UNIT 7 DIGITISATION: CONCEPT, NEED, METHODS AND EQUIPMENT

Structure

7.0 Objectives

7.1 Introduction

7.2 Digitisation: Basics

7.2.1 Definition

7.2.2 Need for Digitisation

7.3 Selection of Materials for Digitisation

7.4 Steps in the Process of Digitisation

7.4.1 Scanning

7.4.2 Indexing

7.4.3 Storing

7.4.4 Retrieving

7.5 Digitisation: Input and Output Options

7.5.1 Scanning as Image Only

7.5.2 Optical Character Recognition (OCR) and Retaining Page Layout

7.5.3 Retaining Page Layout using Acrobat Capture

7.5.4 Re-keying

7.6 Technology of Digitisation

7.6.1 Bit Depth or Dynamic Range

7.6.2 Resolution

7.6.3 Threshold

7.6.4 Image Enhancement

7. 7 Compression

7.7.1 Lossless Compression

7.7.2 Lossy Compression

7.7.3 Compression Protocols

7.8 File Formats and Media Types

7.8.1 Formats and Encoding used for Text

7.9 Tools of Digitisation

7.9.1 Scanners

7.9.2 Scanning Software

7.10 Digitisation of Audio and Video

7.11 Organising Digital Images

7.12 Digital Library Softwares

7.13 Planning and Implementation

7.13.1 Feasibility

7.13.2 Planning the Project

7.13.3 Purchase of Hardware and Software

7.13.4 Selection of Material for Digitisation and ‘Born Digital’

7.13.5 Placement and Training of Manpower
7.13.6 Content Creation
7.13.7 Execution of the Project
7.14 Summary
7.15 Answers to Self Check Exercises
7.16 Keywords
7.17 References and Further Reading

7.0 OBJECTIVES

After reading this Unit you will know about the following concepts relate to the technology of digitisation:

- digitisation: basics, concept and need;
- steps in the process of digitisation;
- technology of digitisation;
- compression;
- Optical character recognition (OCR);
- file formats;
- tools of digitisation; and
- organising digital images.

7.1 INTRODUCTION

All recorded information in a traditional library is analogue in nature. The analogue information can include printed books, periodical articles, manuscripts, cards, photographs, vinyl disks, video and audiotapes. However, when analogue information is fed into a computer, it is broken down into 0s and 1s changing its characteristics from analogue to digital. These bits of data can be re-combined for manipulation and compressed for storage. Voluminous encyclopaedias that take-up yards of shelf-space in analogue form can fit into a small space on a computer drive or stored on to a CD ROM disc, which can be searched, retrieved manipulated and sent over the network. One of the most important traits of digital information is that it is not fixed in the way that texts printed on a paper are. Digital texts are neither final nor finite, and are not fixed either in essence or in form except, when it is printed out as a hard copy.

Flexibility is one of the chief assets of digital information. An endless number of identical copies can be created from a digital file, because a digital file does not decay by copying. Moreover, digital information can be made accessible from remote location simultaneously by a large number of users.

Digitisation is the process of converting the content of physical media (e.g., periodical articles, books, manuscripts, cards, photographs, vinyl disks, etc.) into digital format. In most library applications, digitisation normally results in documents that are accessible from the web site of a library and thus, on the Internet. Optical scanners and digital cameras are used to digitise images by translating them into bit maps. It is also possible to digitise sound, video, graphics and animations, etc.

Digitisation is not an end in itself. It is the process that creates a digital image from an analogue image. Selection criteria, particularly those, which reflect user needs, are of paramount importance. Therefore, the principles that are applicable in traditional collection development are applicable when materials are being selected for digitisation. However, there are several other considerations related to technical, legal, policy, and resources that become important in a digitisation project.

Digitisation is one of the three important methods of building digitised collections. The other two methods include providing access to electronic resources (whether free or licensed) and creating library portals for important Internet resources.

7.2 DIGITISATION: BASICS

7.2.1 Definition

The word "digital" describes any system based on discontinuous data or events. Computers are digital machines because at their most basic level they can distinguish between just two values, 0 and 1, or off and on. All data that a computer processes must be encoded digitally as a series of zeroes and ones.

The opposite of digital is analogue. A typical analogue device is a clock in which the hands move continuously around the face. Such a clock is capable of indicating every possible time of the day. In contrast, a digital clock is capable of representing only a finite number of times (every tenth of a second, for example).

As mentioned before, a printed book is analogue form of information. The contents of a book need to be digitised to convert it into digital form. Digitisation is the process of converting the content of physical media (e.g., periodical articles, books, manuscripts, cards, photographs, vinyl disks, etc.) to digital formats.

Digitisation refers to the process of translating a piece of information such as a book, journal articles, sound recordings, pictures, audio tapes or videos recordings, etc. into bits. Bits are the fundamental units of information in a computer system. Converting information into these binary digits is called digitisation, which can be achieved through a variety of existing technologies. A digital image, in turn, is composed of a set of pixels (picture elements), arranged according to a pre-defined ratio of columns and rows. An image file can be managed as a regular computer file and can be retrieved, printed and modified using appropriate software. Further, textual images can be OCRed so as to make its contents searchable.

An image of the physical object is captured using a scanner or digital camera and converted into digital format that can be stored electronically and accessed via computers. The process of digitisation, however does not stop at scanning of physical objects, a considerable amount of work is involved in optimising usage of digitised documents. Sometimes, these post- scanning processes are often assumed in the meaning of digitisation. At other times the word "digitisation" is used in restricted sense to include only the process of scanning.

7.2.2 Need for Digitisation

Digitising a document in print or other physical media (e.g., sound recordings) makes the document more useful as well as more accessible. It is possible for a user to conduct a full-text search on a document that is digitised and OCRed. It is possible to create hyperlinks to lead a reader to related items within the text itself as well as to external resources. Ultimately, digitisation does not mean replacing the traditional library collections and services; rather, it serves to enhance them.

A document can be converted into digital format depending on the objective of digitisation, end user, availability of finances, etc. While the objectives of digitisation initiatives differ from organisation to organisation, the primary objective is to improve the access. Other objectives include cost savings, preservation, keeping pace with technology and information sharing. The most significant challenges in planning and execution of a digitisation project relate to technical limitations, budgetary constraints, copyright considerations, lack of policy guidelines and lastly, the selection of materials for digitisation.

While new and emerging technologies allow digital information to be presented in innovative ways, the majority of potential users are unlikely to have access to sophisticated hardware and software. Sharing of information among various institutions is often restricted by the use of incompatible software.

One of the main benefits of digitisation is to preserve rare and fragile objects by enhancing their access to multiple users simultaneously. Very often, when an object is rare and precious, access is only allowed for a certain category of people. Going digital could allow more users to enjoy the benefit of access. Although, digitisation offers great advantages for access like, allowing users to find, retrieve, study and manipulate material, it cannot be considered as a good alternative for preservation because of ever changing formats, protocols and software used for creating digital objects.

There are several reasons for libraries to go for digitisation and there are as many ways to create the digitised images, depending on the needs and uses. The prime reason for digitisation is the need of the user for convenient access to high quality information. Other important considerations are:

Quality Preservation: The digital information has potential for qualitative preservation of information. The preservation-quality images can be scanned at high resolution and bit depth for best possible quality. The quality remains the same inspite of multiple usages by several users. However, caution needs to be exercised while choosing digitisation for preservation of information.

Multiple Referencing: Digital information can be used simultaneously by several users at a time.

Wide Area Usage: Digital information can be made accessible to distant users through the computer networks over the Internet.

Archival Storage: Digitisation is used for restoration of rare material. The rare books, images or archival material are kept in digitised format as a common practice.

Security Measure: Valuable documents and records are scanned and kept in digital format for safety and security.

Self Check Exercise

1) Define digitisation. What are the major benefits of digitisation?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

7.3 SELECTION OF MATERIALS FOR DIGITISATION

To begin the process of digitisation, first of all, we need to select documents for digitisation. The proccess of selection of material for digitisation invloves identification, selection and prioritisation of documents that are to be digitised. If an organisation generates contents, strategies may be adopted to capture data that is "born digital". If documents are available in digital form, it can be easly converted into other formats. If the selected material is from the external sources, IPR issues need to be resolved. If material being digitised is not available in public-domain, then it is important to obtain permission from the publishers and data suppliers for digitisation. The IPR issues must be addressed early in the selection process. Getting permissions from the publishers and individuals could be time consuming, difficult and may involve negotiation and payment of copyright fees. Moreover, decision may be taken whether to OCR the digitised images. Documents selected for digitisation may already be available in digital format. It is always economical to buy e-media, if available, than their conversion. Moreover, over-sized material, deteriorating collections, bound volumes of journals, manuscripts, etc. would require highly specialized equipment and highly skilled manpower.

The documents to be digitised may include text, line art, photographs, colour images, etc. The selection of documents needs to be reviewed very carefully considering all the factors of utility, quality, security and cost. Rare and much–in-demand documents and images are selected as first priority without considering the quality. Factors that may be considered for selecting appropriate media for digitisation include the following:

Audio: The sound quality has to be checked and required corrections made together by the subject expert and computer sound editor.

Video: The video clippings are normally edited on Beta max tapes, which can be used for transferring on to digital format. While editing colour tone, resolution is checked and corrected.

Photographs: The selection of photographs is very crucial process. High resolution is required for photographic images and slides. Also, the quality and future needs are to be checked and the copyright aspects are to be taken care of.

Documents: Documents which are much in demand, too fragile to handle, and rare in availability are reviewed and selected for the process. If the correction of literary value demands much input, then documents are considered for publication rather than digitisation. Moreover, the purpose of all digitisation is related to increased access to digitised materials and value addition. The first consideration for digitisation of documents should be intellectual significance of contents in terms of quality, authority, uniqueness, timeliness, and demand. The intellectual contents, physical nature of the source materials, number of current and potential users are therefore, major considerations.

Self Check Exercise

2) Describe criteria used for selection of material for digitisation.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

....

7.4 STEPS IN THE PROCESS OF DIGITISATION

The following four steps are involved in the process of digitisation. Software, variably called Document Image Processing (DIP), Electronic Filing System (EFS) and Document Management System (DMS), provides all or most of these functions.

7.4.1 Scanning

Electronic scanners are used for acquisition of an electronic image into a computer from its original that may be a photograph, text, manuscript, etc. An image is "read" or scanned at a predefined resolution and dynamic range. The resulting file, called "bit-map page image" is formatted (image formats described elsewhere) and tagged for storage and subsequent retrieval by the software package used for scanning. Acquisition of image through fax card, electronic camera or other imaging devices is also feasible. However, image scanners are most important and most commonly used component of an imaging system for the transfer of normal paper-based documents.

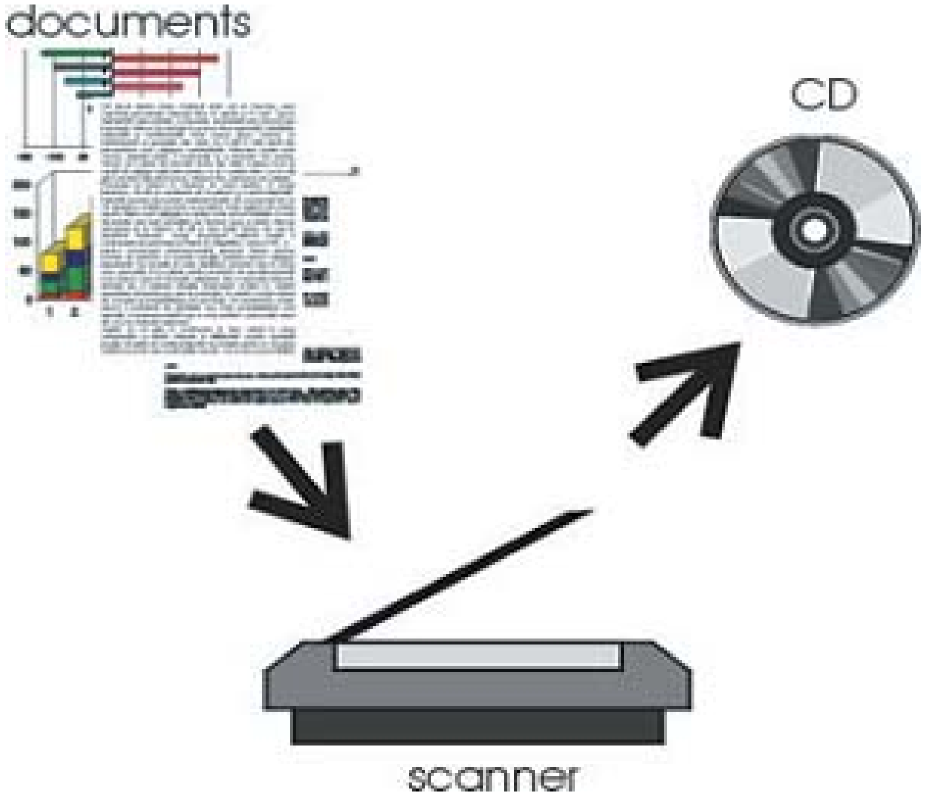


Fig. 7.1: Scanning using a Flatbed Scanner

Steps in the Process of Scanning using a Flatbed Scanner

Step 1 Place picture on the scanner's glass

Step 2 Start scanner software

Step 3 Select the area to be scanned

Step 4 Choose the image type

Step 5 Sharpen the image

Step 6 Set the image size

Step 7 Save the scanned image using a desirable format (GIF or JPEG)

Self Check Exercise

3) Describe the steps involved in the process of scanning a document using a flatbed scanner.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

7.4.2 Indexing

If converting a document into an image or text file is considered as the first step in the process of imaging, indexing these files comprises the second step. The process of indexing scanned images involves linking of the database of scanned images to a text database. Scanned images are just like a set of pictures that need to be related to a text database describing them and their contents. An imaging system typically stores a large amount of unstructured data in a two file system for storing and retrieving scanned images. The first is traditional file that has a text description of the image (keywords or descriptors) along with a key to a second file. The second file contains the document location. The user selects a record from the first file using a search algorithm. Once the user selects a record, the application program keys into the location index, finds the document and displays it.

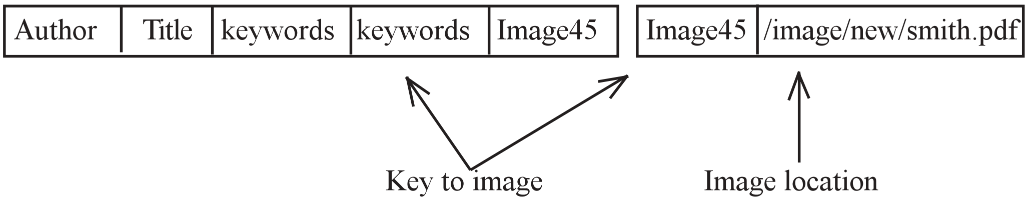


Fig. 7.2: Two File System in a Image Retrieval System

Most of the document imaging software packages through their menu driver or command driven interface, facilitate elaborate indexing of documents. While some document management systems facilitate selection of indexing terms from the image file, others allow only manual keying in of indexing terms. Further, many DMS packages provide OCRed capabilities for transforming the images into standard ASCII files. The OCRed text then serves as a database for full-text search of the stored images.

7.4.3 Storing

The most tenacious problem of a document image relates to its file size and, therefore, to its storage. Every part of an electronic page image is saved regardless of the presence or absence of ink. The file size varies directly with scanning resolution, the size of the area being digitised and the style of graphic file format used to save the image. The scanned images, therefore, need to be transferred from the hard disc of scanning workstation to an external large capacity storage devices such as an optical disc, CD ROM/DVD ROM disc, snap servers, etc. While the smaller document imaging system may use offline media, which need to be reloaded when required, or fixed hard disc drives allocated for image storage, larger document management systems use auto-changers such as optical jukeboxes and tape library systems. The storage required by the scanned images varies and depends upon factors such as scanning resolution, page size, compression ratio and page content. Further, the image storage device may be either remote or local to the retrieval workstation depending upon the imaging system and document management system used.

7.4.4 Retrieving

Once scanned images and OCRed text documents have been saved as a file, a database is needed for selective retrieval of data contained in one or more fields within each record in the database. Typically, a document imaging system uses at least two files to store and retrieve documents. The first is traditional file that has a text description of the image along with a key to the second file. The second file contains the document location. The user selects a record from the first-file using a search algorithm. Once the user selects a record, the application program keys into the location index, finds the document and displays it. Most of the document management systems provide elaborate search possibilities including use of Boolean and proximity operators (AND, OR, NOT) and wild cards. Users are also allowed to refine their search strategy. Once the required images have been identified their associated document image can quickly be retrieved from the image storage device for display or for getting printed output.

7.5 DIGITISATION: INPUT AND OUTPUT OPTIONS

A document can be converted into digital format depending on the objective of digitisation, end users, availability of finances, etc. There are four basic approaches that can be adapted to transform from print to digital:

- Scanning as Image Only
- OCRing and Retaining Page Layout
- Retaining Page Layout using Acrobat Capture; and
- Re-keying the Data

7.5.1 Scanning as Image Only

'Image only' is the lowest cost option in which each page is an exact replica of the original source document. Several digital library projects are concerned with providing digital access to materials that already exists in printed media in traditional libraries. Scanned page images are practically the only reasonable solution for institutions such as libraries for converting existing paper collection (legacy documents) without having access to the original data in computer processible formats convertible into HTML / SGML or in any other structured or unstructured text. Scanned page images are natural choice for large-scale conversions for major digital library initiatives. Printed text, pictures and figures are transformed into computer-accessible forms using a digital scanner or a digital camera in a process called document imaging or scanning. The digitally scanned images are stored in a file as bit-mapped page images, irrespective of the fact that a scanned page contains a photograph, a line drawing or text. A bit-mapped page image is a type of computer graphic, literally an electronic picture of the page which can most easily be equated to a facsimile image of the page and as such they can be read by humans, but not by the computers, understably "text" in a page image is not searchable on a computer using the present-day technology. An image-based implementation requires a large space for data storage and transmission.

Capturing page image format is comparatively easy and inexpensive, therefore, it is a faithful reproduction of its original maintaining page integrity and originality. The scanned textual images, however, are not searchable unless it is OCRed, which in itself, is highly error prone process specially when it involves scientific texts. Options of technology for converting print to digital are given separately.

If OCR is not carried out, the document is not searchable. Most scanning softwares generate TIFF format by default, which, can be converted into PDF using a number of software tools. Scanning to TIFF / PDF format is recommended only when the requirement of project is to make documents portable and accessible from any computing platform. The image can be browsed through a table of contents file composed in HTML that provides link to scanned image objects.

7.5.2 Optical Character Recognition (OCR) and Retaining Page Layout

The latest versions of both Xerox's TextBridge and Caere's Omnipage incorporate technology that allow the option of maintaining text and graphics in their original layout as well as plain ASCII and word-processing formats. Output can also include HTML with attributes like bold, underline, and italic which are retained.

Retaining Layout after OCR

A scanned document is nothing more than a picture of a printed page. It cannot be edited or manipulated or managed based on their contents. In other words, scanned documents have to be referred to by their labels rather than characters in the documents. OCR (Optical Character Recognition) programs are software tools used to transform scanned textual page images into word processing file. OCR or text recognition is the process of electronically identifying text in a bit-mapped page image or set of images and generate a file containing that text in ASCII code or in a specified word processing format leaving the image intact in the process.

7.5.3 Retaining Page Layout using Acrobat Capture

The Acrobat Capture 2.0 provides several options for retaining not only the page layout but also the fonts, and to fit text into the exact space occupied in the original, so that the scanned and OCRed copy never over- or under-shoots the page. Accordingly, it treats unrecognisable text as images that are pasted in its place. Such images are perfectly readable by anyone by looking at the PDF file, but will be absent from the editable and searchable text file. In contrast, ordinary OCR programs treat unrecognised text as wild or some other special character in the ASCII output. Acrobat Capture can be used to scan pages as images, image +text and as normal PDF, all the three options retain page layout.

i) Image Only: Image only option has already been described in Section 7.5.1

ii) Image + Text: In image+text solutions, OCRed text is generated for each image where each page is an exact replica of the original and left untouched, however, the OCRed text sits behind the image and is used for searching. The OCRed text is generally not corrected for errors since; it is used only for searching. The cost involved is much less than PDF Normal. However, the entire page is a bitmap and neither fonts nor line drawings are vectorised, so the file size of Image + Text PDFs is considerably larger than the corresponding PDF Normal files and pages will not display as quickly or cleanly on screen.

iii) PDF Normal: PDF normal gives the clear view on-screen display. It is searchable, with significantly smaller file size than Image+Text. The result is not, however, an exact replica of the scanned page. While all graphics and formatting are preserved, substitute fonts may be used where direct matches are not possible. It is a good choice when files need to be posted on to the web or otherwise delivered online. If during the Capture and OCR process, a word cannot be recognised to the specified confidence level, Capture, by default, substitutes a small portion of the original bitmap image. Capture "best guess" of the suspect word lies behind the bitmap so that searching and indexing are still possible. However, one cannot guarantee that these bitmapped words are correctly guessed. In addition, the bitmap is somewhat obtrusive and detracting from the 'look' of the page. Further, Capture provides option to correct suspected errors left as bit-mapped image or leave them untouched.

7.5.4 Re-keying

A classic solution of this kind would comprise of keying-in the data and its verification. This involves a complete keying of the text, followed by a full re-keying by a different operator, the two keying-in operations might take place simultaneously. The two keyed files are compared and any errors or inconsistencies are corrected. This would guarantee at least 99.9% accuracy, but to reach 99.955% accuracy level, it would normally require full proof-reading of the keyed files, plus table lookups and dictionary spell checks.

Fig. 7.3: Re-keying-in as an Option for Digitisation

7.6 TECHNOLOGY OF DIGITISATION

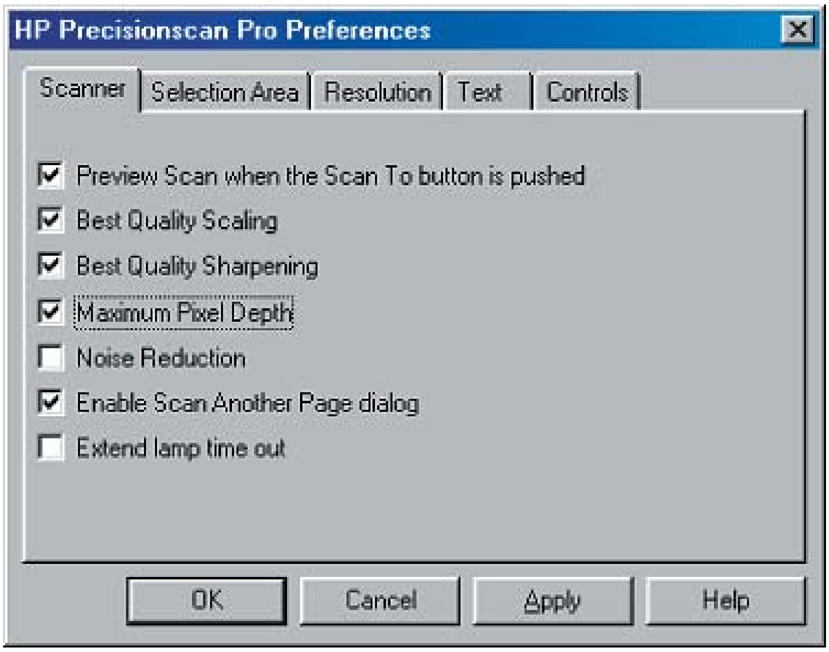
Digital images, also called "bit-mapped page image" are "electronic photographs" composed or set of bits or pixels (picture elements) represented by "0" and "1". A bit-mapped page image is a true representation of its original in terms of typefaces, illustrations, layout and presentation of scanned documents. As such, information or contents of "bit-mapped page image" cannot be searched or manipulated unlike text file documents (or ASCII). However, an ASCII file can be generated from a bit-mapped page image using an optical character recognition (OCR) software such as Xerox's TextBridge and Caere's OmniPage. The quality of digital image can be monitored at the time of capture by the following factors:

- Bit depth / Dynamic Range
- Resolution
- Threshold
- Image Enhancement

Terminology associated with technological aspects of digitisation described below is given in the keywords. Students are advised to understand the terminology, specially bit, byte and pixel before going through the Unit.

7.6.1 Bit Depth or Dynamic Range

The number of bits used to define each pixel determines the bit depth. The greater the bit depth, the greater the number of gray scale or colour tones that can be represented. Dynamic range is the term used to express the full range of total variations, as measured by a densitometer between the lightest and the darkest of a document. Digital images can be captured at varied density or bits per pixel depending upon (i) the nature of source material or document to be scanned; (ii) target audience or users; and (iii) capabilities of the display and print subsystem that are to be used. Bitonal or black & white or binary scanning is generally employed in libraries to scan pages containing text or the drawings. Bitonal or binary scanning represents one bit per pixel (either "0" (black) or "1" (white). Gray scale scanning is used for reliable reproduction of intermediate or continuous tones found in black & white photographs to represent shades of grey. Multiple numbers of bits ranging from 2-8 are assigned to each pixel to represent shades of grey in this process. Although each bit is either black or white, as in the case of bitonal images, but bits are combined to produce a level of grey in the pixel that is, black, white or somewhere in between.



☑ Preview Scan when the Scan To button is pushed
☑ Best Quality Scaling
☑ Best Quality Sharpening
☑ Maximum Pixel Depth
□ Noise Reduction
☑ Enable Scan Another Page dialog
□ Extend lamp time out

Fig. 7.4: Setting Bit Depth in Precisionscan Pro Scanning Software

Lastly, colour scanning can be employed to scan colour photographs. As in the case of grey-scale scanning, multiple bits per pixels typically 2 (lowest quality) to 8 (highest quality) per primary colour are used for representing colour. Colour images are evidently more complex than grey scale images, because it involves encoding of shades of each of the three primary colours, i.e., red, green and blue (RGB). If a coloured image is captured at 2 bits per primary colour, each primary colour can have 2^2 or 4 shades and each pixel can have 4^3 shades for each of the three primary colours. Evidently, increase in bit depth increases the quality of image captured and the space required to store the resultant image. Generally speaking, 12 bits per pixel (4 bits per primary colour) is considered minimum pixel depth for good quality colour image. Most of today's colour scanners can scan at 24-bit colour (8 bit per primary colour).

Table 7.1: No. of Bits Used for Representing Shades in Colour and Gray-scale Scanning

Sl. No.	No. of Bits	No. of bits/shades	No. of shades	No. of Shades/pixel
1	2	2	$2^2 = 4$	$4^3 = 64$
2	4	3	$2^3 = 8$	$8^3 = 512$
3	8	4	$2^4 = 16$	$16^3 = 4096$
4	16	5	$2^5 = 32$	$32^3 = 32768$
4	32	6	$2^6 = 64$	$64^3 = 262144$
5	64	7	$2^7 = 128$	$128^3 = 2097152$
6	128	8	$2^8 = 256$	$256^3 = 16777216$

Self Check Exercise

4) Why is “Bit depth” not important for bitonal scanning?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

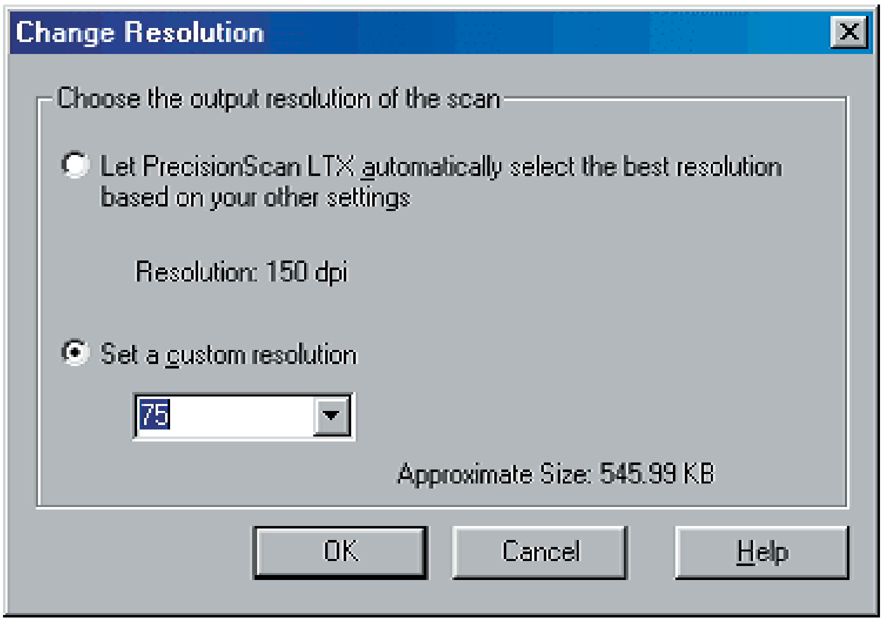
....

7.6.2 Resolution

The resolution of an image is defined in terms of number of pixels (picture elements) in a given area. It is measured in terms of dots per inch (dpi) in case of an image file and as ratio of number of pixels on horizontal line x number of pixel in vertical lines in case of display resolution on a monitor. The higher the dpi set on the scanner, the better the resolution and quality of image and larger the image file.

Regardless of the resolution, the quality of an image can be improved by capturing an image in grayscale. The additional gray-scale data can be processed electronically to sharpen edges, file-in characters, remove extraneous dirt, remove unwanted page strains or discoloration, so as to create a much higher quality image than possible with binary scanning alone. A major drawback in gray scale is that there is large amount of data capture. It may be noted that continuing increase in resolution will not result in any appreciable gain in image quality after some time, except for increase in file size. It is thus important to determine the point where sufficient resolution has been used to capture all the significant details present in the source document.

The black and white or bitonal images (textual) are scanned most commonly at 300 dpi that preserve 99.9% of the information content of a page and can be considered as adequate access resolution. Some preservation projects scan at 600 dpi for better quality. A standard SVGA/VGA monitor has a resolution of 640 × 480 lines while the ultra-high monitors have a resolution of about 2048 × 1664 (about 150 dpi).



Change Resolution

Choose the output resolution of the scan

○ Let PrecisionScan LTX automatically select the best resolution based on your other settings

Resolution: 150 dpi

● Set a custom resolution

75

Approximate Size: 545.99 KB

Fig. 7.5: Setting-up Resolution Manually

Self Check Exercise

5) What is resolution? Is print resolution different from scanning resolution?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

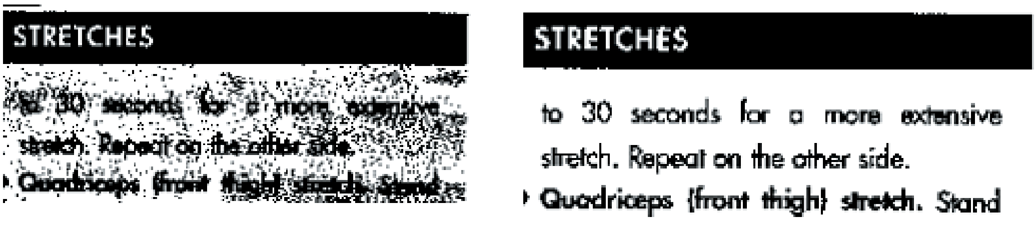
....

7.6.3 Threshold

The threshold setting in bitonal scanning defines the point on a scale, usually ranging from 0 - 255, at which gray values will be interpreted as black or white pixels. In bitonal scanning, resolution and threshold are the key determinants of image quality. Bitonal scanning is best suited to high-contrast documents, such as text and line drawings. Gray scale or colour scanning is required for continuous tone or low contrast for documents such as photographs. In gray scale/colour scanning both resolution and bit depth combine to play significant role in image quality.

In line art mode, every pixel has only two possible values. Every pixel will either be black or white. The Line art **Threshold** control determines the decision point about brightness determining if the sampled value will be a black dot or a white dot. The normal threshold default is 128 (the midrange of the 8-bit 0 - 255 range). Image intensity values above the threshold are white pixels, and values below the threshold are black pixels. Adjusting threshold is like a brightness setting to determine what is black and what is white.

Threshold for text printed on a coloured background or cheap-quality paper like newsprint has to be kept at lower range. Reducing threshold from 128 to about 85 would greatly improve the quality of scan. Such adjustments would also improve the performance of OCR software.



STRETCHES

to 30 seconds for a more extensive stretch. Repeat on the other side.

▸ Quadriceps (front thigh) stretch. Stand

STRETCHES

to 30 seconds for a more extensive stretch. Repeat on the other side.

▸ Quadriceps (front thigh) stretch. Stand

Threshold Setting = 128 Threshold Setting = 85

Fig. 7.6: Threshold Setting in Bitonal Scanning

7.6.4 Image Enhancement

Image enhancement process can be used to improve scanned images at a cost of image authenticity and fidelity. The process of image enhancement is, however, time consuming; it requires special skills and would invariably increase the cost of conversion. Typical image enhancement features available in a scanning or image editing software include filters, tonal reproduction, curves and colour management, touch, crop, image sharpening, contrast, transparent background, etc. In a page scanned in gray-scale, the text /line art and half tone areas can be decomposed and each area of the page can be filtered separately to maximise its quality. The text area on page can be treated with edge sharpening filters, so as to clearly define the character edges, a second filter could be used to remove the high-frequency noise and finally another filter could fill-in broken characters. Gray-scale area of the page could be processed with different filters to maximise the quality of the halftone.

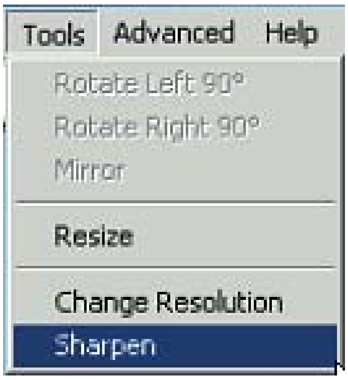


Fig. 7.7: Sharpening Image using HP Precisionscan Pro

7.7 COMPRESSION

Image files are evidently larger than textual ASCII files. It is thus necessary to compress image files so as to achieve economic storage, processing and transmission over a network. A black & white image of a page of text scanned at 300 dpi is about 1 mb in size whereas a text file containing the same information is about 2-3 kb. Image compression is the process of reducing size of an image by abbreviating the repetitive information such as one or more rows of white bits to a single code. The compression algorithms may be grouped into the following two categories:

7.7.1 Lossless Compression

The conversion process converts repeated information as a mathematical algorithm that can be decompressed without loosing any details into the original image with absolute fidelity. No information is “lost” or “sacrificed” in the process of compression. Lossless compression is primarily used in bitonal images.

7.7.2 Lossy Compression

Lossy compression process discards or minimises details that are least significant or which may not make appreciable effect on the quality of image. This kind of compression is called “lossy” because when the image that is compressed using “lossy” compression techniques is decompressed, it will not be an exact replica of the original image. Lossy compression is used with gray-scale/colour scanning.

Compression is a necessity in digital imaging but more important is the ability to output or produce the uncompressed true replica of images. This is especially important when images are transferred from one platform to another or are handled by software packages under different operating systems.

Uncompressed images often work better than compressed images for different reasons. It is thus suggested that scanned images should be either stored as uncompressed images or at the most as lossless compressed images. Further, it is optimal to use one of the standard and widely supported compression protocols than a proprietary one, even if it offers efficient compression and better quality. Attributes of original documents may also be considered while selecting compression techniques. For example ITU G-4 is designed to compress text where as JPEG, GIF and ImagePAC are designed to compress pictures. It is important to ensure migration of images from one platform to another and from one hardware medium to another. It may be noted that highly compressed files are more prone to corruption than uncompressed files.

Self Check Exercise

6) What is image compression? Describe the types of image compression.

Note : i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

7.7.3 Compression Protocols

The following protocols are commonly used for bitonal, gray scale or colour compression:

TIFF-G4

International Telecommunication Union (ITU Group 4) is considered as de facto standard compression scheme for black and white or bitonal images. An image created as a TIFF and compressed using ITU-G4 compression technique is called a Group-4 TIFF or TIFF-G4 and is considered as defacto standard for storing bitonal images. TIFF- G-4 is a lossless compression scheme. Joint Bi-level Image Group (JBIG) (ISO-11544) is another standard compression technique for bitonal images.

JPEG (Joint Photographic Expert Group)

JPEG (Joint Photographic Expert Group) is an ISO-10918-I compression protocol that works by finding areas of the image that have same tone, shade, colour or other characteristics and represents this area by a code. Compression is achieved at loss of data. Preliminary testing indicates that a compression of about 10 or 15 to one can be achieved without visible degradation of image quality.

LZW (Lenpel-Ziv Welch)

LZW compression technique uses a table-based lookup algorithm invented by Abraham Lempel, Jacob Ziv, and Terry Welch. Two commonly-used file formats in which LZW compression is used are the Graphics Interchange Format (GIF) and Tag Image File Format (TIFF). LZW compression is also suitable for compressing text files. A particular LZW compression algorithm takes each input sequence of binary digit of a given length (for example, 12 bits) and creates an entry in a table (sometimes called a "dictionary" or "codebook") for that particular bit pattern, consisting of the pattern itself and a shorter code. As input is read, any pattern that has been read before the results in the substitution of the shorter code effectively compresses the total amount of input to something smaller. The decoding program that uncompresses the file is able to build the table itself by using the algorithm as it processes the encoded input.

Self Check Exercise

7) How does LZW compression protocol work? Which file formats use LZW compression protocol?

Note : i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

OCR (Optical Character Recognition)

OCR (Optical Character Recognition) programs are software tools used to transform scanned textual page images into word processing file. OCR or text recognition is the process of electronically identifying text in a bit-mapped page image or set of images and generate a file containing that text in ASCII code or in a specified word processing format leaving the image intact in the process. The OCR is performed in order to make every word in a scanned document readable and fully searchable without having to key-in everything in the computer manually. Once a bit-mapped page image has gone through the process of OCR, a document can be manipulated and managed by its contents, i.e., using the words available in the text.

OCR does not actually convert an image into text but rather creates a separate file containing the text while leaving the image intact. There are four types of OCR technology that are prevailing in the market. These technologies are: matrix matching, feature extraction, structural analysis and neural network.

i) **Matrix / Template Matching:** Compares each character with a template of the same character. Such a system is usually limited to a specific number of fonts, or must be "taught" to recognise a particular font.

ii) **Feature Extraction:** Can recognise a character from its structure and shape (angles, points, breaks, etc.) based on a set of rules. The process claims to recognise all fonts.

iii) **Structural analysis:** Determines characters on the basis of density gradations or character darkness.

iv) **Neural Networking:** Neural networking is a form of artificial intelligence that attempts to mimic processes of the human mind. Combined with traditional OCR techniques plus pattern recognition, a neural network-based system can perform text recognition and "learn" from its success and failure. Referred to as "Intelligent Character Recognition", a neural network-based system is being used to recognize hand-written text as well as other traditionally difficult source material. Neural network ICR can contemplate characters in the context of an entire word. Newer ICR combines neural networking with fuzzy logic.

The image scanner optically captures text images to be recognised. Text images are processed with OCR software and hardware. The process involves three operations: document analysis (extracting individual character images), recognising these images based on (i) their template stored in the OCR database; (ii) structure and shape (angles, points, breaks, etc.) (iii) density gradations or character darkness and (iv) contextual processing. The output interface is responsible for communication of OCR system that results, to the outside world.

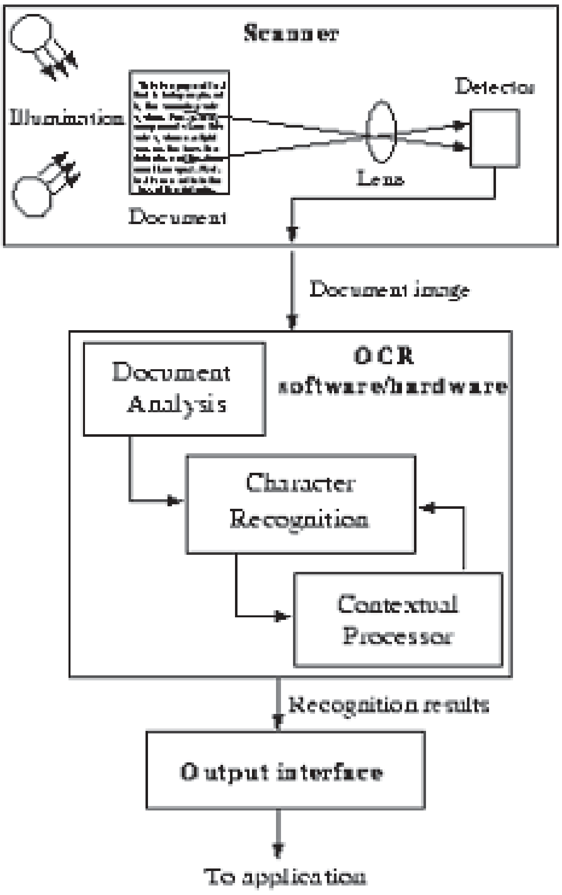


Fig. 7.8: OCR Technology

Several software packages now offer facility of retaining the page layout after it has been OCRed. The process for retaining the page layout is software dependent. Caere's Omnipro offers two ways of retaining page layout following OCR. It calls them True Page Classic and True Page Easy. True Page Classic places each paragraph within a separate frame of a word processor into which the OCR output is saved. If one wishes to edit anything subsequently, then the relevant paragraph box may need to be resized. However, Easy Edit facilitates editing of pages without the necessity of resizing the boxes although there are greater chances of spillage over the page. Xerox Text Bridge offers similar feature called DocuRT which is broadly equivalent to True Page Easy edit. The process of OCR dismantles the page, OCRs it, and then reassembles it in such a way that all the component parts such as tabs, columns, table, graphics can be used in a text manipulation package such as word processor.

There is a little doubt about the fact that OCR is less accurate than rekeying-in the data. At an accuracy ratio of 98%, a page having 1800 characters will have 36 errors per page on an average. It is therefore, imperative to cleanup after OCR unless original scanned image will be viewed as a page and OCR is being used purely to create a searchable index on the words that will be searched via a fuzzy retrieval engine like Excalibur, which is highly tolerant to OCR errors.

Another possibility for cleaned-up OCR is use of a specialist OCR system such as, Prime Recognition. With production OCR in mind, Prime OCR licenses leading to recognising engine and passes the data through several of them using voting technology along with artificial intelligence algorithms. Although it takes longer initially, but saves time in the long run and Prime contends that it improves the result achieved by a single engine by 65 – 80 %. The technology is available at a price depending upon the number of search engines that one would like to incorporate. Michigan Digital Library production services used Prime OCR for placing more than two million pages of SGML – encoded text and the same number of page images on the web.

Self Check Exercise

8) What is OCR? Why is it important to OCR a digitised image?

Note : i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

7.8 FILE FORMATS AND MEDIA TYPES

A defined arrangement for discrete sets of data that allow a computer and software to interpret the data is called a file format. Different file formats are used to store different media types like text, images, graphics, pictures, musical works, computer programs, databases, models and designs, video programs and compound works combining many types of information. Although, almost every type of information can be represented in digital form, a few important file formats for text and images typically applicable to a library-based digital collections are described here. However, every object in a digital library needs to have a name or identifier which distinctly identifies its type and format. This is achieved by assigning file extensions to the digital objects. The file extensions in a digital library typically denote formats, protocols and rights management that are appropriate for the type of material. Names of file formats applicable in digital library and their file extensions are given in Table 7.2.

Self Check Exercise

9) What are file formats? How is " unstructured text" different from " structured text"?

Note : i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

7.8.1 Formats and Encoding Used for Text

Text and image-based contents of a digital library can be stored and presented as (i) simple text or ASCII (American Standard Code for Information Interchange); (ii) unstructured text; (iii) Structured text (SGML or HTML or XML); (iv) page description language and (v) page image formats.

Simple Text or ASCII

Simple text or ASCII (American Standard Code for Information Exchange) is the most commonly used encoding scheme used for facilitating exchange of data from one software to another or from one platform to another. "Full-text" of articles from many journals are available electronically through online vendors like DIALOG and STN in this format since more than two decades.

Simple text or ASCII is compact, economic to capture and store, searchable, inter-operable and is malleable with other text-based services. On the other hand, the simple text or ASCII cannot be used for displaying complex tables or mathematical formulas. Photographs, diagrams, graphics, special characters cannot be displayed in ASCII. ASCII format does not store text formatting information, i.e., italics, bold, font type, font size or paragraph justification information. Simple text or ASCII in many ways is inadequate to represent many journal articles because of the reasons mentioned above. Although simple text or ASCII is extremely useful for searching and selection, its inability to capture the richness of the original makes it an interim step to structured text formats.

Structured Text Format

Structured text format attempts to capture the essence of documents by "marking-up" the text so that the original form could be recreated or even produce other forms such as ASCII. Structured text formats have provision for imbed images, graphics and other multimedia formats in the text. SGML (Standard Generalized Markup Language) is one of the most important and popular structured text format. ODA (Office Document Architecture) is a similar and competing standard. SGML is an international standard (ISO, 1986) around which several related standards are built. SGML is a flexible language that gave birth to HTML (Hyper-Text Markup Language), de facto markup language of the World Wide Web, to control the display format of documents and even the appearance of the user interface for interacting with the documents. Like simple text or ASCII, structured text can be searched or manipulated. It is highly flexible and suitable both for electronic and paper production. Well-formated text increases visual presentation volume of textual, graphical and pictorial information. Structured formats can easily display complex tables and equations. Moreover, structured text is compact in comparison to the image-based formats, even after including imbedded graphics and pictures.

Creation of structured text, if rekeyed, is always too expensive on a production basis. However, creation of structured text is generally integrated with the production of printed artifacts. SGML is in fact, a format generated as a by-product of printed artifacts generated electronically.

Besides SGML and HTML, there are other formats used in digital library implementation. TeX, used for formatting highly mathematical text is one such format which allows greater control over the resulting display of document, including reviewing the formatting of errors.

Page Description Language (PDL)

Page Description Languages (PDLs), such as Adobe's PostScript and PDF (Portable Document Format) are similar to image but the formatted pages displayed to the user are text-based rather than image-based. PostScript and PDF formats can easily be captured during the typesetting process. PostScript is especially easy to capture since most of the systems automatically generate it and conversion program, called Acrobat Distiller, can be used to convert PostScript file into PDF files. The documents stored as PDF require Acrobat Reader at the user's end to read or print the document. The Acrobat Reader can be downloaded free of cost from the Adobe's Web Site.

Acrobat's Portable Document Format (PDF) is a by-product of PostScript. Adobe's page-description language had become the standard way to describe pages electronically in the graphics world. While PostScript is a programming language, PDF is a page-description format.

PDF can have two formats: (i) Text-based PDF that uses outline font technology of PostScript PDL (Page Description Language) from Adobe to describe format of a page; (ii) raster-scanned image PDF without the text output of OCR (Optical Character

Recognition). The image PDF is essentially equivalent to TIFF or CCITT G4 formats or to a photograph where text characters cannot be manipulated by the computer. Besides, an image-based PDF may be converted into text-based PDF once it goes through the process of OCR. In this process, scanned image is replaced by the text with fonts and layout matching with the scanned document.

Self Check Exercise

10) How is Page Description Language (PDL) different from structured text?

Note : i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

....

Page Image Format

The digitally scanned images are stored in a file as a bit-mapped page image, irrespective of the fact that a scanned page contains a photograph, a line drawing or text. The bit-mapped page image can be created in dozens of different formats depending upon the scanner and its software. National and international standards for image-file formats and compression methods exist to ensure that data will be interchangeable amongst systems. An image file stores discrete sets of data and information allowing a computing system to display, interpret and print the image in a pre-defined fashion. An image file format consists of three distinct components, i.e., **header** which stores information on file identifier and image specifications; **Image data** consisting of look-up table and image raster and lastly, **footer** that signals file termination information. While bit-mapped portion of a raster image is standardised, it is the file header that differentiates one format from another.

TIFF (Tagged Image File Format) is the most commonly used page image file format and is considered to be the de facto standard for bitonal images. Some image formats are proprieary developed by commercial vendors and require specific software or hardware for display and printing. Images can be coloured, grey-scale or black and white (called bitonal). They can be uncompressed (raw) or compressed using several different compression algorithms.

Table 7.2: File Formats Used in a Digital Library

Abbreviation	Format	File Extension
	File Format for Unstructured Text	
ASCII	American Standard Code for Information Interchange	.txt
	File Format for Structured Text	
SGML	Standard Generalized Markup Language	.sgml
HTML	Hypertext Markup Language	.html
XML	Extended Markup Language	.xml
PDF	Portable Document Format (Adobe)	.pdf
PostScript	PostScript (Adobe)	.ps
TEX	Texture Format	.txt

Abbreviation	Format	File Extension
File Format for Images		
PDF	Portable Document Format	.pdf
BMP	Bit Map Page (Windows)	.bmp
IMG	Ventura Publisher	.img
JPEG	Joint Photographic Expert Group	.mpg
JFIF	JPEG File Format	.jfif
PCP	PC Paint (B&W Mode)	.pcp
PCX	PC Paint Brush (Color & B&W)	.pcx
PSD	Photoshop	.psd
TGA	True Vision Targa	.tga
PNG	Portable Network Graphic	.png
TIFF	Taged Image File Format	.tif
TIFF-G4	Taged Image File Format with Group 4 Fax Compression	.tif
SPIFF	Still Picture Interchange File Format	.spf
PCD	Photo CD (Kodak)	.pcd
Audio and Video File Format		
WAVE	Waveform Audio (Microsoft)	.wav
AIFF	Audio Interchange Format	.aif
VoC	Creative Voice	.voc
MIDI	Musical Instrument Digital Interface	.midi
SND	Sound	.snd
AU	Audio (Sun Microsystems)	.au
RAF	Real Audio Format (Progressive Networks)	.ra
AVI	Audio Visual Interleave	.avi
FLA	Macromedia Flash Movie	.fla
FLC	AutoDesk FLIC Animation	.flc
MOV	Quicktime for Windows Movie	.mov
MPEG	Motion Picture Expert Group	.mpg
MP2	MPEG Audio Layer 2	.mp2
MP3	MPEG Audio Layer 3	.mp3

7.9 TOOLS OF DIGITISATION

Digital imaging is an inter-linked system of hardware, software, image database and access sub-system with each having their own components. Tools used for digitisation include several core and peripheral systems. An image scanning system may consist of a stand-alone workstation where most or all the work is done on the same workstation or as a part of a network of workstations with imaging work distributed and shared amongst various workstations. The network usually includes a scanning station, a server and one or more editing, retrieval stations. A typical scanning workstation for a small, production-level project could consist of the following:

- Hardware (Scanners, computers, data storage and data output peripherals)
- Software (image capturing and image editing)
- Network (data transmission)
- Display and Printing technologies

This Unit concentrates on scanners and scanning software as important components of the scanning system.

7.9.1 Scanners

Digital scanners are used to capture digital images from analogue media such as printed pages or a microfiche / microfilm at a predefined resolution and dynamic range (bit range). There are two types of image scanners: vector scanner and raster scanners. The vector scanners scan an image as a complex set of x,y coordinates. Vector images are generally used in Geographical Information Systems (GIS). The display software for the vector image interprets the image as function of coordinates and other included information to produce an electronic replica of the original drawing or photograph. Vector images can be zoomed in portion to display minute details of a drawing or a map. Maps, engineering drawings, and architectural blueprints are often scanned as vector images. Raster images are captured by raster scanners by passing lights (laser in some cases) down the page and digitally encoding it row by row. Multiple passes of lights may be required to capture basic (as a set of bits known as bit map) colours in a coloured image. Raster scanners are used in libraries to convert printed publications into electronic forms. Majority of electronic imaging systems generate raster images. The scanners used for digitizing analogue images into digital images come in a variety of shapes and sizes.

How Scanner Works?

Scanners are equipped with a lamp that moves with the scanner head to light-up the object being scanned. Most scanners use a cold cathode florescent lamp or a xenon lamp. The scan head is made up of the mirrors, lens, filter, and charge-coupled devices (CCD) array. A belt that is connected to the stepper motor makes the scan head move. A stabilizing bar prevents wobbling during the pass. The mirrors reflect what is being scanned into the lens and the image is then focused through a filter on the CCD array. Three smaller images of the original are made by the lens. These images then go through a color filter and onto a section of the CCD array. The data is then combined into a single image.

While selecting a scanner, one should consider resolution, sharpness, and rate of image transfer. The resolution is measured in dots per inch (dpi). The average scanner has at least 300x300 dpi. The number of sensors in a row of the CCD array determines a scanner‘s dpi. Sharpness depends on how bright the lamp is and the quality of the lens. Image transfer depends on the connection used to connect the scanner to the computer. The slowest is the parallel port. Universal Serial Bus or USB scanners are affordable, easy to use, and have good speed.

The hardware required for a scanner is a connector such as a USB. The software required is a driver. The driver is needed to communicate with the scanner. TWAIN is the language spoken by scanners. Any program that supports TWAIN can acquire a scanned image.

There are following types of Scanners:

- Flatbed Scanners – right angle, prism and overhead flatbed
- Sheet-Feed Scanners
- Drum Scanners
- Digital Cameras
- Slide Scanners
- Microfilm Scanners
- Video Frame Grabbers
- Hand-held scanners

The type of scanner selected for an imaging project would be influenced by the type, size and source of documents to be scanned. Many scanners can handle only transparent material, whereas others can handle only reflective materials.

Flatbed Scanners: Flatbed scanners are most common, and widely used scanners that look like a photocopier and are used in much the same way. Source material in a flatbed scanner is placed face down for scanning. The light source and charge-coupled devices (CCDs) move beneath the platen, while the document remains stationary as in the case of photocopying machine. Flatbed scanner comes in various models like right-angle, prism and planetary/overhead to handle bound volumes and books. Flatbed scanner can scan usually a document at 600 dpi. Many flatbed scanners however, offer higher resolution.

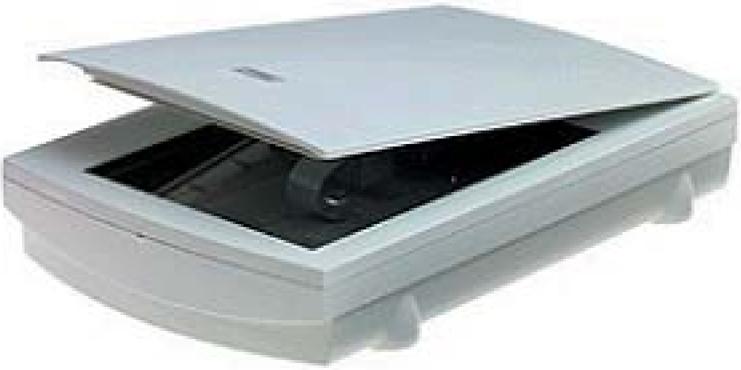


Fig. 7.9: Flatbed Scanner

Sheet feed Scanners: In a sheet-feed scanner, as is indicated in the name, document is fed over a stationary CCD and light source via roller, belt, drum, or vacuum transport. In contrast to a flat-bed scanner, sheet-feed scanner has optional attachment of auto- feed uniform-sized stack of documents to be scanned.

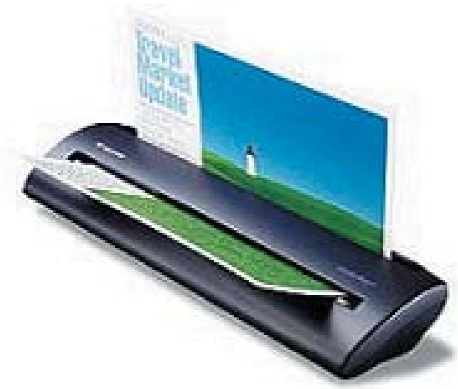


Fig. 7.10: Sheet feed Scanners

Drum Scanners: Source material in a drum scanner is wrapped on a drum, which is then rotated past a high-intensity light source to capture the image. Drum scanners offer superior image quality, but require flexible source material of limited size that can be wrapped around the photosensitive drum. Drum scanners are specially targeted for graphic arts market. Drum scanners offer highest resolution for grey scale and colour scanning. Drum Scanner uses Photo-Multiplier (Vacuum) Tubes (PMTs) instead of CCDs, which offer a greater bit depth (12 to 16 bits).

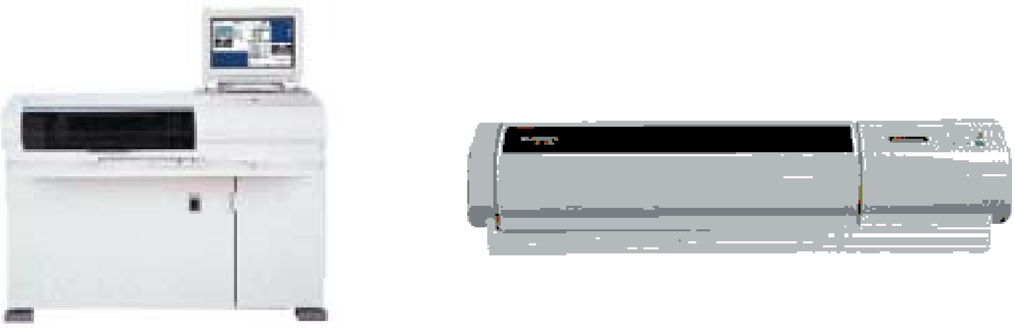


Fig. 7.11: Drum Scanners

Digital Cameras: Digital cameras mounted on copy cradle resemble microfilming stand. Source material is placed on the stand and the camera is cranked up or down in order to focus the material within the field of view. Digital cameras are most promising scanner development for library and archival applications.

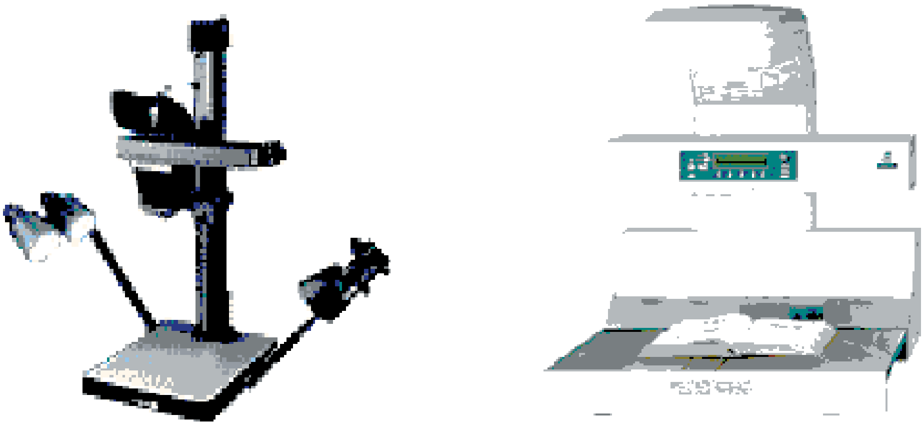


Fig. 7.12: Digital Camera

Slide Scanner: Slide scanners have a slot in the side to accommodate a 35mm slide. Inside the box, the light passes through the slide to hit a CCD array behind the slide. Slide scanners can generally scan only 35mm transparent source materials.

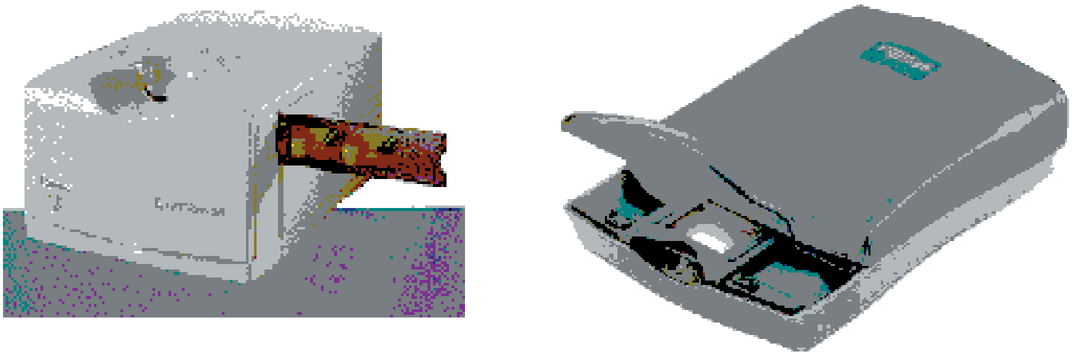


Fig. 7.13: Drum Scanner

Microfilm Scanner: Specially targeted to library/archival application, microfilm scanners have adapters to convert roll film, fiche, and aperture cards into the same model.

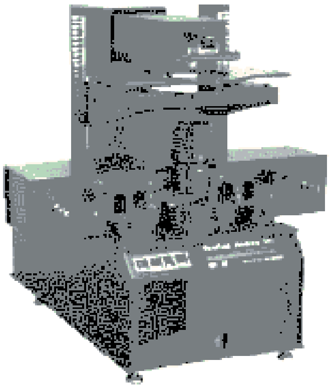


Fig. 7.14: Microfilm Scanner

Video Frame Grabber or VideoDigitiser: Video digitisers are circuit boards placed inside a computer and attached to a standard video camera. Any thing that is filmed by the video camera is digitised by the video digitiser.

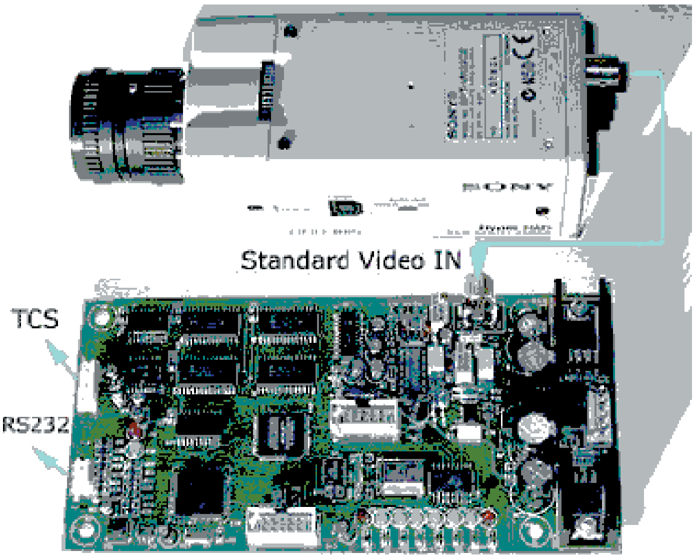


Fig. 7.15: Video Graber

Hand-held Scanners

Hand-held scanners are used for scanning selective sections of data. It may require multiple pass to capture large area. Moreover, a user should have a steady hand while moving the scanner over the document to be scanned. These scanners are normally used for circulation work in a library.

7.9.2 Scanning Software

The scanning software is used for scanning the image and capturing it in the computer. This software is provided by the manufacturer of the product to the buyers. These drivers translate the instructions into commands, which the scanner understands.

Image Editing Applications

Image editing applications are used once the process of scanning the image is over and the image is available in the computer for further manipulation. Most image editing software offer features like image editing, sharpening, filter, cropping, colour adjustments, forms conversion, resising, etc. Most image editing software can also be used for capturing the images.

7.10 DIGITISATION OF AUDIO AND VIDEO

The songs or speeches that we generally listen from tape recorders or radio are in an analogue form. The analogue sound tracks can be digitized by attaching an audio player to a system through an audio capture card so as to record the sound to the system. The audio files can be saved as wav, mp3, midi, etc. MP3 format is highly compact and the sound quality is better in comparison to other formats. Audio files can be further processed using noise reduction software.

Like audio, video capture also requires a video capture card with input from video cassette player (VCP / VCR), TV antenna, cable or movie camera. The digitised files can be saved as mov, avi, mpg file formats.

7.11 ORGANISING DIGITAL IMAGES

A disc full of digital images without any organisation, browse and search options may have no meaning except for one who created it. Scanned images need to be organised in order to be useful. Moreover, images need to be linked to the associated metadata to facilitate their browsing and searching. The following three steps describe the process of organising the digital images:

Organise the scanned image files into disc hierarchy that logically maps the physical organisation of the document. For example, in a project on scanning of journals, create a folder for each journal, which, in turn, may have folder for each volume scanned. Each volume, in turn, may have a subfolder for each issue. The folder for each issue, in turn, may contain scanned articles that appeared in the issue along with a content page, composed in HTML providing links to articles in that issue.

Name the scanned image files in a strictly controlled manner that reflects their logical relationship. For example, each article may be named after the surname of first author followed by a volume number and an issue number. For example, file name "smithrkv5n1.pdf" conveys that the article is by "R.K. Smith" that appeared in volume 5 and issue no.1. The file name for each article would, therefore, convey a logical and hierachial organisation of the journal.

Describe the scanned images file internally using image header and externally using linked descriptive metadata files. The following three types of metadata are associated with the digital objects:

i) **Descriptive Metadata:** Include content or bibliographic description consisting of keywords and subject descriptors.

ii) **Administrative or technical Metadata:** Incorporates details on original source, date of creation, version of digital object, file format used, compression technology used, object relationship, etc. Administrative data may reside within or outside the digital object and is required for long-term collection management to ensure longevity of digital collection.

iii) **Structural Metadata:** Elements within digital objects facilitate navigation, e.g., able of contents, index at issue level or volume level, page turning in an electronic book, etc.

The simplest and least effective method for providing access is through a table of contents and linking each item to its respective object / image. Content pages of issues of journals done in HTML would offer browsing facility. Full-text search to HTML pages or OCRed pages can be achieved by installing one of the free Internet search engines like Oingo Free Search (http://www.oingo.com/oingo_free_search/products.html); Swish-E (http://www.berkeley.edu/SWISH-E/);WhatyoUseek (http://intra.whatuseek.com/); Excite (http://excite.com/) and Google (http://www.google.com).

Large scanning projects would, however, require a back-end database storing images or links to the images and metadata (descriptive / administrative). Back-end database used by most document management systems holds the functionality required by most web applications. Important management systems like File Net have now integrated their database with HTML conversion tools. Further, some of the document management systems have also signed up with Adobe to incorporate Acrobat and Acrobat Capture into their web-based document management systems. These databases entertain queries from users through "HTML forms" and generate search results on the fly. Several digital library packages are now available as "open source" or "free-ware" that can be used not only for organzing the digital objects but also for their search and retrieval.

7.12 DIGITAL LIBRARY SOFTWARES

Several digital library softwares are currently available like, Greenstone Digital Library (GSDL), Dspace, Eprints, Fedora, etc., which are available freely for download on the Internet.There are some commercial Digital Library Softwares available but none has been used on large scale in comparision to the ones mentioned above. We will provide here a brief account of digital library softwares available in public domain.The description is an outline of the features of the respective softwares.

Dspace

Dspace (www.dspace.org) has been developed in partnership between Hewlett Packard (HP) and Massachusetts Instiute of Technology (MIT). Development is still in progress but as an instituional repository software. Dspace is making its mark, with an increasing number of institutions around the globe installing, evaluating, and using the package.The latest stable version is 1.2 available for download at the Dspace web site.

Currently, the original developers undertake most of the core development, but a growing technical user base is generating suggestions for future releases as well as looking for producing some add-on modules. In addition the Dspace Federation is guiding the transition of this software to a more community-wide open-source development model.

Dspace captures, stores, indexes, preserves, and, redistributes the intellectual output of a university's research faculty into digital formats. Dspace accepts all forms of digital materials including text, images, video, and audio files. Possible content includes: articles and preprints; technical reports; working papers; conference papers; e- theses; datasets(statistical, geospatial, matlab, etc.); images (visual, scientific, etc.); audio files; video files; learnning objects; and reformatted digital library collections. The back end technologies used include: Apache, Tomcat, OpenSSL/mod_ssl; Java1.3, JSP1.2, Servlet 2.3; PostgreSQL7, JDBC (RDBMS); CNRI handle System5 (persisten ids); Lucene 1.2 (index/ search).

Prerequisites

Dspace depends upon the Java programming language and the PostgreSQL open source database system. It also requires a number of additional Java- based elements to be installed: Tomcat, which is a Java based server; a number of Jave code libraries; and the Ant, a Java compiler. It is recommended that Dspace be installed on a Linux or a Unix machine.It requires an experienced system administrator to do the prerequisite installation.

E-prints

GNU Eprints 2.x is a free software which creates online archives (*http://software.eprints.org/*) . The default configuration creates a research paper archive.With its origin in the scholarly communication movement, e-print default configuration is geared to research papers but it can be adapted to other purposes and content. It was developed in the Intelligent Agents, Multimedia Group at the Electronics and Computer Science Departmen of the University of Southampton.

GNU Eprints is freely distributed to the GNU General Public License. The latest version is 2.3 and is available for download at http://software.eprints.org/download.php

Prerequisites

- Any computer capable of running GNU/Linux or similar operating system. The faster, the better, but any Intel Pentium II processor will give good performance.
- A GNU operating system. GNU/Linux (a very advanced and free UNIX- like operating system) works just fine, and is in fact the development platform
- Apache WWW server
- Perl programming language, also a number of additional modules
- mod_perl module for Apache, which significantly increases the performance of Perl scripts
- My SQL Databes

Greenstone Digital Library (http://greenstone.org)

Greenstone Digital Library is an open-source software available under the terms of the GNU General Public License. It has the ability to serve digital library collections and build new collections. It provides a new way of organising information and publishing it on the Internet or on CD-ROM. The Greenstone Digital Library software is produced by the New Zealand Digital Library Project at the University of Waikato, and distributed in cooperation with UNESCO and the Humanities Library Project. The New Zealand Digital Library Web site (http://nzdl.org) contains numerous example collections, all created with the Greenstone software, which are publicly available for anyone to peruse The Greenstone runs on Windows and Unix platforms. The distribution includes ready-to-use binaries for all versions of Windows and for Linux. It also includes complete source code for the system, which can be compiled using Microsoft C++ or gcc. Greenstone works with associated software that is also freely available: the Apache Web server and PERL.

Ganesha Digital Library

Ganesha Digital Library version 3.1 (GDL) (http://gdl.itb.ac.id/) is another open source software developed under Indonesian Digital Library Network (IndonesiaDLN). Ganesha Digital Library enables institutions or individuals to share their knowledge and also access and utilise available knowledge in the Indonesian 'giant memory' through the network of IndonesiaDLN digital libraries. The software is available in three publisher editions: Personal, Internet Cafe, and Institution. Released under the terms of the GNU GPL, (GNU General Public License).

7.13 PLANNING AND IMPLEMENTATION

Digitisation is the first step towards building a digital library. It is highly specialised and cost-intensive activity that requires inputs from diverse branches of knowledge. It is important that objectives, needs and the purpose of digitisation are established clearly. The digitisation proposal should therefore define its goals (objectives) scope, feasiblity, benefits, costs, time required for the developmental phase, implementation issues, deliverables and target users. It may be desirable to continue with traditional collections and also acquire collections in digital media; (ii) buy access to electronic resources; and (iii) develop subject gateways or library portals, instead of undertaking digitisation project. This approach would save the cost and efforts on digitisation and other recurring administrative costs. However, once a decision for digitisation is taken, due importance should be given to factors such as sustenance, reusability, interoperability, verification and documentation both for users as well as for the developer. The steps given below may be considered as pre-requisites. Careful planning of digitisation would bring activity to the project, save cost, time and human efforts. The Planning would include the following steps :

7.13.1 Feasibility

First, it is necessary to conduct a feasiblity study of the digitisation project. The feasiblity should be established not only in terms of the availability of tools, and expertise, but also the factors like volume/number of documents to be covered in the process of digitisation, target audience, demand for material to be digitised and user's requirements. The study should also assess whether the library can take-up the project in-house or should it be out-sourced.

7.13.2 Planning the Project

The planning of the project needs to cover the following areas :

Managerial Planning

Managerial planning would essentially involve the process of sequencing various taks, their time management and project monitoring. Activities that need managerial planning may include conducting feasiblity study, procurement of equipment, recruitment of manpower, digitisation (whether out-sourced or done-in-house), IPR and rights management issues, integration and organisation of content, finding market, launching and marketing of services. Flow diagrams, PERT, CPM and SWOT analysis and other management techniques may be deployed at this stage.

Hardware and Software Planning

In this, the requirements of hardware and software for the servers and network componenets may be worked out with their financial implications and network components. Connectivity and the bandwidth required for hosting the digitised collection may also be planned. The technical specifications should be worked out much before the actual process of digitisation commences irrespective of whether digitisation is done in-house or out-sourced. For this, the existing types, formats, standards and practices are to be reviewed first. Draft specifications are to be prepared and these are to be tested with sample data. Necessary modifications may be made in the specifications based on this testing laid down for metadata creation for digital objects as well as for the digital collection. Digital objects and digital collections typically require descriptive (keywords / descriptors), structural (navigation, content pages etc.) and administrative (formats, compression, standards, etc.) metadata.

Human Resources Planning

Human resources needed to be worked out in terms of staff time involved, training of existing staff and recruitment of new staff with desired skills. Human resource planning would depend on whether the library is going for in-house digitisation or for outsourcing the process of digitisation.Project management continues to be an important issue even if the digitisation work is outsourced. The management of the project may be divided in groups with responsiblities defined.Communication betweeen the groups and a reporting structure may be laid down to facilitate unambiguous communication among the groups and the staff.

Financial planning

Financial planning is cruicial. Cost of migration from one medium to another and from one computer to another may be built-in. Cost of hosting the services and their maintenance should also be planned besides other aspects mentioned above.

7.13.3 Purchase of Hardware and Software

Choice of technology and the equipment required may be made. These include storage and back up devices, network equipment regained, software for search and access and other related items. The software may be acquired or developed inhouse. The following steps may be followed in this regard:

i) Acquire and install hardware and software;

ii) Acquire and install the network required for hosting the digitised collection. Consider bandwidth requirements that depend upon the media offered by the digital library. While simple text requires relatively low bandwidth to deliver content, images and video require large bandwidth; and

iii) Acquire and install other components

7.13.4 Selection of Material for Digitisation and 'Born Digital'

In the process of execution of the project, the first task is to identify, select, and to prioritise the documents that are to be digitised.If the organisation is itself creating contents, strategies are to be laid down to capture 'born digital' data. If documents are available in digital form, they can be easily converted to other formats.If the selected material is from external sources, IPR issues need to be resolved.It is necessary to obtain permission from the publishers and data suppliers for digitisation, if material being digitised is not available in public domain.Moreover, decision may be taken whether to OCR the digitised images. Documents selected for digitisation may already be available in digital format. It is always economical to buy e-media, if available than their conversion. Moreover, oversized material, deteriorating collections, bound volumes of journals, manuscripts, etc., would require highly specialized equipment and highly specialised manpower.

7.13.5 Placement and Training of Manpower

Since the entire job of developing and or maintaining a digital library is a highly skilled one, there should be no compormise or slackening in the quality of intake or selection of manpower for the job. Also, even if good quality manpower is positioned, they usually need training to upgrade and sharpen their skills for this job. So, necessary training, should form a component of the execution of the project.

7.13.6 Content Creation

The steps involved in content creation include the following :-

i) Conversion of datasets that are 'born digital', for example, convert MS Word file into PDF;

ii) Conversion of the existing printed sections into digital format (digitisation); and

iii) Identification of vendors if the digitisation work is to be outsourced.

7.13.7 Execution of the Project

Once the equipment and software and other infrastructure facilties are installed or positioned, and the priorities of the documents for digitisation laid down, the execution of the project starts. The library may use digital library software like greenstone Digital library, or Dspace, etc.

Self Check Exercise

11) How do you digitise audio and video? What are the devices used for it?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of the Unit.

....

....

....

....

....

7.14 SUMMARY

Digitisation is the process of converting the content of physical media (e.g., periodical articles, books, manuscripts, cards, photographs, vinyl disks, etc.) into digital format. In most libraries, digitisation normally results in making documents accessible from the web sites of the libraries. Optical scanners and digital cameras are used to digitise images by translating them into bit maps. It is also possible to digitise sound, video, graphics, animations, etc.

Digitisation is the first step in the process of building digital libraries. Digitisation is also used for achieving preservation and archiving although it is not considered a good option for this purpose. It is highly labour-intensive and cost-intensive process that involves several complexities including copyright and IPR issues. However, digital objects offer numerous benefits in terms of accessibility and search. The documents to be digitised may include text, line art, photographs, colour images, etc. The selection of document needs to be made carefully considering all the factors of utility, quality, security and cost. Rare and much-in-demand documents and images are selected as first priority even though the quality is not good.

The process of digitisation involves four steps, namely, scanning, indexing, storage and retrieval. A scanned document is nothing more than a picture of a printed page. It cannot be edited or manipulated or managed based on their contents. In other words, scanned documents have to be referred by their labels rather than characters in the documents. OCR (Optical Character Recognition) programs are software tools used to transform scanned textual page images into word processing files. OCR or text recognition is the process of electronically identifying the text in a bit-mapped page image or set of images and generate a file containing text in ASCII code or in a specified word processing format leaving the image intact in the process.

The quality of digital image can be monitored at the time of capture by four factors, namely: i) bit depth / dynamic range; ii) resolution; iii) threshold; and iv) image enhancement. The Unit describes these parameters in detail. Image files are evidently larger than textual files. It is thus necessary to compress image files. Image compression is the process of reducing size of an image by abbreviating the repetitive information such as one or more rows of white bits to a single code. The compression algorithms may be grouped as lossless compression and lossy compression. The Unit also describes compression technology and protocols.

Text and image-based contents of a digital library can be stored and presented as: i) simple text or ASCII (American Standard Code for Information Interchange; ii) unstructured text; iii) structured text (SGML or HTML or XML); iv) page description language and v) page image formats. The Unit also describes file formats in digitised collection.

An image scanning system would normally consist of a stand-alone workstation where most or all the work is done on the same workstation or as a part of a network of workstations with imaging work being distributed and shared amongst various workstations. The network usually includes a scanning station, a server and one or more editing and retrieval stations. A typical scanning workstation for a small production-level project, could consist of the following:

- Hardware (scanners, computers, data storage and data output peripherals)
- Software (image capturing and image editing)
- Network (data transmission)
- Display and Printing technologies

This Unit describes scanners and scanning software as important components of the scanning system.

7.15 ANSWERS TO SELF CHECK EXERCISES

1) Digitisation is the process of converting the content of physical media (e.g., printed materials, images, photographs, microforms, etc.) into digital format. It refers to the process of translating a piece of information such as a book, journal articles, sound recordings, pictures, audio tapes or videos recordings, etc. into bits. Converting information into these binary digits is called digitisation. Digitising a document in print or other physical media (e.g., sound recordings) makes the document more useful as well as more accessible. It is possible for a user to conduct a full-text search on a document that is digitised and OCRed. It is possible to create hyperlinks to lead a reader to related items within the text itself as well as to external resources. Ultimately, digitisation does not mean replacing the traditional library collections and services; rather, it serves to enhance them.

2) The process of selection of material for digitisation invloves identification, selection and prioritisation of documents that are to be digitised. The selection of document need to be reviewed very carefully considering all the factors of utility, quality, security and cost. Rare and much in demand documents and images are selected as first priority without considering the quality. If the selected material is from the external sources, IPR issues need to be resolved. If material being digitised is not available in

public-domain then it is important to obtain permission from the publishers and data suppliers for digitisation. Moreover, decision may be taken whether to OCR the digitised images. Documents selected for digitisation may already be available in digital format. It is always economical to buy e-media, if available, than their conversion. Moreover, over-sized material, deteriorating collections, bound volumes of journals, manuscripts, etc. require highly specialized equipment and highly skilled manpower. The documents to be digitised may include text, line art, photographs, colour images, etc.

3) Steps in the Process of Scanning using a Flatbed Scanner are as follows:

Step 1. Place picture on the scanner's glass

Step 2. Start scanner software

Step 3. Select the area to be scanned

Step 4. Choose the image type

Step 5. Sharpen the image

Step 6. Set the image size

Step 7. Save the scanned image using a desirable format (GIF or JPEG)

4) Bitonal or black & white or binary scanning is used in libraries to scan pages containing text or the drawings. Bitonal or binary scanning represents one bit per pixel (either "0" (black) or "1" (white). Gray scale scanning, on the contrary, is used for reliable reproduction of intermediate or continuous tones found in black & white photographs to represent shades of grey. Multiple numbers of bits ranging from 2-8 are used to for each pixel to represent shades of grey. Although each bit is either black or white, as in the case of bitonal images, but bits are combined to produce a level of grey in the pixel that is, black, white or somewhere in between. Since only one bit (either black or white) is used to represent a pixel in bitonal scanning, bit depth is not important in bitonal scanning.

5) The resolution of an image is defined in terms of number of pixel (picture elements) in a given area. It is measured in terms of dot per inch (dpi) in case of an image file and as ratio of number of pixel on horizontal line x Number of pixel in vertical lines in case of display resolution on a monitor. Higher the dpi is set on the scanner, the better the resolution and quality of image and larger the image file. (printed resolution versus screen resolution)

6) Image compression is the process of reducing size of an image by abbreviating the repetitive information such as one or more rows of white bits to a single code. The compression algorithms may be grouped into the following two categories:

Lossless Compression: The conversion process converts repeated information as a mathematical algorithm that can be decompressed without loosing any details into the original image with absolute fidelity. No information is "lost" or "sacrificed" in the process of compression. Lossless compression is primarily used in bitonal images.

Lossy Compression: Lossy compression discards or minimizes details that are least significant or which may not make appreciable effect on the quality of image. This kind of compression is called "lossy" because when the image that is compressed using "Lossy" compression techniques is decompressed, it will not be an exact replica of the original image. Lossy compression is used with grey-scale/ colour scanning.

7) LZW compression technique uses a table-based lookup algorithm invented by Abraham Lempel, Jacob Ziv, and Terry Welch. A particular LZW compression algorithm takes each input sequence of binary digit of a given length (for example, 12 bits) and creates an entry in a table (sometimes called a "dictionary" or "codebook") for that particular bit pattern, consisting of the pattern itself and a

shorter code. As input is read, any pattern that has been read before the results in the substitution of the shorter code effectively compresses the total amount of input to something smaller. The decoding program that uncompresses the file is able to build the table itself by using the algorithm as it processes the encoded input. LZW compression is suitable for compressing both graphical as well as text files. Two commonly-used file formats in which LZW compression is used are the Graphics Interchange Format (GIF) and Tag Image File Format (TIFF).

8) OCR (Optical Character Recognition) or text recognition is the process of electronically identifying text in a bit-mapped page image or set of images and generate a file containing that text in ASCII code or in a specified word processing format leaving the image intact in the process. OCR programs are software tools that are used to transform scanned textual page images into word processing file. The process of OCR is used in order to make every word in a scanned document readable and fully searchable without having to key-in everything in the computer manually. Once a bit-mapped page image has gone through the process of OCR, a document can be manipulated and managed by its contents, i.e. using the words available in the text. OCR does not actually convert an image into text but rather creates a separate file containing the text while leaving the image intact.

9) A defined arrangement for discrete sets of data that allow a computer and software to interpret the data is called a file format. Different file formats are used to store different media types like text, images, graphics, pictures, musical works, computer programs, databases, models and designs video programs and compound works combining many type of information. Every file in electronic format needs a name or identifier that distinctly identifies its type and format. The file extensions are used to denote formats, protocols and right management that are appropriate for the type of material. For example file with extension .doc can be recognized as a Microsoft word file and a file with an extension .gif can be recognized as an image file stored in Graphical Interface Format (GIF).

Simple text or ASCII is a compact and economic file format that is used to capture and store textual data. However, the simple text or ASCII can neither be used for displaying complex tables or mathematical formulas nor can it be used to display photographs, diagrams, graphics, and special characters. Text or ASCII format does not store text formatting information, i.e., italics, bold, font type, font size or paragraph justification information. Simple text or ASCII is, therefore, inadequate to represent journal articles because of these reasons. Structured text, on the other hand, attempts to capture the essence of documents by "marking-up" the text so that the original form can be recreated or even produce other forms such as ASCII. Structured text format have provision display italics, bold, font size, font type and paragraph justification. It is capable of imbedding images, graphics and other multimedia formats in the text. SGML (Standard Generalized Markup Language), HTML, .DOC are some of the important and popular structured text format. Like simple text or ASCII, structured text can be searched or manipulated. It is highly flexible and suitable both for electronic and paper production. Well-formated text increase visual presentation of textual, graphical and pictorial value of information. Structured formats can easily display complex tables and equations. Moreover, the structured text is compact in comparison to the image-based formats, even after including imbedded graphics and pictures.

10) Page description Language (PDLs), such as Adobe's PostScript and PDF (Portable Document Format) is a page description format that essentially has all the information required to display or print that page. PDLs retain all the information about what that page is supposed to look like and how it should be printed. In addition to fonts, PDF represents all the other visual aspects of the page including line breaks, layout, white space, graphics, colours; i.e. every visual feature of the page. A PDF is essentially similar to image but the formatted pages displayed to the user are text-based rather than image-based. PostScript and PDF formats can easily be captured

during the typesetting process. A PDL can be generated from almost all structured formats using various software packages although reverse process is not possible. Moreover, a structured text file can be edited and manipulated, a PDL file can not be edited or manipulated directly.

11) An audio file that we generally listen from tape recorders or radio is in an analogue form. The analogue sound tracks can be digitised by attaching an audio player to a system through an audio capture card so as to record the sound to the system. The audio files can be saved as .wav, mp3, midi, etc. MP3 format is highly compact and the sound quality is better in comparison to other formats. Audio files can be further processed using noise reduction software.

Like audio, video capture also requires a video capture card with input from video cassette player (VCP / VCR), TV antenna, and cable or movie camera. The digitised files can be saved as .mov, .avi, and .mpg file formats.

7.16 KEYWORDS

ASCI : American Standard Code for Information Interchange or ASCI is a standard coding technique for representing computer information.

Analogue : A term used to describe a signal, such as the human voice and electric current, whose value varies continuously with time or transmission method, such as the traditional telephone network, which carries source signals as electrical waves. Compared with digital systems, an analogue telephone line carries data at low speed; it also requires a modem to convert the computer's digital output into a form (analogue) which it can handle.

Bit and Byte : Bit is short term for binary digit, the smallest unit of information on a machine. A single bit can hold only one of two values: 0 or 1. More meaningful information is obtained by combining consecutive bits into larger units. A byte is composed of 8 consecutive bits.

Bit Depth : The number of bits used to represent each pixel. The greater the bit depth, the more colours or grey-scales can be represented. For example, a 24-bit colour scanner can represent 2 to the 24th power (16.7 million) colours.

Bitonal : Each pixel in a bitonal image is represented by a single bit, i.e. black and white. Textual documents and line drawings are scanned in bitonal, i.e. a pixel is either black or white whereas each pixel for a picture (black and white or colour) may contain 2-8 bit per primary colour.

Capture : A term used in document imaging for scanning of a document or any other artefact.

Crop : The process of elimination of a portion of a picture, illustration or photograph that contain unnecessary material or to highlight a certain area of the image.

Digitisation : The process of converting data and information into digital format is called digitisation. It is synonymous with scanning, it is the conversion from printed paper, film, or some other media, to an electronic form where the page is represented as either black and white dots, or color or grayscale pixels.

Document Scanning : Document scanning is the process by which print and film documents are fed into a scanner and converted into electronic documents. During the scanning process documents can be OCRed and indexed to insure quick retrieval at a later date.

Dots Per Inch (DPI) : Dots per inch (dpi) indicates the resolution of images or printers. The more dots per inch, the higher the resolution

Dynamic Range : The number of colours or shades of grey that can be represented by a pixel. Dynamic range is a measurement of the number of bits used to represent each pixel in a digital image. 1-bit or bitonal means that a pixel can either be black or white. Bitonal imaging is good for black and white images, such as line drawings and text. However, scanning in grey-scale rather than bitonal may produce a better looking image. 8-bit color or 8-bit grey-scale means that each pixel can be one of 256 shades of colour or one of 256 shades of grey. 24-bit colour means that each pixel can be one of 16.8 million colours.

Halftone : A method of generating an image that requires varying densities or shades to accurately render the image. This is achieved by representing the image as a pattern of dots of varying size. Larger dots represent darker areas, and smaller dots represent lighter areas of an image.

Image Sharpening : Scanned images can be adjusted to increase edge contrast and artificially enhance the overall quality of image. Most paint and colour manipulation programs have special tools to selectively sharpen isolated areas of an image.

Metadata : Data about data, or information known about the image in order to provide access to the image. Usually includes information about the intellectual content of the image, digital representation data, and security or rights management information.

Optical Character Recognization (OCR) : The OCR refers to the process of scanning text using a scanning device from a printed page into an image and translate it into a computer processible format , i.e. an ASCII file. OCR systems include an optical scanner for reading text and sophisticated software for analyzing images.

Pixels : The term "pix" means "part of a picture" and "el" means "from element". In bitonal (black and white) display, each pixel can have only one bit, i.e. either black or white, whereas in a gray-scale display, each pixel can have three numerical values for three colour, i.e. Red, Green, Blue (RGB) to represent the colour. These three RGB components can be represented by three 8-bit numbers for each pixel. Three 8-bit bytes (one byte for each of RGB) is called 24 bit colour. Each 8 bit RGB component can have 256 possible values, ranging from 0 to 255. For example, three values like (250, 165, 0), meaning (Red=250, Green=165, Blue=0) to denote one Orange pixel. The pixels are most commonly used to represent images as a computer file. Pixel are of a uniform size and shape.

Portable Document Format (PDF) : Portable Document Format is a type of formatting that enables files to be viewed on a variety computers regardless of the program originally used to create them. PDF files retain the "look and feel" of the original document with special formatting, graphics, and colour intact. A special program or print driver (Adobe Distiller or PDF Writer) is used to convert a file into PDF format. The Acrobat Reader program available free from the Adobe site.

PostScript : A page description language developed and marketed by Adobe Systems. PostScript can be used by a wide variety of computers and printers, and is the dominant format used for desktop publishing. Documents in PostScript format are able to use the full resolution of any PostScript printer, because they describe the page to be printed in terms of primitive shapes which are interpreted by the printer's own controller. PostScript is often used to share documents on the Internet because of this ability to work on many different platforms and printers.

Primary Colour : Colours that are basis for all other colour combinations. Primary colours are red, green and blue (RGB).

RGB : Red, Green and Blue (RGB) are the colours that are basis for all other colour combinations. They are called primary colours.

Resolution : Resolution refers to the number of pixels contained on a display monitor, printers and bit-mapped graphic images. In the case of printers and images, it is indicated the number of dots per inch. For graphics monitors, the screen resolution signifies the number of dots (pixels) on the entire screen. Resolution generally refers to the sharpness and clarity of an image.

Scanner : An optical input device that uses light–sensing equipment to capture an image on paper or some other subject. The image is translated into digital signals that can then be manipulated by optical character recognition (OCR) software or graphics software. Scanners come in a number of types, including flatbed (scan head passes over a stationary subject), feed (subject is pulled across a stationary scan head), drum (subject is rotated around a stationary scan head), and hand-held (user passes device over a stationary subject).

TIFF : Tag Image File Format (TIFF) is a common format for exchanging raster (bitmapped) images between application programs. Usually identified with the ".tiff" or ".tif" filename extension, the format was developed in 1986 by an industry committee chaired by the Aldus Corporation (now part of Adobe). One of the more common image formats, TIFFs are common in desktop publishing, faxing, and medical imaging applications.

Threshold : The minimum level at which a signal of any kind can be detected, either by the human senses or by using any electronic instrumentation. In image processing, threshold is a specified grey level used for producing a binary image.

7.17 REFERENCES AND FURTHER READING

Arms, William Y. (2000). *Digital Libraries*. The MIT Press: Cambridge, MA.

Arms, W.Y. (1995). Key Concepts in the Architecture of the Digital Library. *D-lib Magazine*.

Haigh, S. (1996). Optical Character Recognition (OCR) as a Digitization Technology. *Network Notes*. no. 37.

IMLS: A Framework of Guidance for Building Good Digital Collections.

(http://www.imls.gov/pubs/forumframework.htm)

Jantz, Ronald. (2001). Technological Discontinuities in the Library: Digital Projects That Illustrate New Opportunities for the Librarian and the Library. *IFLA Journal* 27, 74-77.

Kenney, Anne R. and Stephen Chapman. (1996). *Digital Imaging forLibraries and Archives.* Ithaca: Dept. of Preservation and Conservation, Cornell University Library.

Kessler, Jack. (1996). *Internet Digital Libraries: The International Dimension*. Boston: Artech House Publishers.

Lesk, Michael. (1997). *Practical Digital Libraries: Books, Bytes and Bucks*. San Fransisco: Morgan Kaufmann Publishers.

NorthEast Document Conservation Center. *NEDCC Handbook for Digital Projects: A Management Tool for Preservation and Access.*

Noerr, Peter (2000). *Digital Library Tool Kit.* U.S.A.: Sun Microsystems.

(http://www.sun.com/products-n-solutions/edu/libraries/digitaltoolkit.html)

Ostrow, Stephen. Digitizing Historical Pictorial Collections for the Internet. *CLIR* Feburary 1998. (http://www.clir.org/pubs/reports/ostrow/pub71.html)

Rosenfeld, Louis and Morville, Peter(1998). *Information Architecture*. Cambridge: O'Reilly,

Tyson, Jeff. (2003). *How Scanners Work: How Stuff Works.* (http://www.howstuffworks.com)

Townsend, Sean, [et. al.]. *Digitising History.* (http://hds.essex.ac.uk/g2gp/digitising_history/index.html)