International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 5

Different File Formats used in Digital Preservation

Padmavati S Tubachi
Librarian, DPM's Shree Mallikarjun College
Canacona, Goa, India

ABSTRACT

Digitization is the process where information contained in analogues items like images, text, audio and video are captured and converted into a digital format. In this format information is organized into discrete units of data called bit/s that can be separately addressed as byte which is a multiple bit group called byte/s) Every type of information is represented in digital form. Digital information can be saved on any medium that is able to represent the binary digits (bits) 0 and 1. The meaningful sequence of bits with no intervening spaces, punctuation or formatting is called as bit stream. A file is nothing more than a sequence of bits and the file format is interpreting the bit stream (Barve, S 2007). Digital information is produced in a variety of standard and proprietary formats like ASCII, Common image formats, word processing, spreadsheets, database documents, formulae, charts, multimedia files, sound and video. (Ramareddy), (Angadi, M 2004). A defined arrangement for discrete sets of data that allow a computer and software to interpret the data is called a file format. A file format is a specific format in which a file is saved (Prasad, A. R. D).

Keywords: *Digitization, File Formats*

INTRODUCTION

Since the advent of computers in the early part of the last century, the society has been moving into the electronic world at an increasingly rapid pace. Now a day's many libraries are adopting the digital library system either by digitizing the existing collection or acquiring born digital collection since digitizing information makes it easier to preserve, access, and share and information can be made available to people worldwide. Digitization is the process where information contained in analogues items like images, text, audio and video are captured and converted into a digital format. In this format information is organized into discrete units of data called bit/s that can be separately addressed as byte which is a multiple bit group called byte/s. This is a binary data that computers and other devices with computing capacity can process. (Rouse, M 2007). Every type of information is represented in digital form. Digital information can be saved on any medium that is able to represent the binary digits (bits) 0 and 1. The meaningful sequence of bits with no intervening spaces, punctuation or formatting is called as bit stream. A file is nothing more than a sequence of bits and the file format is interpreting the bit stream (Barve, S 2007). Digital information is produced in a variety of standard and proprietary formats like ASCII, Common image formats, word processing, spreadsheets, database documents, formulae, charts, multimedia files, sound and video. (Ramareddy), (Angadi, M 2004). A defined arrangement for discrete sets of data that allow a computer and software to interpret the data is called a file format. A file format is a specific format in which a file is saved (Prasad, A. R. D).

1. DEFINITION

Webopedia defines it as a format for encoding in a file; each different type of file has a different file format. The file format specifies first whether the file is a binary or ASCII files and second how the information is organized (Webopedia N. D)

Brown (2006) defined file format as 'The internal structure encoding of a digital object which allows it to be processed, or to be rendered in human accessible form, A digital object may be a file or a bits ream embedded within a file. (Barve, S 2007)

NDIIP website defines formats as packages of information that can be stored as data files or sent via network as data streams (aka bit streams, byte streams)

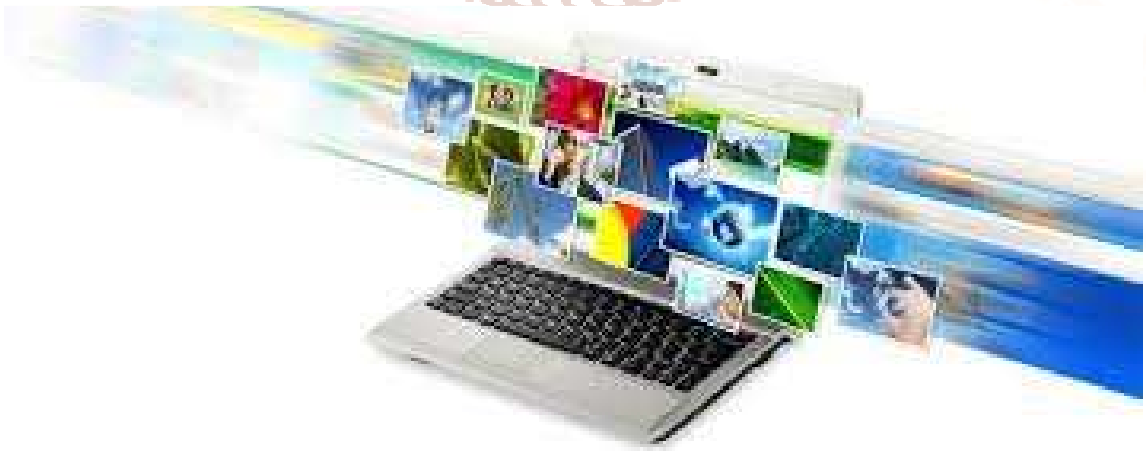
Every type of information is represented in digital form. Different file formats specify how binary digits represent the intellectual content created by a digital object's creator. Every object in a digital format needs to have a name or identifier which distinctly identifies its type and format; this is achieved by assigning file extensions to the digital objects. The file extensions denote formats, protocol and rights management that are appropriate for the type of material. Every time when a document or a graphic is created in a computer, item is saved with a particular file format. E.g. **.doc** file format identifies it as a MS word file. Different applications (programs) store data in different formats. Different file formats are designed to perform different/particular tasks/functions like open…, Save…, Save as…, import…, export…, etc. The accessibility of that information is highly vulnerable in today's rapidly evolving technological environment because file formats encode information into forms that can only be processed and rendered comprehensible by very specific combination of hardware and software.

When a file is saved using a particular program, that program assigns its own individual format that is known as a native file format. File format appear as a 3 letter suffix or 'extension' after the name of the file. Filename extensions are usually noted in parentheses if they differ from the format name or abbreviation. e.g. photoshop=.psd, MS excel=xls, Coreldrwa=.cdr.

The files that a program will open or save as other than its native form vary with the individual program that one is using are called Non native forms. E.g**. .txt**

Different file formats are used to store different media types like text, graphics, graphics, pictures, musical works, computer programs, databases, models and designs, video programs and compound works combining many types of information. Digital file types describe the types and characteristics of the files produced from digitization of original record materials as well as the standard or most common data formats that are used to store digitized records. There are different categories of file formats available for different applications. The official categorization of file formats is the MIME type, provided by IANA. Main categories of formats are: Application, Audio, Images, Message, Model, multipart, text, and video (Barve, S 2007).

2. FILE FORMATS THAT ARE USED IN THE DIGITAL LIBRARY ARE:



	Format	Abbreviation	File extension
File format for Unstructured text	American standard code for Information Interchange	ASCII	.txt
File formats for structured Text 	Standard Generalised Markup Language	SGML	.sgml
	Hypertext Markup Language	HTML	.html
	Extended Markup Language	XML	.xml
	Portable Document Format	PDF	.pdf
	PostScript	PostScrift	.ps
	Texture Format	TEX	.txt
File formats for Images 	Portable Document Format	PDF	.pdf
	Bit Map Page (Windows)	BMP	.bmp
	Ventura Publisher	IMG	.img
	Joint Photographic Expert Group	JPEG	.jpeg
	PC Paint (B & W mode)	PCP	.pcp
	PC Paint Brush (Color & B & W)	PCX	.pcx
	Photoshop	PSD	.psd
	True Vision Targa	TGA	.tga
	Portable Network Graphics	PNG	.png
	Tagged Image File Format	TIFF	.tiff
	Tagged Image File format with Group 4 Fax Compression	TIFF-G4	.tif
	Still Picture Interchange File format	SPIFF	.spf
	Photo CD (Kodac)	PCD	.pcd
Audio Video File Formats 	Waveform Audio	WAVE	.wav
	Audio Interchange Format	AIFF	.aif
	Creative Voice	VoC	.voc
	Musical Instrument digital Interface	MIDI	.midi
	Sound	SND	.snd
	Audio	AU	.au
	Real Audio format (Progressive Network)	RAF	.ra
	Audio Visual interface	AVI	.avi
	Macromedia Flash Movie	FLA	.fla
	Motion Picture Expert Group	MPEG	.mpg
	MPEG Audio layer 2	MP2	.mp2
	MPEG Audio layer 3	MP3	.mp3
	QuickTime for Windows Movie	MOV	.mov
	Autodesk FLIC Animation	FLC	.flc

3. TOOLS TO MANAGE FILE FORMATS

- Format Identification for Digital Objects (FIDO): Command line tool to identify the file formats of digital objects and is designed for simple integration into automated workflows.
- Bit Curator Access: Open source software that supports the provision of access to disk images Webinar on using Bit Curator.

- Apache Tika; Toolkit detects and extracts metadata and text from over a thousand different file types (such s PPT, XLS and PDF)
- BWF Meta Edit: Free, open source tool that supports embedding, validating and exporting of metadata in Broadcast WAVE Format files.

I. EVALUATION OF DIFFERENT TYPES OF FILE FORMATS:

The aim of digital preservation is to ensure that records are filed and made accessible throughout time. File formats require proper hardware and software in order to interpret and display the data for user. But continuous technological changes in software and hardware make old formats become unreadable and unusable. There are several file formats which are available today are incredibly complex, making the binary code meaningless to a user if the required software is not available to interpret the format.

1. CRITERIA FOR EVALUTION:

The selection of file formats for creating digital documents should be determined not only by the immediate requirement but also long time preservation. Selected format should also meet the requirement for both preservation of authenticity and ease of access. Following criteria should be considered when selecting the file formats.

- Authenticity
- Support
- Documentation quality
- Disclosure
- Stability
- Intellectual property rights
- Complexity
- Viability
- Re-usability
- Interoperability
- Metadata support
- Open standards
- Processability
- Permanancy
- Ubiquity

While selecting file formats quality and functionality factors should be considered. Some of them include:
Text: Support for integrity of document structure and navigation, Support for integrity of layout, font and other design features, Support for rendering for mathematics, formulae, diagrams etc.

Still Images: Clarity (support for high image resolution), color maintenance, support for graphic efforts and typography, Support for multispectral bands.

Sound: Fidelity (support for high audio resolution), Support for multiple channels (including note based e.g., MIDI), Support for downloadable or user defined sounds, samples and patches.

Moving Images: Clarity (Support for high image resolution), Fidelity, Support for multiple sound channels.

A. Text formats: Text and image based contents are stored and presented as

a. Unstructured or simple text.
b. Structured text
c. Page description language.

Text document formats are the simplest form of digital data. Denoted by the ".txt" file extension, these documents contain only a simple string of characters and are devoid of more complex information. **.doc, .pdf, .wpd, .ps**

Simple Text: ASCII is used as a format for facilitating exchange of data from one software to another.

Advantages: It is compact, economic to capture & store, searchable, interoperable.

Disadvantages: It cannot be used for displaying complex tables, or mathematical formulae, photographs or diagrams, does not store text formatting information i. e font, font size, italics etc.

b. Structured text formats: structured text formats attempt to capture the essence of document by marking up the text so that the original form could be recreated. SGML is one of the structured text. It is flexible language de facto Markup language to control the display. They can easily display complex tables and equations.

TeX is used for formatting highly mathematical text. It is one such format that allows greater control over the formatting of errors.

Office application file formats are used for word processing, spreadsheets and PowerPoint presentation. Most common office suite used is Microsoft Office.

The file formats used are binary in nature and not publicly published. MS Word, MS Excel and MS PowerPoint use the DOC, XLS and PPT formats. Word and Excel can save their data in Rich Text Format file format. This is a binary file format for cross platform document interchange.

Other software that is used as office application is Open Office and Star Office:

Open Office: It is fully fledged open source office application suite comprising word processor, spreadsheet, presentation software, graphic editor and a database program. All the files are ae compressed and stored as a single zip compressed file. It is available on multiplatform.

Star Office: It shares the same code base as Open Office. It is released under proprietary commercial license.

Adobe's Portable Document Format: PDF is a file format developed by Adobe System for secure and reliable electronic distribution and exchange. The format is able to preserve the look and integrity of the original document regardless of the application and platform used to create it evem if it contains complex combination of tet, graphics and images.PDF is very useful as a format for multiplatform document exchange and distribution and for sharing information.

Some office document formats are:

i. Microsoft applications-DOC, XLS (Spreadsheet), PPT (PowerPoint)
ii. Open Office - SXW (text), SXC (Spreadsheet), SXI (presentation),
iii. OASIS, ISO/ IEC – ODT (text), ODS9 spreadsheet), ODP (presentation)
iv. PDF (text & presentation) -Adobe

B. Graphic Files/Image Files:
A type of digital object that is created from the digitization of still image (textual document and photographs) originals. Anything on the page that is not actual text, from simple drawing to fully active images. A still image is data in which a grid or raster of picture elements (pixels) has been mapped to represent a visual subject, e.g., the page of a book or a photograph.

Graphic styles may be divided into two major types:
Raster Graphics: Simply a collection of dots known as 'pixels'. Hence they are also called 'bitmaps. The color of each pixel is described by one or more information channels – separated into the primary hues – Red, Green and Blue or in a single stream of color mapped data. Raster images are simple images and are hence most suitable for interoperability.

Disadvantages: They do not scale well and scaling may lead to a loss of resolution and hence poorer picture quality

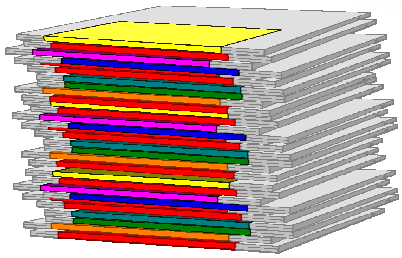
Vector Graphics: An image as a collection of vector equations.

Advantages: no break in figures, smooth curves and lines irrespective of the size of the image or resolution
Disadvantage: They take longer to draw require more storage space.

Some of the Image File formats to store scanned images: gif, jpej, tiff, bmp

1. TIFF – Tagged Image File Format: TIFF stands for Tagged Image File Format. This format was designed to overcome the problem of application dependence. It was originally designed to become the standard format. The format was intended to be capable of handling just about any possibility. This file format is generally used when graphic files need to be moved between different computer types (For example: PC to Mac and vice-versa).

Advantages:
- Allows for high resolution



- Highly flexible – there are several possibilities of how a .tif image can be saved.
- Is supported by most scanning and image editing software
- Format works well for both on-screen display and print of photographs
- Format works well in page-layout programs as it allows editing of file attributes
- Can differentiate between types of images in three categories – Black & White, Gray scaled, Color
- Supports 24-bit true color
- Multiple images and data in the same file
- Tags in file header (information on size, compression)
- Loss-less format, useful for archival images
- Platform independent
- Format useful for future modification – can edited without compression loss

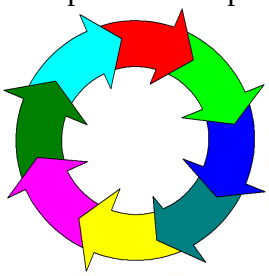
Disadvantage:

- tif files from IBM PCs are usually different from those of Macs
- The format is meant to be a standard. But many manufacturers have tried to 'improve' on it and it is no longer a standard
- No application can claim to support all the variations of .tif files
- Files tend to be larger than many other formats and hence takes longer to print tif files from IBM PCs are usually different from those of Macs.
- Size of images is very high.

2. JPEG – Joint Photographic Expert Group: JPEG stands for 'Joint Photographic Experts Group' that designed this format for high compression. It is one of the most popular image formats on the web. It discards extra data –or data beyond what the eye can see and hence has good compression capabilities.

Advantages:

- It is a web standard, strongest format for web images and printing images.
- Images are small. Quality is superior.
- It is quick to transfer over the Internet. Best method for online viewing
- It is flexible, allows user to choose between image size and picture quality.
- Useful to incorporate large number of files.
- It supports millions of colors.
- Variety of compression capability. Picture quality remains accurate at lesser levels of compression & usually better compressions in photographs.



Disadvantage

- Lossy compression format
- Very high levels of compression result in loss of picture quality
- Repeated saving in this format results in loss of quality. If repeated saving is required, it should be first saved repeatedly in another format and final image is to be saved in only in .jpg

3. WINDOWS BITMAP (BMP): Bitmap files or .bmp files are the standard Windows Raster format. These file lay emphasis on quick display. It hence stores images in the uncompressed form. The obvious trade off is that bmp files are space eaters.

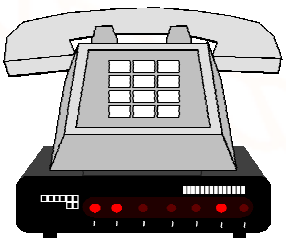
Advantage:

- Very quick download time
- Supports 1,4,8,24 bits of color per pixel

Disadvantages:

- Uncompressed – hence occupies more storage space
- Transfer from other formats to bmp format is not handled well – loss of picture quality

4. GIF – Graphic Interchange Format: One of the most popular graphic file formats on the Internet, Graphic Interchange Files (.gif) was developed by CompuServe with the main purpose of archiving information. .gif images are usually scanned stand-alone pictures that are not 'drawn' using an application program."



Advantages:

- Highly compressible
- Very useful format when large number of images are to be incorporated
- A standard web format - most browsers have a .gif viewer
- Small size allows quick transmission over the net

- Later versions of gif support transparent colors (which may be used in the background of WebPages). Also support animations
- Very old format
- Lossless compression format
- Less storage space
- Strong candidate for graphic art and drawing.

Disadvantage:

- Not a good choice for non-web images
- Supports only 256 colors – if true-color is desired, it is better to use .tif or .jpeg
- If a higher resolution file is converted to the .gif format, the extra colors that .gif cannot handle are thrown away from the image, resulting in poorer quality
- Limited to 256 colors

4. JAS: JAS format is from JASC Inc. This file format is designed to create the smallest possible image file for 24 bits per pixel color images and 8 bits per pixel gray scale images.

JAS results in "lossy" transformations. Saving and retrieving an image using the JAS file format will result in some loss of image data. The amount of loss is dependent on the compression level that one has selected in the application. By using the lowest possible value for the file compression amount of loss can be reduced.

Advantages:

- Very high level of compression

Disadvantages:

- Supports only up to 24 bits
- There is some loss of data during compression – poorer quality

5. MAC files: An application specific format, MAC files are used in Macintosh Mac Paint application.

Advantages:

- Yields small file sizes
- It is still readable by most Mac machines

Disadvantage:

- Supports only 1-bit per pixel – only monochrome files can be converted to .MAC files MAC format requires a fixed image size (576x720). Images converted to .MAC will be cropped to fit that size

6. PBM Plus files: PBM files are Portable Bitmap files, popular in UNIX

Advantages:

- Allows for extensive manipulation of grayscale and color bitmap images.
- Three libraries- pbm, pgm, ppm allow conversion of bitmaps to every other popular graphic file format

7. PCX: This file format was developed by Zsoft for its Paintbrush software. This format is used by IBM computers. Version 5 is most popular and supports 1, 4,8,24 bits per pixel.

Advantages:

- .pcx format is supported by more applications than any other format
- Highly flexible
- Accommodates any size and any number of colors
- Image compression is an integral part of this format

Disadvantages:

- Causes some difference in images if using different versions

8. PICT files: This is the standard Apple Macintosh graphic file format

Advantages:

- Accepted by many applications
- May be imported/exported using clipboard (cut, copy, paste) to almost any text or graphics program

Disadvantages:

- Used only by Macs

9. RAS: Sun Microsystems developed the 'Raster' file format. There are three specific types of .RAS files

10. RAW Files: This flexible format consists of a stream of bytes that describes the color information in the file. Each pixel is described in binary format where 0=black and 255=white.

This format is used to transfer documents between different applications

11. TGA FILES: Targa files or TGA files were developed by True vision. It is widely used by high-end paint programs and ray tracing packages

Advantages:
- Can handle images with up to 16 million unique colors
- May be saved in the compressed on uncompressed forms

Disadvantages:
- Not as widely used as .PCX or .TIFF files
- Was designed for use on systems that run MS-DOS color applications

12. WPG FILES: Word Perfect Graphic files or WPG is an application specific file format. It first appeared with Word Perfect 5.0 and changed with Word Perfect 5.1

Advantages:
- Can contain bitmap, line art and vector images

Disadvantages:
- An application other than Word Perfect is used to view a WPG file containing bitmapped and vector elements, vector elements will be discarded
- Supports only up to 8-bits of color per pixel

13. DjVu – déjà vu (a free file format)
- Useful for compressing documents with s continuation of text and images
- File format to save scanned images especially with text.
- Advanced technology for image layer separation of text and images.
- High quality readable images, stored in minimum space – useful for web.
- Progressive loading – useful for web.
- Format used for Million books project

14. AMIGA INTERCHANGE FILE FORMAT (IFF) is a application specific format developed by Commodore Amiga for transfer to and from its own computers. This data format is designed for storage and exchange and manipulation of data between many different programs

Advantages:
- Image files can be easily processed by several programs sequentially, one after the other
- When used with pictures, it preserves quality of color
- .iff stores sound data, text and configuration data

Disadvantages:
- Can be used only on Amiga computers

15. PNG – Portable Network Graphics
- A new format
- Created to improve on GIF format
- Supports 24-bit color or grayscale
- Provides for variety of transparency
- Lossless data compression
- Disadvantage
- New so old software does not support

16. COMPUTER GRAPHICS METAFILE: is an ANSI standard graphic file format used to exchange vector data. It stores vector information as opposed to other file formats like gif, bmp, etc that store raster information. As a result, vector files allow viewing with pan, zoom and loss of detail. cgm files can be viewed directly on the web using a web browser plug-in

Advantage:
- Can be used as a format to transfer vector graphics such as post script files from Unix programs to Mac programs without a loss of resolution
- Widely supported by UNIX & Windows.
- Beginning to be supported by Macs

Disadvantages:
- Cannot be used with native browsers such as Netscape
- **.cgm** files are very large (at least 10 times larger than .gif files)

17. XPM: The XPM (XpixMap) format is a de facto standard for creating icon pixmaps for use in GUis based on the X window system. It consist of ASCII image format and a C library, the XPM format defines how to store color images in a portable way while the associated library provides a set of functions to store and retrieve images to and from XPM format.

18. SVG: The Scalable VECTOR GRAPHICS: This is meant for Vector graphics. It is used for geometrical primitives such as points, lines, curves and polygons to represent images in graphics.SVG consist of an XML based file formats and a programming API for graphical applications.

Advantages:

- It offers a way based on open standards to render graphic optimally on all types of devices.

Disadvantage;

- The usage of SVG on the Web is limited.

C. AUDIO FILES: There are two major groups of audio file formats .
Those using lossless compression like WAV, FLAC
Those using lossy compression like MP3, Ogg Vorbis

1. **AU FILES**: Most commonly found on the web, it is required by PC users to load applications such as Waveform Hold and Modify to play these files. Macs need different sound applications to play this file type

2. **MIDI FILES**: This is used by files following the Musical Instrument Digital Interface standard. These are used mostly in audio control in Multimedia industry. MIDI file specification allows for lengths to be specified as a variable number of bytes.

3. **AIFF FILES**: Audio Interchange File Format (aiff) was developed by Apple. Although it was originally made for Macs, now it can be used by other platforms too. It is a very good audio file format for use on the Internet. It can also be used in Multimedia authoring on both Macs and Windows

4. **MPEG layer 3**: (.mp3)It is currently the most popular of the audio file formats. Its hallmark is its CD-quality of music. MP3 allows for very high levels of compression. A minute of music may constitute approximately 1 Mb file. An MP3 player is that is readily available on both Macs and Windows is required to play this file type. It is a popular lossy compression audio format.

Advantages: Can produce good production of original. Music files encoded with MP3 are popular on music exchange and download sites on the Internet due to the relatively small size of the files and the wide availability of free software on PCs that allow easy creation, sharing, collecting and playing MP3 files.

Disadvantages: MP3 makes use of patented technology and so software and devices that support it are subject to royalty payments in those countries that recognize software patent.

5. **VOC**: Creative Lab's Sound Blaster uses the .VOC file. They are designed for storing digitized voice data and hence the name. They can however also handle any digitized sound in any of a variety of formats. The VOC files have a two part structure. The header block which defines the contents of the file, the data block which actually contains the audio information

6. **WAV**: Wave files are a commonly used file format on Windows machines. The WAV format is most commonly used with an uncompressed, lossless storage method (pulse code method resulting in comparatively large audio files. It can be used on the Internet and is good for multimedia authoring. Advantage: It is flexible and handles both compressed and uncompressed storage formats. Lossless storage method

7. **Real audio**: RealAudio is a proprietary audio format developed by Real Networks for low bandwidth usage. Many Radio stations use RealAudio to stream their programs over the Internet.

Advantages: Variety of audio codecs from low bit rate to high fidelity formats, streaming audio format.

8. **Ogg Vorbis**: the format originated from the Xiph.Org foundation. It uses the Vorbis lossy

audio compression scheme. The audio is wrapped up in Ogg container format. Is a compressed audio format. It is free of patents and royalty payments.

D. VIDEO FILES: Video files have become most popular with films being available and viewed on VCDs & DVDs. It is important to be aware of the video file formats are these are the most disk space-occupying (bulky) types. Video is an affordable medium that is easier to use but I is subjected to detererіotion. Video formats are subjected to change. Some of the video file formats are Mpeg, avi, mov etc

1. **AUDIO-VIDEO INTERLEAVED** file format was developed by Microsoft. An AVI player and drivers are required to play this format. They are readily available both in Mac and Windows machines. With the player, AVI plays full motion picture video with audio in a small window at about 15 frames per second

Advantages:
- AVI comes with Windows, so no drivers need to be obtained and the built in media player, although the drivers are better for faster machines and will improve quality.
- AVI is a popular standard; many videos have been produced in the format because of its non requirement of drivers.
- The quality of AVI files with good drivers and good hardware can be quite impressive.
- The majority of AVI files have audio

Disadvantages:
- AVI's are no longer developed by Microsoft and are left to be developed by a third party.
- Frame rates are not as high as other video formats (such as MPEGs)
- Not much development going on in this file format

2. **MOV/.MOVIE FILES**: Movie files are the common format used in QuickTime movies, the Mac native video platform.

3. **QT FILES**: QuickTime files.

The latest version is used on Macs today

4. **RAM FILES**: Developed by Real networks for streaming video. This requires Real Player for viewing

5. **MPG/.MPEG FILES**: The standard Internet format uses MPEG compression scheme. This format can be used on Macs by converting into QuickTime movies using applications such as 'Sparkle

Advantages:
- Defines "Rights Expression Language" standard
- Sharing digital rights/permissions/restrictions for content from content creator to consumer
- XML based file system
- Can communicate machine readable license information in a "ubiquitous, unambiguous and secure" manner.
- The main objective of the MPEG-21 is to define the technology needed to support users to exchange, access, consume, trade or manipulate Digital Items in an efficient and transparent way.

6. **MOTION PICTURE FILE**: A high resolution moving image recording often with synched audio produced from either original physical format. Bit-depth, pixel array, frame rate per second and color encoding.

7. **DPX**: Digital Moving Picture Exchange: The DPX file format is a pixel based (raster) image in which each content frame is a separate data file linked by metadata to play in the correct sequence.

8. **DCP**: Digital Cinema Package is a collection of digital files used to store and convey digital cinema audio, image and data streams.

SPREADSHEETS: XLS, ODS, SXC

DATABASES: Microsoft database (MDB)

POWERPOINT: PPT, ODP, SXI

MARKUP LANGUAGES: HTML, HTM, SGML, XHML, XML

COMPRESSION: Zone Information Protocol (ZIP)

OTHER: Executable (EXE)

II. CONCLUSION:

Choosing a suitable file format for data preservation and sharing is vital for the sustainability of future access and reuse of data. File format types should be considered and decided upon before the

commencement of data collection. eg Information lost by storing data using a lossy image, sound or video format cannot be recovered. Migrating data from an unsuitable format to a more sustainable option is always difficult and expensive, and may in some cases be impossible. Uncompressed non-lossy file formats take up a lot more storage space that needs to be taken into account when budgeting for storage. File formats apply to documents, images, audio files, video files and research for simple integration into automated workflow.

III. REFERENCES:

1. http://www.webopedia.com.
2. http://www.archives.gov.
3. https://www.kb.nl.
4. http://www.digitalpreservation.gov.
5. http://www.dpbestflow.org.
6. https://en.wikibooks.org/wiki.
7. Angadi, M. "Planning and implementing digital libraries." *CUG-lecture series.* March 17, 2014.
8. Barve, Sunita. http://dlissu.pbworks.com.
9. Kim, Yunhyong, and Seamus Ross. "Digital forensic formats: Seeking a digital presesvation storage container format for web archiving." *The international journal of digital curation*, 2012.
10. *Selecting formats for digital preservations: lessons learned.* www.niso.org.