Information Management Resource Kit

Module on Digitization and Digital Libraries

UNIT 4. CREATION AND MANAGEMENT OF DIGITAL DOCUMENTS

LESSON 3. BASIC FACILITIES AND REQUIREMENTS FOR DIGITIZATION

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 equipment, software, human resources and funds required to digitize hardcopy documents.

Introduction

We all agree that we should start digitizing our documents. We've also decided which documents to digitize.
Now, let’s work out what we need to do …

You will need various types of resources in order to digitize documents to include in a digital library.

What resources do you need?

How much will they cost?

In this lesson, we will give you some guidelines to help you determine what you will need.

Requirements

The following table lists the types of resources you may need to digitize your documents:

Equipment	• Scanners, computers and storage devices • Audio and video capture equipment (if you are handling recordings)
Software	• Scanning • Optical character recognition • Word processing • Spellchecking • Image management • Video and audio capture (if you are handling recordings)
Human resources	Personnel and skills
Funds	To cover: • salaries • equipment • software • running costs, and so on

Let's analyze each of these items in detail...

Equipment

The first thing you need is 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	WHEN TO USE
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**.

Computer

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 volume 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**.

CD-writer

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); or

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 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 (or tapes or USB drives).

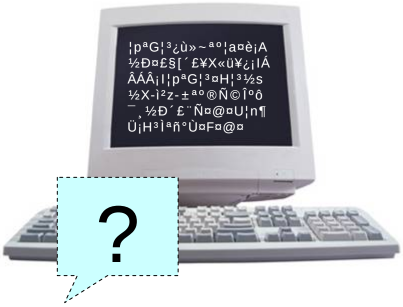
Software

You will need the following types of **software** (those marked * are free) :

Software type	Purpose	Examples
Scanning and OCR	To convert the hardcopy image to a digital one, and then into text that a word processor can understand A 'lite' version of scanning and OCR software is normally provided when you buy a scanner	ReadIris, OmniPage, FineReader
Word processor and spellchecker	To correct text errors and to optimize page layout	Microsoft Word, Corel WordPerfect, OpenOffice*
File conversion	To convert files from one format to another	Microsoft Word Many open source converters available
Image management	To view, modify and manage images	CompuPic, Kudo, ACDSee, Irfanview
Image editing	To edit images	Adobe PhotoShop, Corel PhotoPaint, Microsoft PhotoDraw ImageMagic*
PDF creation	To create PDF documents	Adobe Acrobat , PDF-PHP*, PDFCreator*, PDF995*, CutePDF Writer*
PDF viewing	To read PDF documents	Adobe Reader*

Language

You may be dealing with languages that use Roman scripts with a lot of **accented characters** (such as á, å, etc.) and **non-Roman scripts** (Arabic, Chinese, Cyrillic, etc.). If so, the software you are using might have problems recognizing, correcting and representing characters in these scripts.



You can take the following precautions to solve these problems:

- seek OCR software that is **specific for your language**;
- set up a language-specific **dictionary** in your spellchecking or word processing program (you can create a language-specific dictionary for Microsoft Word); and
- if you are not using Unicode, find programs that **convert** from other encoding systems to Unicode. This means that users will not have to download special fonts to read the text.

Personnel

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

A **manager** to coordinate the team and the digitization workflow.

Staff assigning **metadata**. Skilled librarians familiar with the subject matter are best for this task.

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.

Personnel

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

Staff to do **scanning**, **OCR**, **proofreading** and **layout**. They must have computer skills, and understand the language of the documents they are working with. They do not have to be experts in the subject matter of the documents – though this is an advantage.
Scanning, OCR, proofreading and layout require **diligence** and **concentration**. The best results and productivity come during a limited number of hours each day. It may be best to organize it on a **part-time** basis. If you have full-time staff, employ only experienced, highly motivated and quality-conscious people.

Logistical and **secretarial** staff. They will have to obtain the documents, clean and sort them if necessary, cut the bindings and rebind them (if you do this), and return them to their original location.

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

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

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

Equipment and software	Scanner, computers, software, office furniture.
Document acquisition	Registration, categorization, mailing and transport costs, staff time.
Scanning	Staff time, photocopying (if you photocopy documents before scanning them).
OCR, proofreading and layout	Staff time, consumables (disks, paper).
Metadata assignment	Staff time (depends on the number of documents, the difficulty of the subject, and the salaries of the specialists).
Management and overhead	Management, overhead, staff training.
Contingency	Additional, unanticipated expenses.

On the next screens we will focus on costs for scanning and OCR, but remember to cover all the costs listed in the table!

Costs

The total cost 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 skilled at using computers in your organization and enter this amount (in dollars) in the following field:

US $

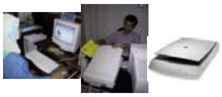
To calculate the estimated cost of scanning **per page**, click on the Scanning Costs button (see annex_costs.pdf):

Scanning Costs

To calculate the estimated cost of OCR, proofreading and layout **per page**, click on the OCR Costs button (see annex_costs.pdf):

OCR Costs

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 $

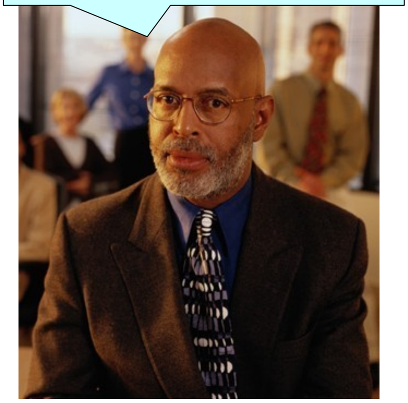
Then, click on the icons to view the estimated costs for each scenario.

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

These estimates are based on Loots et al., From Paper to Collection, 2004.

Outsourcing

Now that we can estimate costs and staffing considerations, we can determine the best overall approach!



Taking the previous scenarios as a starting point, you can determine the best approach and combination of resources for your needs. You may want to consider outsourcing the job.

Outsourcing could be the best choice if:

- you have a “one-off” job, not an ongoing activity; and

- you must scan many pages but you cannot justify buying an expensive professional scanner.

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

Guidelines and procedures

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

You may use them as tools for your job.

Software needed to digitize documents (see Software-M.pdf)

Cost categories (see costs_tot.pdf)

Summary

To digitize hardcopy publications, you will need adequate equipment, software, human resources and funds.

The type and amount of equipment you will need depends on the **number of pages** to be digitized.

You will need a variety of **software**, but you can get much of this for free if you are willing to use **open source** software.

Dealing with different **languages** in hardcopy documents is an issue you have to consider.

The costs of the digital library depend on the **number of pages** to be scanned and the **salary cost** of skilled staff. Consider **outsourcing** this task if you cannot do it in-house.

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

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 key issue did Mr. Touré not mention?

Please type your answer in the box and press Check Answer.

Exercise 2

Mr. Touré has listed several items to include in his digital library budget:

- Equipment and software
- Document acquisition
- Scanning
- OCR, proofreading and layout
- Management and overhead
- Contingency

What has he forgotten?

Please type your answer in the box and press Check Answer.

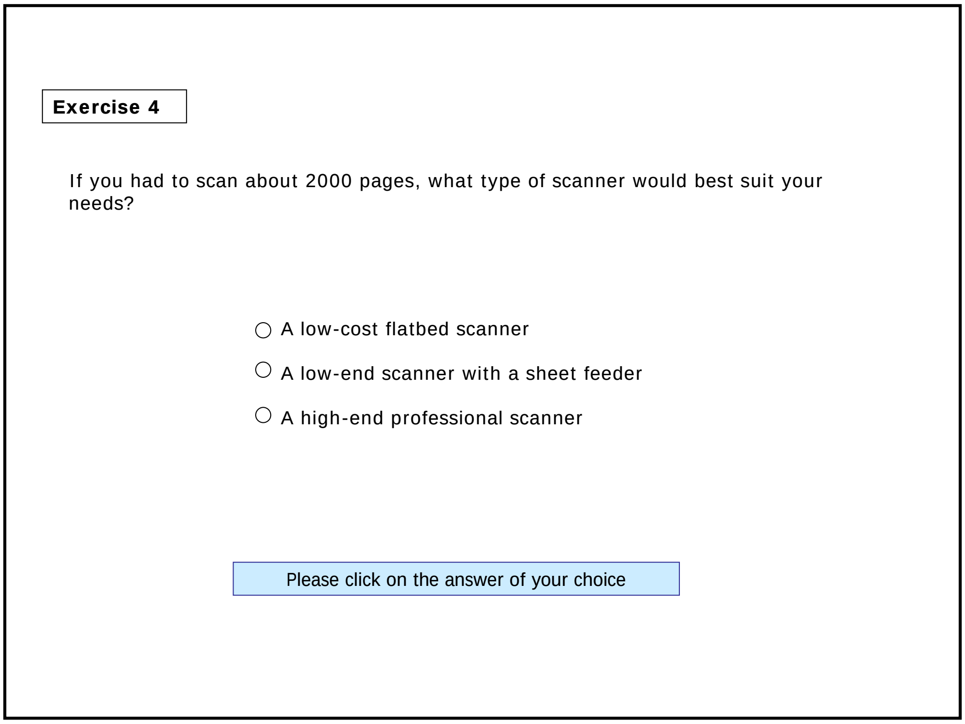
Exercise 3

My Word Processor does not handle the language of these documents, so there’s no point in trying to spellcheck them…

○ True

○ False

Please click on the answer of your choice



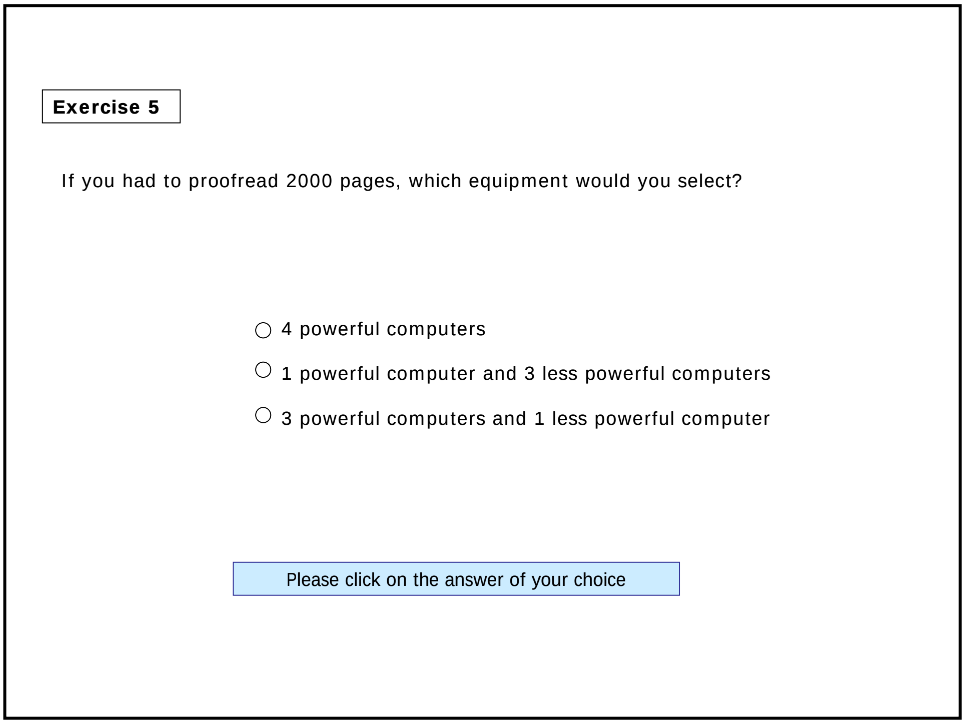
Exercise 4

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



Exercise 5

If you had to proofread 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

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.

Please select the options of your choice (2 or more)
and press Check Answer

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)
PDF995: (http://www.pdf995.com)
Adobe Reader: (http://www.adobe.com)
Guide to Digital Scientific Artwork: (http://www.mlab.nl/GtoDSA/Start.htm)
CompuPic: (http://www.photodex.com)
ACDSee: (http://www.acdsystems.com/English/index.htm)
Irfanview: (http://www.irfanview.com)
PDF-PHP: (http://sourceforge.net/projects/pdf-php)
PDFCreator: (http://sourceforge.net/projects/pdfcreator)
CutePDF Writer: (http://www.cutepdf.com/Products/CutePDF/writer.asp)

Open source software is freely available from a number of websites. Here is a list of them:

OpenOffice.org - OpenOffice is an open source (free) suite of software available in various languages which includes a word processor, spreadsheet, presentation and drawing software with PDF capabilities: (http://www.openoffice.org)
The UNESCO Free Software Portal: www.unesco.org/webworld/portal_freesoft/index.shtml: (http://www.unesco.org/webworld/portal_freesoft/index.shtml)
SourceForge.net: http://sourceforge.net: (http://sourceforge.net/)
Freshmeat: http://freshmeat.net: (http://freshmeat.net/)
Open Source System for Libraries: http://www.oss4lib.org: (http://www.oss4lib.org)
The World Wide Web Consortium (W3C): http://www.w3.org: (http://www.w3.org/)
Open Source and Linux News and Software: http://osdir.com: (http://osdir.com/)

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.