
UNIT 6 DIGITISATION PROCESS

Structure

- 6.0 Objectives
- 6.1 Introduction
- 6.2 Digitisation of Print Based Documents
 - 6.2.1 Capturing Print Based Document
 - 6.2.2 Digitising
- 6.3 Video Digitisation
 - 6.3.1 Video Capturing
 - 6.3.2 Video Digitisation Process
- 6.4 Audio Digitisation
 - 6.4.1 Audio Capturing
- 6.5 Audio/Video Compression
- 6.6 Audio/Video Streaming
- 6.7 File Formats and Content Creation
- 6.8 Summary
- 6.9 Answers to Self Check Exercises
- 6.10 Keywords
- 6.11 References and Further Reading

6.0 OBJECTIVES

After going through this Unit, you will be able to:

- Understand the digitisation process of text, audio and video;
- Know different types of file formats; and
- Explain the file compression process.

6.1 INTRODUCTION

A digital library may contain materials that are born digital, such as e-journals and e-books, or may contain materials that were originally produced in another form but subsequently digitised. The process of digitising materials involves different steps depending upon material, technology and requirement. Various technical issues, like hardware and software, file formats and file compression and then the post processing requirements for making the digitised file accessible to end-user will be discussed.

6.2 DIGITISATION OF PRINT BASED DOCUMENTS

Once you have taken decision as to what needs to be digitised, the first step is to capture the documents available in print or analogue form for conversion into digital form. In the case of print based material, it is the hard copy of the document which needs to be scanned and digitised. The hard copy can be a paper based document, microforms or projection slides. For audio/ video media conversion is done from the analogue form to digital formats. Capturing devices for print based material include scanners and digital cameras attached with a computer. For audio/ video material

The steps for scanning a document

Step 1: Place the document on the scanner bed

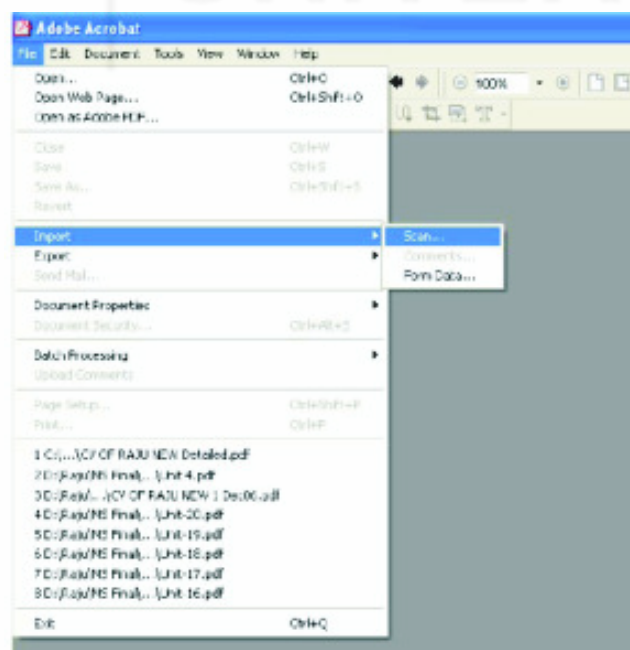
Here we show the process using the Konica Minolta PS 7000 book scanner, which is a superior system for scanning large-sized books, artwork, ledgers and other bound materials. It is a face-up scanning system.



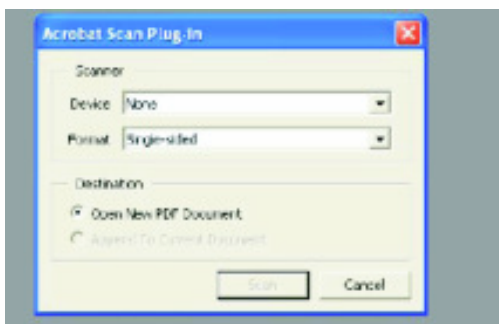
Fig. 6.1: Book Scanner

Step 2: Open the Adobe Acrobat

Click on **File>>Import>>Scan...**



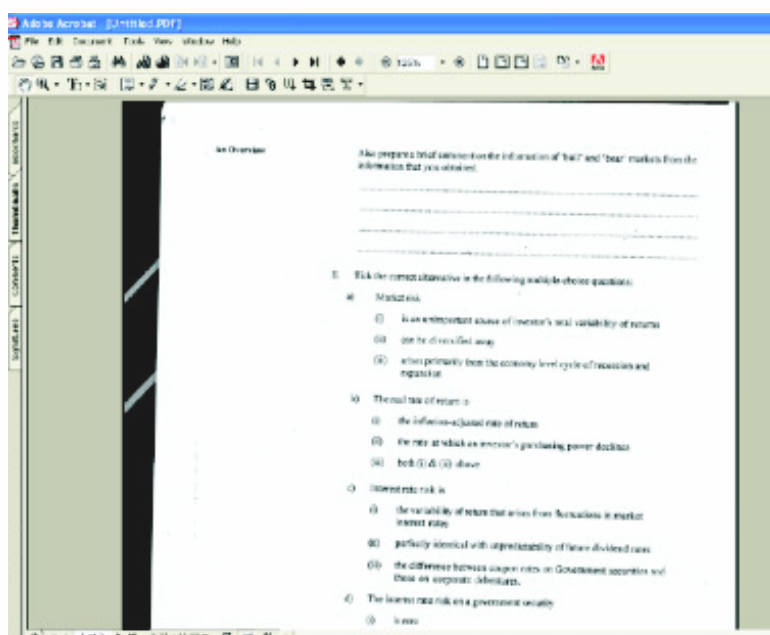
Fill in the information for device, format and destination in the dialogue box that appears



To scan the documents click on the **Scan All** option. From the **Minolta PS7000 Scanner Setup Dialog Box** that appears.



Click on **Done** option from the **Minolta PS7000 Scanner Setup Dialog Box** which shows the file like this:



Save the file as PDF version giving .pdf extension. To change the resolution, Click on Scan Setting >> Resolution (DPI) from the Minolta PS7000 Scanner Setup Dialog Box. To change the Scan Area click on, Scan Setting >> Scan Area. You can also change the Brightness and Contrast of the scanned file by using the drag button from the right panel. If you want to change the Image Type then click on Scan Setting >> Image Type. You can also change the Brightness and Contrast of the scanned file by using the drag button from the right panel. Scanned pages can be saved as individual files or as a complete document by appending them to the current document while scanning.

6.2.2 Digitising

The process of digitisation involves capturing the physical or analogue object through devices like scanners, digital camera, recorder etc., converting them into numerical values in bits and bytes which enables them to be read electronically.

Digitisation of text is possible either through text transcription or using optical character recognition method. Text transcription can be through keying in the text using a keyboard or by voice recognition software. Keyed in text are saved in ASCII format which do not replicate the structure and format of the original text.

OCR software converts image of text captured by a scanner into computer editable text which a word processor can read. The software tries to match the image of each letter against the pattern it recognizes making use of the stored knowledge about the shapes of individual characters. The OCR software has options for either storing the text and graphics in their original layout or converting them into ASCII or word processing format. Omnipage Pro and ABBYY Fine Reader are two commonly used OCR software.

After OCR, you can export the resulting text to a variety of word-processing, page layout, and spreadsheet applications. It also provides the option to save it directly as a PDF file.

Self Check Exercise



- Note:** i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of this Unit.
- 2) Name two commonly used OCR software.

.....
.....
.....
.....
.....

To perform OCR with automatic processing the following steps are to be followed:

- 1) Select all settings needed to process pages. Do this in the following:
 - Get Pages drop-down list
 - Layout Description drop-down list

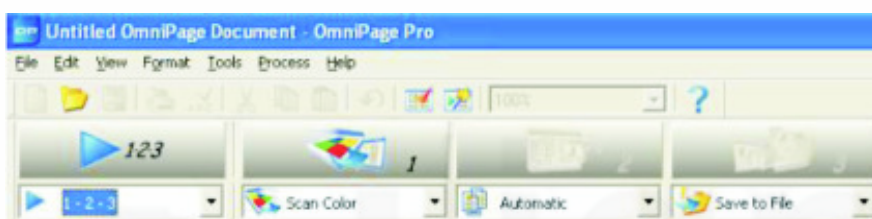
- Export Results drop-down list
- Options dialog box panels (Tools menu)

2) Click the button  or Click on the shortcut icon 

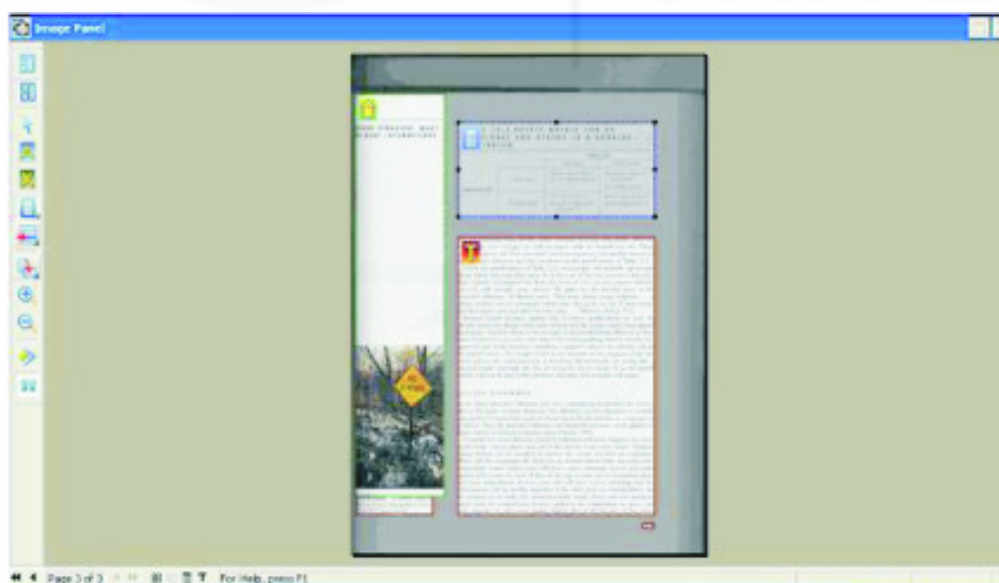
Start button with 1-2-3 selected in the Workflow drop-down list. Your pages will be acquired, auto-zoned and recognized one after the other. Proofing will start if you requested it. When proofing and/or recognition are finished, an export dialog box appears. Select the destination, file type and file name to save the file.

To manually perform the OCR, follow the steps given below.

- 1) Scan the document as an image
 - Launch Omni Page Pro. **Start>Programs>Scansoft OmniPage Pro**
 - The Program will open with the toolbar shown below.



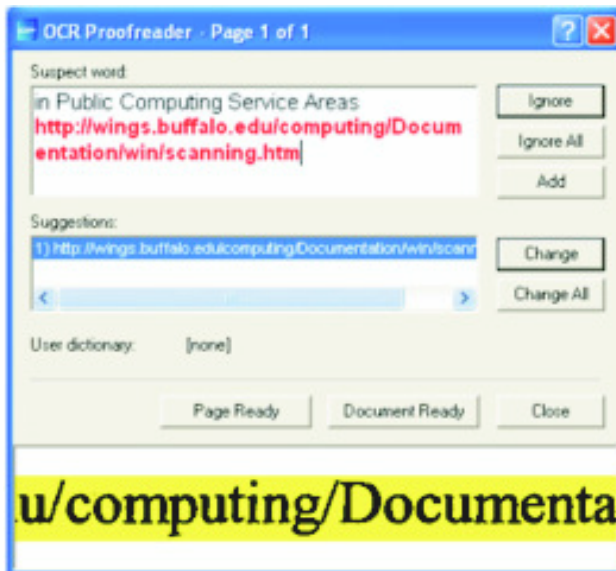
- Place the document to be scanned in the scanner.
 - Click the icon above the Scan Color menu. The Program will scan the document.
- 2) Select for Recognition
 - Once the document opens up in Omni Page, draw a box around the text you want.
 - You can categorize the objects on the scanned image into text, table or image by selecting the appropriate option on the side toolbar.



 → Image

 → Table

 → Text

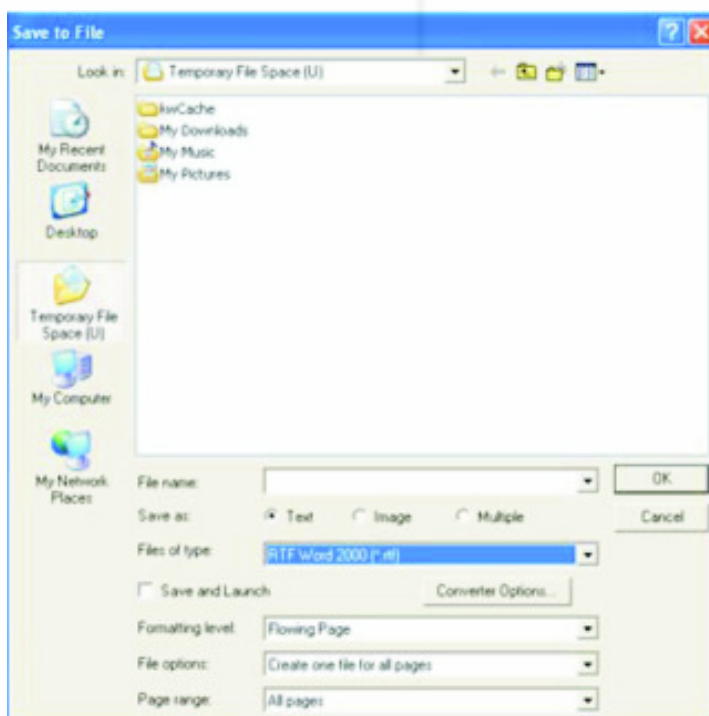


4) Select page layout

Once the proofreading is complete the document is exported to the text editor in **OmniPage**. Here you can edit the text and change the page layout.

5) Save as a file

- To do this click on the icon above the **Save to File menu**.
- Choose the location to save it at and give it a file name and select the file type to save it as. Now you can save the file in the available format you want. The typical formats available are MS-Word document *.doc, PDF *.pdf, HTML *.html, Text *.txt
- Enter your desired file name in the **File Name** text field.
- You can choose a document format from the Files of type pull-down menu. The default selection of RTF Word (*.rtf) is highly recommended, as it can be opened by most of the word processing programs.
- Click **OK** to save the file.



6.3 VIDEO DIGITISATION

Analogue mediums such as vinyl, VHS cassettes, and TVs have now been replaced by superior digital medium, such as CDs, DVDs, and HDTVs. The digital medium provides higher quality content. It also allows exact reproduction from copy to copy, barring any encryption technology implemented to stop copying.

Digital video refers to video being viewed or manipulated in the digital system (for instance on a computer), or sometimes simply video stored in a digital tape format. The video may have originally been analogue source material digitised into a computer, or it may have been stored directly to a digital tape format. Traditionally, digital tape formats were only available at the professional level (D-1, Digital Betacam, etc.), but now that some digital tape formats (DV) have emerged on the consumer scene, there is even more confusion about the generic term “digital video.”

DV (and related DVCAM and DVCPRO) is a digital tape format developed by a consortium of 10 companies as a “consumer” digital video format. There are now over 60 companies in the DVC consortium, including Sony, Panasonic, JVC, Philips, and other similar names you’ve heard before.

6.3.1 Video Capturing

In the simplest terms multimedia capturing can be stated as the process of storing or displaying the video/audio from the devices like Camcorders, Digital Cameras etc to some digital form like that of Monitor or in the binary forms (files).

As we have moved into the 21st Century, traditional analogue mediums such as vinyl, VHS cassettes, and TVs are being replaced by superior digital ones, such as CDs, DVDs, and HDTVs. Not only does digital formats allow for higher quality content, but also allows exact reproduction from copy to copy, barring any encryption technology implemented to stop copying. As computers become faster and disk storage space becomes larger, users are able to more deftly manipulate their digital data taken from analogue mediums and frequently “improve” the original analogue content using various techniques in the digital world.

System Requirements for a beginner multimedia processing system:

- x86-based PC @ 800+Mhz
- 256+MB RAM
- 40+GB of Free HD space (7200 rpm drive)
- Microsoft Windows98/ME/2000/XP
- Sound card with Line-in
- Video Capture card

These are the minimum requirements to perform reliable video capture. It is entirely possible to do video capture with less than this configuration, but good results cannot be guaranteed. Obviously, a faster CPU, more RAM, and more HD space are nothing but a good thing. Windows 9x/ME users should be aware that the FAT32 file system has a limitation preventing files from being larger than 4GB.

Windows machine is strongly recommended since the NTFS file system has no such file size limitation.

Choosing the Right Device to Capture the Video/Audio

One can purchase a video card with video-in support built right onto the card. We require the device which has a built-in “Analogue-to-Digital Conversion with Pass-Through” ability. This feature is quite useful since it will allow us to attach any analogue device (VCR, 8mm camcorder, etc.) to our Handy cam and then stream the digital data over FireWire to our computer.

6.3.2 Video Digitisation Process

Video digitisation is the next step used where the captured data from the analogue/digital device like cam coder is processed and saved in various file formats understandable by Media Players (both hardware and software based).

Software for video digitisation:

1) *VideoLAN*

VLC Player is one of the open source technologies that we are using to do the following things:

- Digitisation of content in various formats
- Re-Digitisation of multimedia video/audio content on LIVE and VOD.

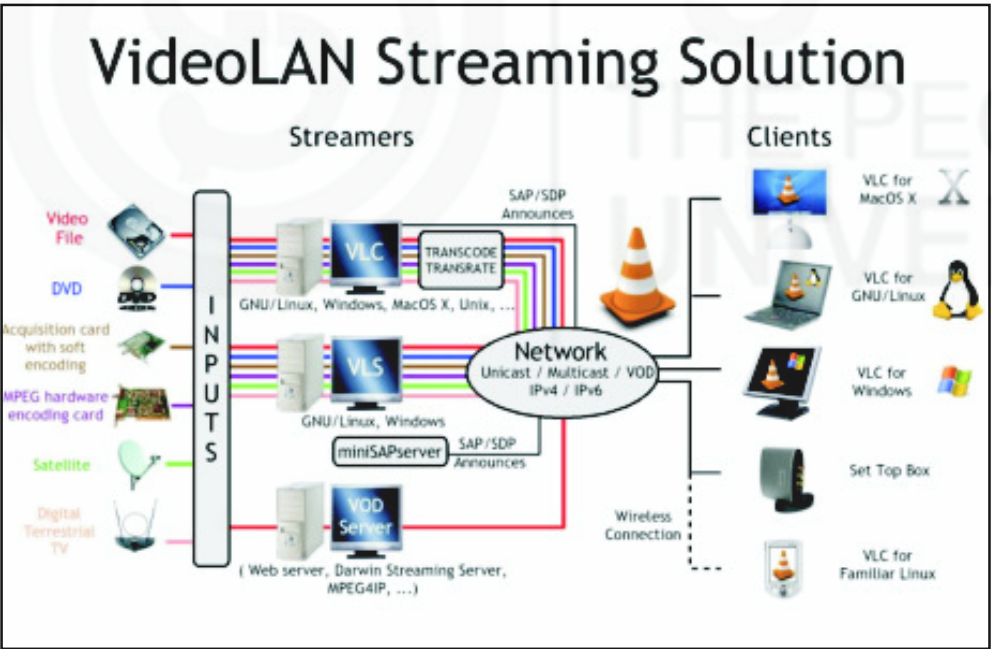
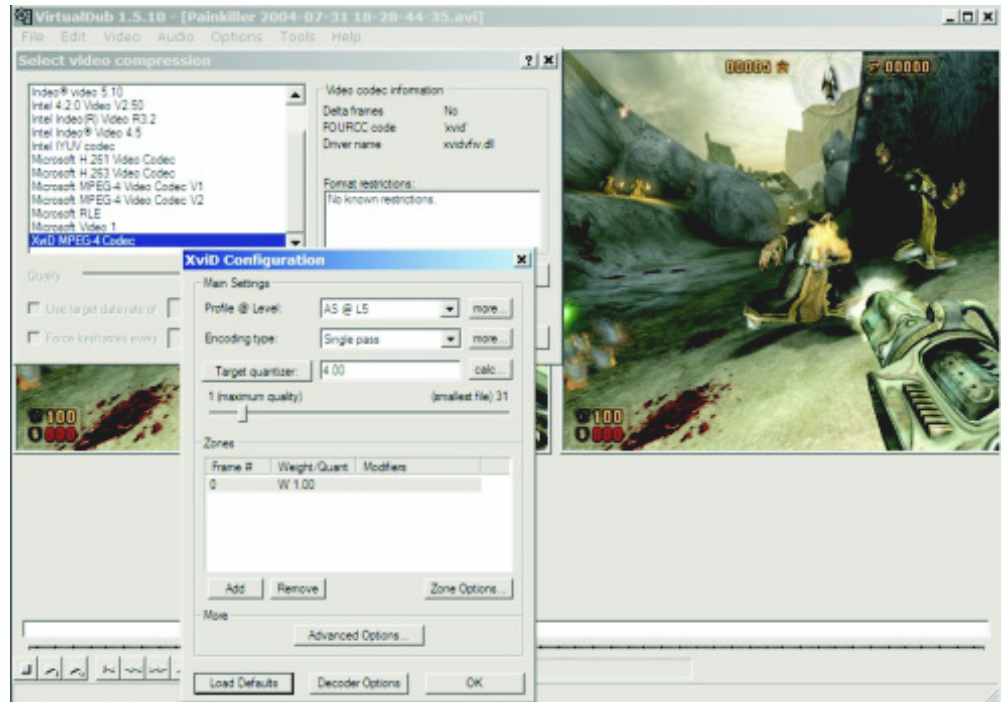


Fig. 6.2: VideoLan Streaming

2) *Virtual DUB*

Virtual Dub is an open source video capture/processing utility for 32-bit Windows platforms, licensed under the GNU General Public License (GPL). It lacks the editing power of a general-purpose editor such as Adobe Premiere, but is streamlined for fast linear operations over video.



It has batch-processing capabilities for processing large numbers of files and can be extended with third-party video filters. VirtualDub is mainly geared toward processing AVI files, although it can read (not write) MPEG-1 and also handle sets of BMP images.

3) *FFmpeg*

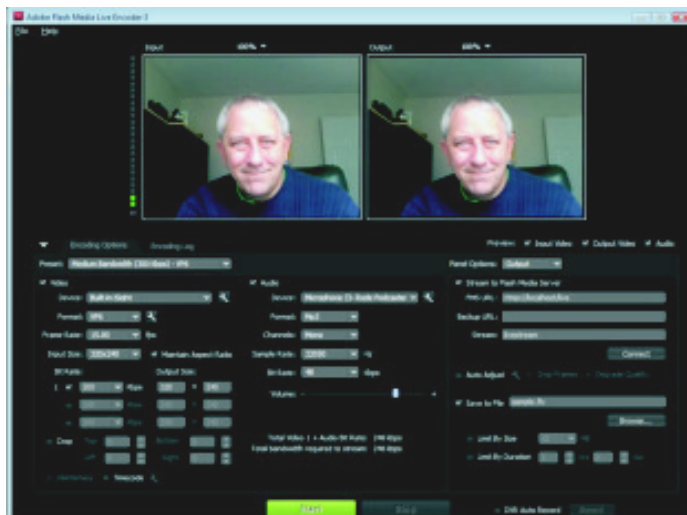
It is a complete Open Source, cross-platform solution to record, convert and stream audio and video. It includes **libavcodec** - the leading audio/video codec library.

```
E:\Youtube>ffmpeg -i "Slam Dunk Bails.flv" -b 350kb "Slam Dunk Bails.mp4"
FFmpeg version SUN-r10464, Copyright (c) 2000-2007 Fabrice Bellard, et al.
configuration: --enable-gpl --enable-pp --enable-swscale --enable-pthreads --
enable-liba52 --enable-avisynth --enable-libamr-nb --enable-libamr-wb --enable-
libfaac --enable-libfaad --enable-libgsm --enable-libmp3lame --enable-libnut --en
able-libogg --enable-libtheora --enable-libvoorbis --enable-libx264 --enable-libx
vid --cpu-i686 --enable-memalign-hack --extra-ldflags--static
libavutil version: 49.5.0
libavcodec version: 51.43.0
libavformat version: 51.12.2
built on Sep 10 2007 10:31:22, gcc: 4.2.1

Seems stream 0 codec frame rate differs from container frame rate: 1000.00 (1000
/1) -> 29.97 (30000/1001)
Input #0, flv, from 'Slam Dunk Bails.flv':
Duration: 00:00:58.7, start: 0.000000, bitrate: 64 kb/s
Stream #0.0: Video: flv, yuv420p, 320x240, 29.97 fps(r)
Stream #0.1: Audio: mp3, 22050 Hz, mono, 64 kb/s
Output #0, mp4, to 'Slam Dunk Bails.mp4':
Stream #0.0: Video: mpeg4, yuv420p, 320x240, q=2-31, 350 kb/s, 29.97 fps(c)
Stream #0.1: Audio: libfaac, 22050 Hz, mono, 64 kb/s
Stream mapping:
Stream #0.0 -> #0.0
Stream #0.1 -> #0.1
Press [q] to stop encoding
frame= 1762 fps=488 q=4.1 Lsize= 3019kB time=58.1 bitrate= 424.0kbits/s
video:2724kB audio:270kB global headers:0kB muxing overhead 0.842219%
```

4) *Adobe Flash Media Encoder*

Adobe® Flash® Media Live Encoder 3 software is designed to enable us to capture live audio and video while streaming it in real time to RED 5 (Open Source) or Flash Media Server software or Flash Video Streaming Service (FVSS).



When high-quality streaming along with a very low bandwidth is our priority, Flash Media Live Encoder 3 can help you broadcast live events and around-the-clock broadcasting such as:

- Sporting events
- Concerts
- Webcasts
- News
- Educational events

6.4 AUDIO DIGITISATION

Analogue audio tapes are available in two formats: open reels and cassettes. They are available in various playing speeds and recoding formats such as mono aural, stereophonic, and quadraphonic with tracking configurations like 2 track or 4 track. To digitise analogue audio data a player needs to be attached with a computer system through audio capture card. This process of analogue to digital conversion of audio data is known as sampling. The process involves sampling the original sound many times per second. The frequency of this sample is measured in Hertz (Hz) and the range of each sample is measured in bits. When digitising sound, the frequency range in kHz determines the sampling rate and the dynamic range i.e., the ratio between lowest and highest sound determines the number of bits per sample.

Various open source products are used for the audio digitisation. Here we are basically using Open Source and Free encoders.

6.4.1 Audio Capturing

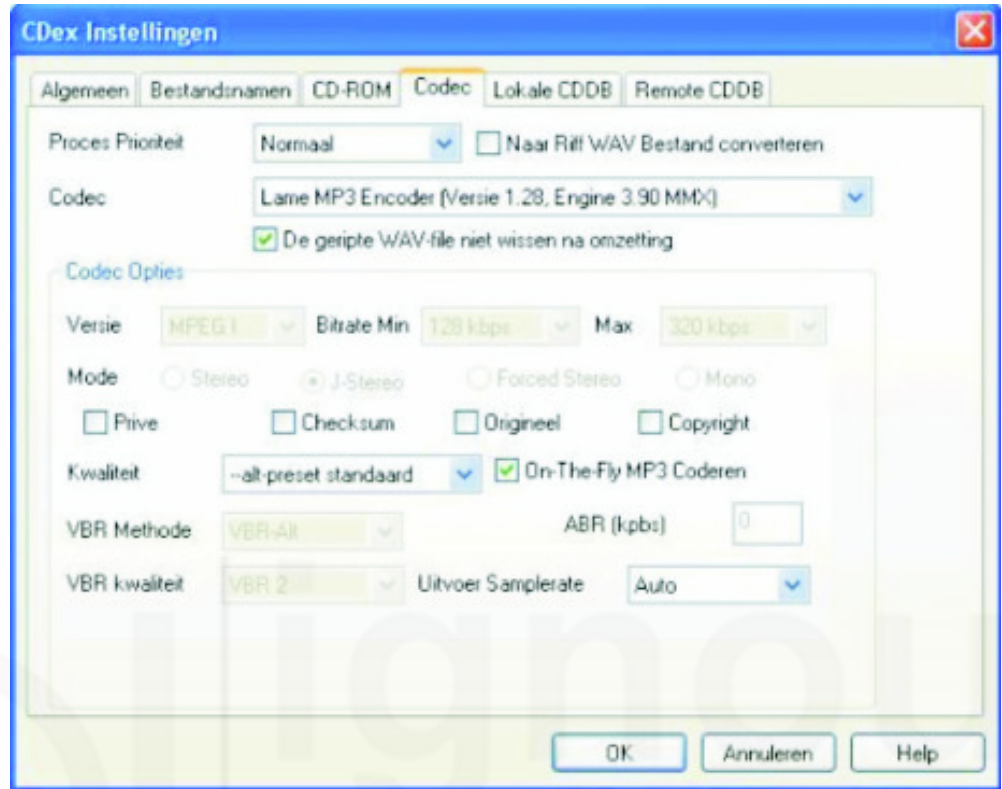
Audio can be captured using microphone. For better quality audio capture and storage of audio data via USB and Portable modes one can use voice recorders like shown in the figure below:



Fig. 6.17: Audio Capturing Devices

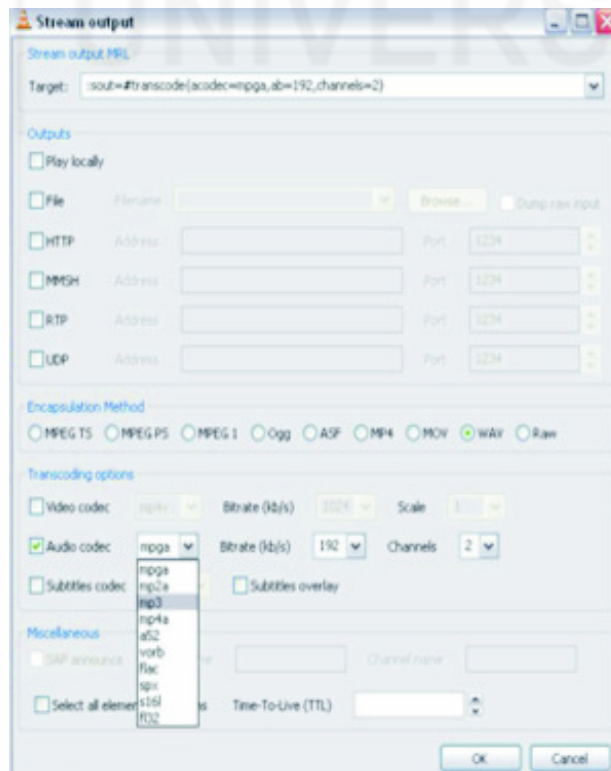
LAME Audio Encoder

LAME is a high quality MPEG Audio Layer III (MP3) encoder licensed under the LGPL. Currently LAME is considered the best MP3 encoder at mid-high bitrates and at VBR



VLC Media Player

As already seen in the Video Processing the VideoLAN can be also used for the audio processing as well.



Self Check Exercise

- Note: i) Write your answers in the space given below.
 - ii) Check your answers with the answers given at the end of this Unit.
- 3) What is LAME encoder?

.....

.....

.....

.....

.....

6.5 AUDIO/VIDEO COMPRESSION

Audio compression algorithms are implemented in computer software as audio codec. A codec is a device or program capable of performing encoding and decoding on a digital data stream or signal. Generic data compression algorithms perform poorly with audio data, seldom reducing file sizes much below 87% of the original, and are not designed for use in real time. Consequently, specific audio “lossless” and “lossy” algorithms have been created. Lossy algorithms provide far greater compression ratios and are used in mainstream consumer audio devices. In addition to the direct applications (mp3 players or computers), digitally compressed audio streams are used in most video DVDs; digital television; streaming media on the internet; satellite and cable radio; and increasingly in terrestrial radio broadcasts.

There are five **MPEG** standards designed with a specific application and bit rate in mind for video compression. They include:

MPEG-1: for Video CD designed for up to 1.5 Mbit/sec application transmitted as .mpg files.

MPEG-2 for the compression and transmission of digital broadcast television between 1.5 and 15 Mbit/sec rate of transmission. Digital Television set top boxes and DVD compression is based on this standard.

MPEG-4 for multimedia and Web compression based on object-based compression technique.

MPEG-7 also called the Multimedia Content Description Interface provides a framework for multimedia content that will include information on content manipulation, filtering and personalization, as well as the integrity and security of the content.

MPEG-21 also called the Multimedia Framework attempts to describe the elements needed to build an infrastructure for the delivery and consumption of multimedia content, and how they will relate to each other. The work on this standard is still on.

Other video compressions are:

DV is a high-resolution digital video format used with video cameras and camcorders. DV images are compressed with a similar but superior technique to motion-JPEG, allowing for higher-quality 5:1 compression. DV video information

is a constant data-rate of about 36 Mbps. The resulting video stream is transferred from the recording device via FireWire (IEEE 1394). IEEE-1394 (“FireWire”) is a communications protocol for high-speed, short-distance data transfer.

H.261 is an ITU standard designed for two-way communication over ISDN lines (video conferencing) and supports data rates which are multiples of 64Kbit/s.

H.263 is based on H.261 with enhancements that improve video quality over modems.

DivX is a software application that uses the MPEG-4 standard to compress digital video, so it can be downloaded over a DSL/cable modem connection in a relatively short time with no reduced visual quality.

6.6 AUDIO/VIDEO STREAMING

With the advent of high end streaming media technology, the concept of doing live/on-demand webcast has gained popularity like never before. Webcasting allows us to extend the reach of audio/video programmes to all corners of the world, with no limitations of physical or geographical boundaries.

Web casting can be either live or on demand. The modalities of these two types of delivery are explained below:

- **Live Webcast:** The transmission of live or pre-recorded audio or video to personal computers that are connected to the Internet. A user who clicks a link to a live clip joins the **live event** in progress. Because the event is happening in real time, fast-forward, rewind, and pause capabilities are not available. Live Web casts are most suitable for high demand live presentations to large geographically dispersed audiences. Participants can attend these virtual presentations from their desktop by visiting a web site. Interaction between instructor and learners occurs in real-time. Participants can use a chat window to type in questions to the presenter during the session. Web casts simulate the look and feel of a live event and can even be recorded for later viewing for those who missed the original web cast. This method is also less expensive than satellite broadcasting.
- **On-Demand Webcast:** Pre-recorded clips are delivered, or streamed, to users upon request. A user who clicks a link to an on-demand clip watches the clip from the beginning. The user can fast-forward, rewind, or pause the clip. Therefore on demand streams can be created from archived live events or recorded clips.

6.7 FILE FORMATS AND CONTENT CREATION

As large amount of document are being digitised and made available online through digital libraries throughout the world, it is pertinent that while archiving documents, physical survival, interpretability, and usability of the data is given importance. For this it is important to give due consideration to encoding standards, file formats and also ensure that the formats are usable and accessible in future. An ideal format for the purpose of archiving would be the one that is a representation rather than a presentation. The most common formats for text archiving are native formats (mostly MS Word), pdf, pdf-a, tex/latex, and xml applications. Other formats that are also prevalent are html, sgml, xhtml. Document formats may be broadly grouped into three types: text based formats, image formats, audio and video formats.

Table 6.1: Standard Digital Formats

Category	Type	Formats	Description
Text formats	Plain text	Text Files (*.txt)	ASCII text files viewed with an editor (such as Edit or Notepad) or with a Word Processor (such as MS Word). Do not contain any kind of formatting on the document (such as bold, italics, font colour, images, etc.).
	Formatted text	<ol style="list-style-type: none"> 1. doc or odf 2. pdf files 	<p>Document files created, viewed and edited using programs such as MS Word or OpenOffice Writer. Formatting features such as bold, italics, justification, adding bullets and numbering, etc., is possible in such formats.</p> <p>Portable Document Format (pdf) was developed by Adobe Systems to transfer formatted documents over the net so that they gave a 'printed document' look and feel. This file type requires Adobe Acrobat Reader which is freely downloadable from the net.</p>
Image formats Audio/ video formats- Audio- Video		<ul style="list-style-type: none"> • Tagged Image File Format (TIFF) 	<ul style="list-style-type: none"> • standard for describing and storing raster image data from scanners, faxes and digital photography applications. It is capable of describing bilevel, grayscale, palette-colour, and full-colour images in several colour spaces. TIFF is extensible, portable and does not favour a particular computer operating system, compiler or processor.
		<ul style="list-style-type: none"> • Graphics Interchange Format (GIF) 	<ul style="list-style-type: none"> • free and open specification for the storage of raster imagery and to facilitate the exchange of digital imagery between different computer platforms and operating systems
		<ul style="list-style-type: none"> • Joint Photographic Experts Group (JPEG) 	<ul style="list-style-type: none"> • JPEG is a standardized lossy image compression mechanism that is designed for compressing full-colour and grayscale images.
		<ul style="list-style-type: none"> • Audio Video Interleave (AVI) 	<ul style="list-style-type: none"> • for storing and playing audio and video data on a PC. The format is limited to a 320 x 240 video resolution and playback rate of 30 fps.
		<ul style="list-style-type: none"> • MPEG-4 	<ul style="list-style-type: none"> • MPEG-4 is built on the MPEG-1, MPEG-2 and Quicktime MOV standards. These files are designed for transmission over a narrow Internet bandwidth,
		<ul style="list-style-type: none"> • Quicktime (MOV) 	<ul style="list-style-type: none"> • The MOV file format was developed by Apple Computer to create, play and stream high-quality audio and video files on both Macintosh and Windows computers using the Quicktime software application
		<ul style="list-style-type: none"> • Real Networks' RealVideo (RM) 	<ul style="list-style-type: none"> • RealVideo was the first streaming video format available on the World Wide Web. A RealVideo clip consists of two parts, a visual track that is encoded with RealVideo codecs (COMpression/DECompression) and an audio track encoded using RealAudio codecs

Table 6.2: Common Formats

Format	File Extension	Notes
XML	.xml	An XML file, validated with DTD or schema specified, is a format suitable for preservation.
SGML	.sgml.sgm	A SGML file, validated, with DTD specified, is suitable for preservation.
HTML	.htm, .html	Hypertext markup language file, which may in principle be validated against a DTD. In practice invalid documents are often produced and used.
XHTML	.xhtml, .htm, .html	XML-conformant HTML file, is required to be well-formed and valid.
DTD	.dtd	Document Type Definition. Defines the rules and syntax applied to a document. To be supplied with an SGML or XML document.
XML Schema	.xsd	An XML schema file. Defines the rules and syntax applied to a document. To be supplied with an XML document.
Pseudo-SGML	.sgm, .sgml, .txt or other	A text file employing some SGML-like formalisms for inserting markup, but not valid SGML. Suitability depends on whether tagging is consistently applied and well-documented, sufficient for later migration.
Various non-SGML encodings in text files	.txt or other	Suitability depends on acceptance as de facto standard in an academic community, plus an assessment of its likely future viability and level of documentation

6.8 SUMMARY

The conversion of analogue sources into digital form and their appropriate storage and processing form an important part of building a digital library. Digitisation is a complex process requiring managerial and technical skills. Proper planning and management help in keeping the cost down, and they also lead to the successful completion of a digitisation project. Digitisation can be carried out in-house or outsourced.

Various technical issues need to be considered in a digitisation project ranging from hardware to software and standards for file formats, file compression and post-processing. Selection of metadata format depends on the nature of the documents as well as the nature and needs of the users.

6.9 ANSWERS TO SELF CHECK EXERCISES

- 1) For converting hard copies into machine readable form three options available are:
 - 1) Keying in the text
 - 2) Scanning and capturing them as image files
 - 3) OCR the files

- 2) Omnipage Pro and ABBYY Fine Reader are two commonly used OCR software.
- 3) LAME is a high quality MPEG Audio Layer III (MP3) encoder licensed under the LGPL. Currently LAME is considered the best MP3 encoder at mid-high bitrates and at VBR.

6.10 KEYWORDS

Charge-coupled device (CCD) : A device for the movement of electrical charge, usually from within the device to an area where the charge can be manipulated, for example conversion into a digital value.

Contact Image Sensors (CIS) : Relatively recent technological innovation in the field of optical flatbed scanners that are rapidly replacing CCDs in low power and portable applications.

Photomultiplier Tubes (PMT) : Members of the class of vacuum tubes, and more specifically vacuum phototubes, are extremely sensitive detectors of light in the ultraviolet, visible, and near-infrared ranges of the electromagnetic spectrum.

6.11 REFERENCES AND FURTHER READING

<http://www.librarydigitisation.com/>

<http://www.records.nsw.gov.au/recordkeeping/advice/designing-implementing-and-managing-systems/digitisation-of-analogue-audio-and-video>

<http://www.jiscdigitalmedia.ac.uk/digitisation>

http://www.tape-online.net/Short_Guidelines_Video_Digitisation.pdf

<http://travesia.mcu.es/portalnb/jspui/bitstream/10421/6742/1/digitisation.pdf>

<http://www.slq.qld.gov.au/about-us/projects-and-partnerships/distributed-collection-of-queensland-memory/digitisation-toolkit/what-is-digitisation>