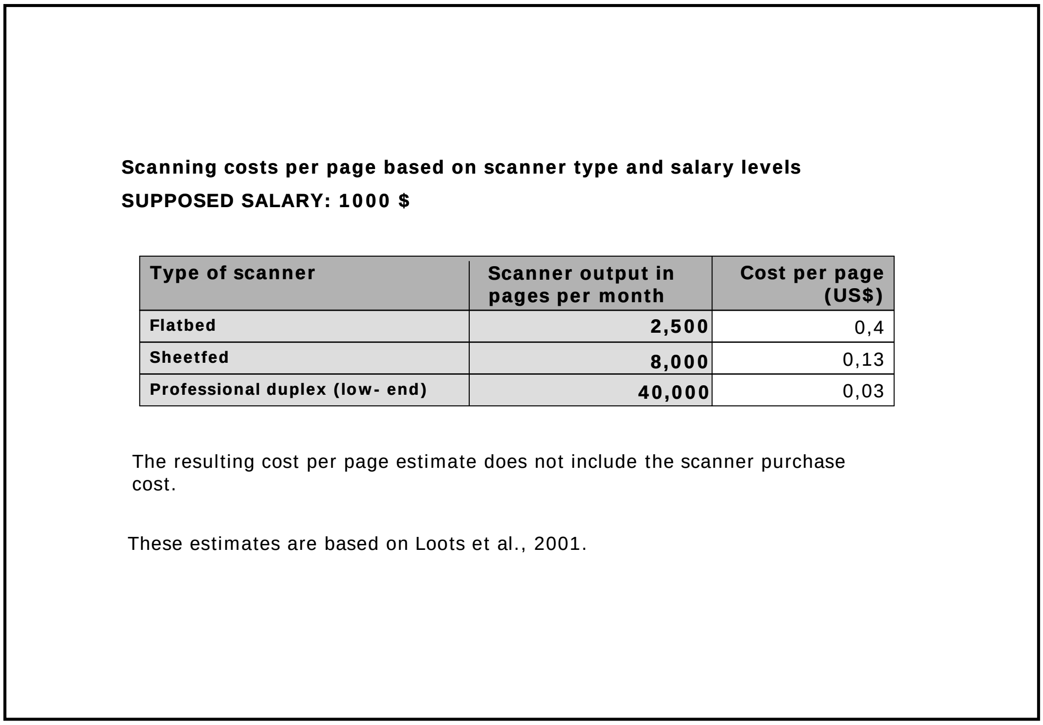
You can calculate the approximate costs of digitizing documents in your organization as follows:

First, you will need to estimate the typical **monthly salary cost** for staff in your organization **skilled at using computers** and enter this amount (in dollars) in the following field:

US $

To calculate the estimated cost of scanning **per page**, click on the Scanning Costs button:

To calculate the estimated cost of OCR, proofreading and layout **per page**, click on the OCR Costs button:



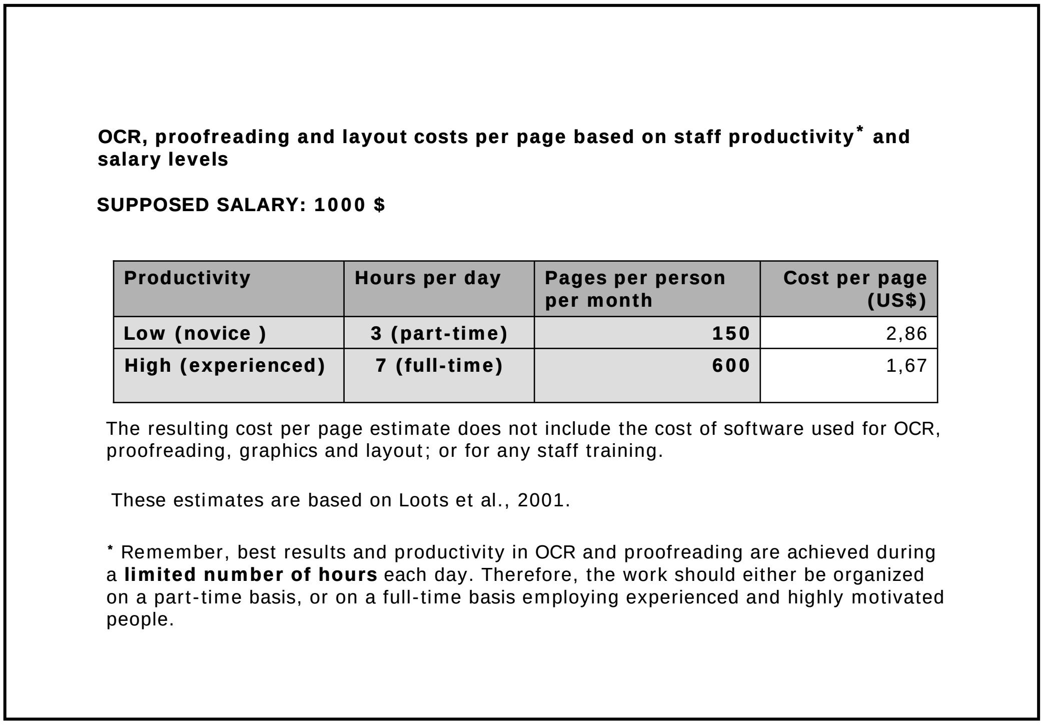
Scanning costs per page based on scanner type and salary levels

SUPPOSED SALARY: 1000 $

Type of scanner	Scanner output in pages per month	Cost per page (US$)
Flatbed	**2,500**	0,4
Sheetfed	**8,000**	0,13
Professional duplex (low- end)	**40,000**	0,03

The resulting cost per page estimate does not include the scanner purchase cost.

These estimates are based on Loots et al., 2001.



OCR, proofreading and layout costs per page based on staff productivity* and salary levels

SUPPOSED SALARY: 1000 $

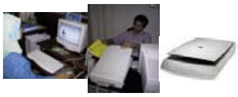
Productivity	Hours per day	Pages per person per month	Cost per page (US$)
Low (novice)	**3 (part-time)**	**150**	2,86
High (experienced)	**7 (full-time)**	**600**	1,67

The resulting cost per page estimate does not include the cost of software used for OCR, proofreading, graphics and layout; or for any staff training.

These estimates are based on Loots et al., 2001.

* Remember, best results and productivity in OCR and proofreading are achieved during a **limited number of hours** each day. Therefore, the work should either be organized on a part-time basis, or on a full-time basis employing experienced and highly motivated people.

Costs



TOTAL COST OF SCANNING AND OCR

As we have seen, the total cost of scanning and OCR depends on the size of the job, and the level of staff and equipment used. For example, while a less powerful scanner has a higher cost of scanning per page, it may be more cost effective than buying a more expensive and powerful scanner for a small to medium-sized job. Now, let’s look at three different cost scenarios which take into account the size of the job and the appropriate scanner to be used.

First, enter the typical **monthly salary cost for staff skilled at using computers** (in US dollars) in the following field:

US $

Then, click on the icons to view the estimated costs for each scenario.

1,000 pages **5,000 pages** **100,000 pages**

These estimates are based on Loots et al., 2001.

Total cost for scanning and OCR (1,000 pages)

SUPPOSED SALARY: 1000 $

1,000 pages represents a part-time job of about one month for scanning, and up-to six months part-time for OCR, proofreading and layout.
A low-cost flatbed scanner and one computer equipped with a CD-R will suffice for this task.

Entries	**Cost (US$)**
Flatbed scanner	300
Scanning	40
OCR, proofreading and layout	286
Total (approximate)	626

The resulting cost estimate assumes that a computer with adequate processing power, storage and back-up device is available. If not, this also needs to be added to the total cost estimate.

1) scanning = 1,000 X cost per page (based on salary costs and use of a flatbed scanner capable of 2500 pages per month as calculated previously).

2) OCR, proof-reading and layout = 1,000 X cost per page (based on low productivity level for OCR, proofreading and layout as calculated previously).

Screen 20

Total cost for scanning and OCR (5,000 pages)

SUPPOSED SALARY: 1000 $

5,000 pages represents a part-time job of less than one month for scanning, and about 33 months part-time, or about 8 months full time for OCR, proofreading and layout. Costs for the later will vary greatly based on staff productivity. A sheetfed scanner and several computers equipped with a CD-R are required for this task.

Entries	**Cost (US$)**
Sheetfed scanner	800
Scanning	63
OCR, proofreading and layout (full time - part time)	833 - 1429
Total (approximate)	1696 - 2292

The resulting cost estimate assumes that computers with sufficient processing power, storage and back-up device are available for scanning and OCR, as well as additional computers for proof-reading and layout. If not, these also need to be added to the total cost estimate.

1) scanning = 5,000 X cost per page (based on salary costs and use of a sheetfed scanner capable of 8,000 pages per month as calculated previously).

2) OCR, proof-reading and layout = 5,000 X cost per page (based on low and high productivity levels for OCR, proofreading and layout as calculated previously).

Total cost for scanning and OCR (100,000 pages)

SUPPOSED SALARY: 1000 $

100,000 pages represents a full-time job of two to three months for scanning, and about 170 months full-time for OCR, proofreading and layout. Novice / low productivity staff should not be considered for this volume of pages. A minimum of a professional low-end duplex scanner and several computers equipped with a CD-R are required for this task.

Entries	**Cost (US$)**
Sheetfed scanner	6,000
Scanning	250
OCR, proofreading and layout	16667
Total (approximate)	22917

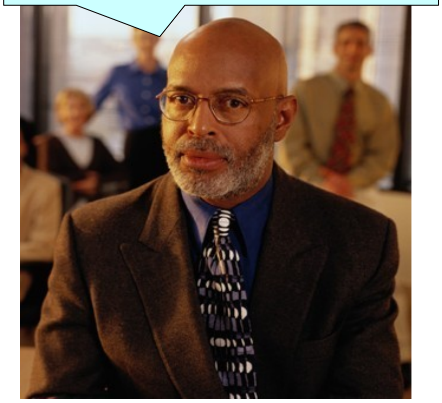
The resulting cost estimate assumes that computers with sufficient processing power, storage (6 to 8 Gbytes) and back-up devices are available for scanning and OCR, as well as additional computers with access to sufficient storage for proofreading, layout and storage of converted documents. If not, these also need to be added to the total cost estimate.

1) scanning = 100,000 X cost per page (based on salary costs and use of a professional low-end duplex scanner capable of 40,000 pages per month as calculated previously).

2) OCR, proof-reading and layout = 100,000 X cost per page (based on high productivity levels for OCR, proofreading and layout as calculated previously).

Outsourcing

Now that we are able to better estimate costs and staffing considerations, our team will be able to determine the best overall approach!



Taking the previous scenarios as a starting point, you can try to determine the best approach and combination of resources for your needs.

But keep in mind that you may also want to consider **outsourcing the job**.

This could be the best choice if you have a "one-off" job, and not an ongoing activity, where the amount of pages to be scanned requires a professional level scanner, but the short-term nature of the job does not justify its purchase.

The costs and staffing commitments required for in-house OCR, proofreading and layout should also be weighed against the cost of outsourcing the work to a professional OCR company.

Summary

- The digitizing process, that allows the conversion of a hardcopy document to electronic format, consists of three phases: scanning, OCR and proofreading.
- When selecting documents to scan, consider how easy they are to convert. Are they up to date? How about copyright and security issues?
- During the planning phase, consider the following issues:

1. the **equipment**: scanners, computers and storage devices;
2. the **software**: scanning, optical character recognition, word processing, spellchecking, image management;
3. the **human resources**: personnel and skills;
4. how much it will **cost**.

Exercises

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

Good luck!

Exercise 1

Define each of the three phases of the digitization process

A SCANNING

B OPICAL CHARACTER RECOGNITION (OCR)

C PROOFREADING

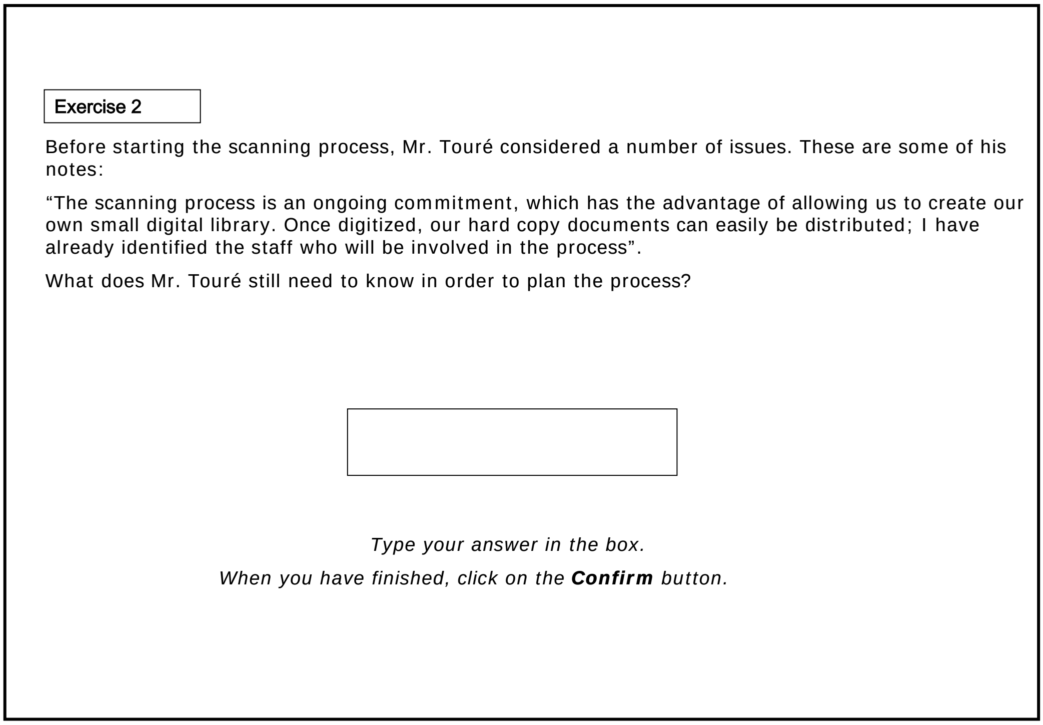
Converting the digital image into a series of letters and numbers that a word processor can read.

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 the confirm button.



Exercise 2

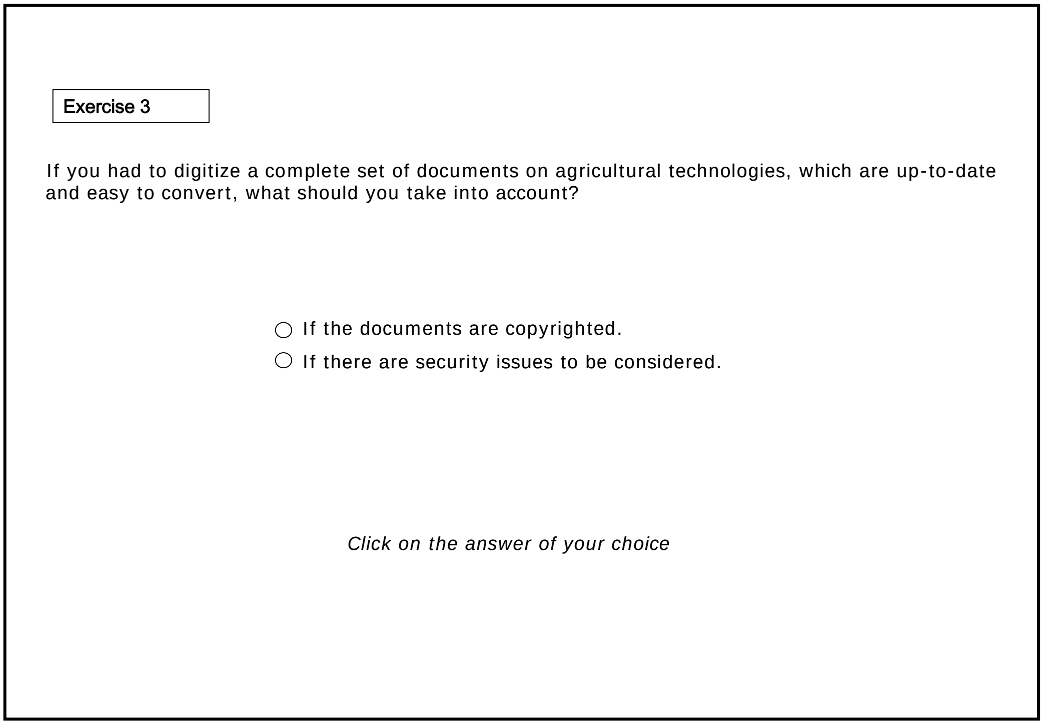
Before starting the scanning process, Mr. Touré considered a number of issues. These are some of his notes:

“The scanning process is an ongoing commitment, which has the advantage of allowing us to create our own small digital library. Once digitized, our hard copy documents can easily be distributed; I have already identified the staff who will be involved in the process”.

What does Mr. Touré still need to know in order to plan the process?

Type your answer in the box.

When you have finished, click on the ***Confirm*** *button.*



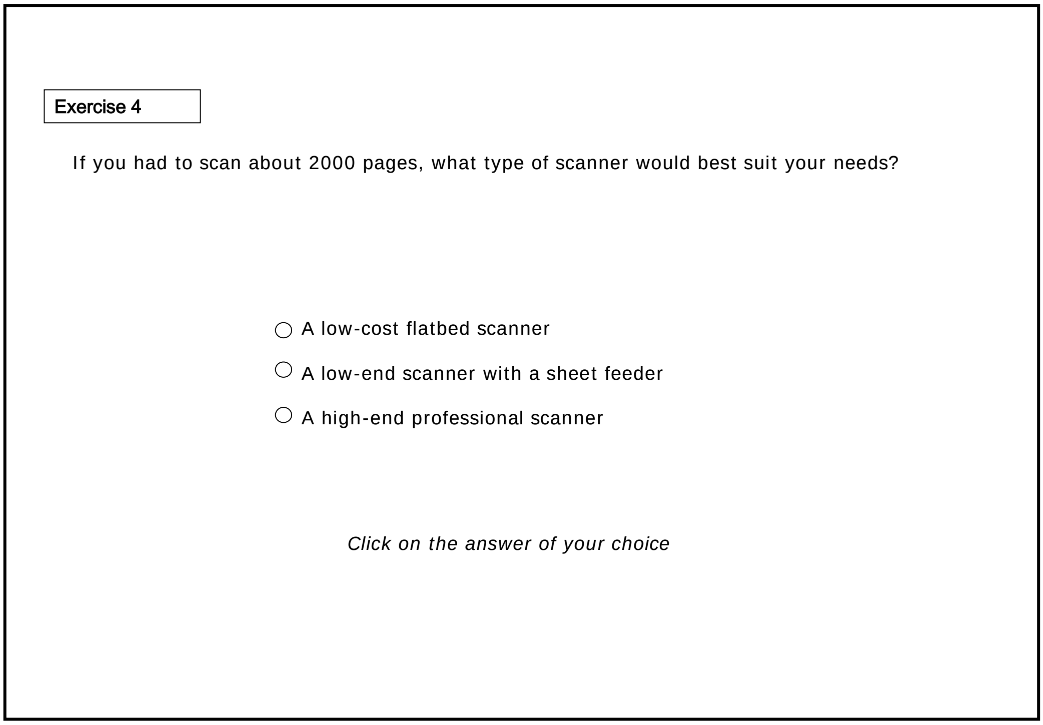
Exercise 3

If you had to digitize a complete set of documents on agricultural technologies, which are up-to-date and easy to convert, what should you take into account?

○ If the documents are copyrighted.

○ If there are security issues to be considered.

Click on the answer of your choice



Exercise 4

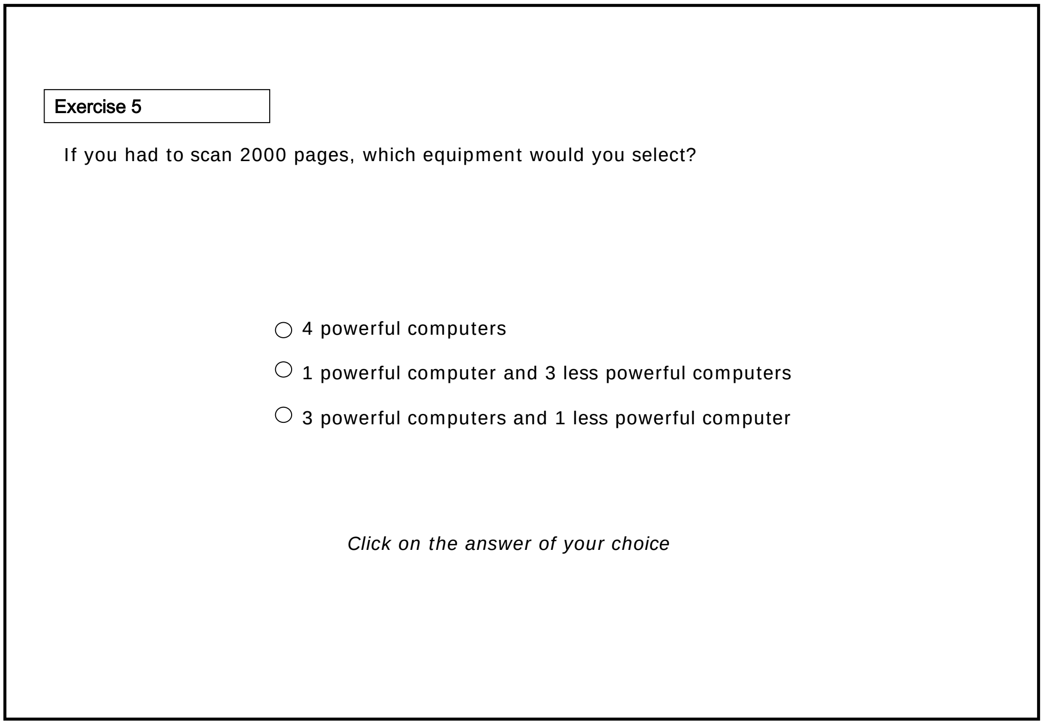
If you had to scan about 2000 pages, what type of scanner would best suit your needs?

○ A low-cost flatbed scanner

○ A low-end scanner with a sheet feeder

○ A high-end professional scanner

Click on the answer of your choice



Exercise 5

If you had to scan 2000 pages, which equipment would you select?

○ 4 powerful computers

○ 1 powerful computer and 3 less powerful computers

○ 3 powerful computers and 1 less powerful computer

Click on the answer of your choice

Exercise 6

Which of these factors will primarily influence the total cost of the scanning process?

□ The number of people who must be involved.
□ The time needed for the process.
□ The number of pages to convert.
□ The number of computers needed.
□ The salary levels of the people doing the work.

Click on the answers of your choice

If you want to know more...

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)
The Digital Library Tool Kit, 3rd Edition. By Peter Noerr. Sun Microsystems. January 2003: (http://www.sun.com/products-n-solutions/edu/whitepapers/digitaltoolkit.html)
Strategies for building digitized collections. Abby Smith. Council on Library and Information Resources. September 2001: (http://www.clir.org/)
A framework for building good digital collections. Institute of Museum and Library Services (IMLS). November 6, 2001: (http://www.imls.gov/scripts/text.cgi?/pubs/forumframework.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
Andrew Hampson et al. Digitisation of exam papers. The Electronic Library, 17,4;Aug 1999;239-46. Discusses complete workflow, project planning and management for digitizing and providing intranet access to exam papers