Introduction to Digitization:
An Overview

July 16th 2008, FIS 2308H

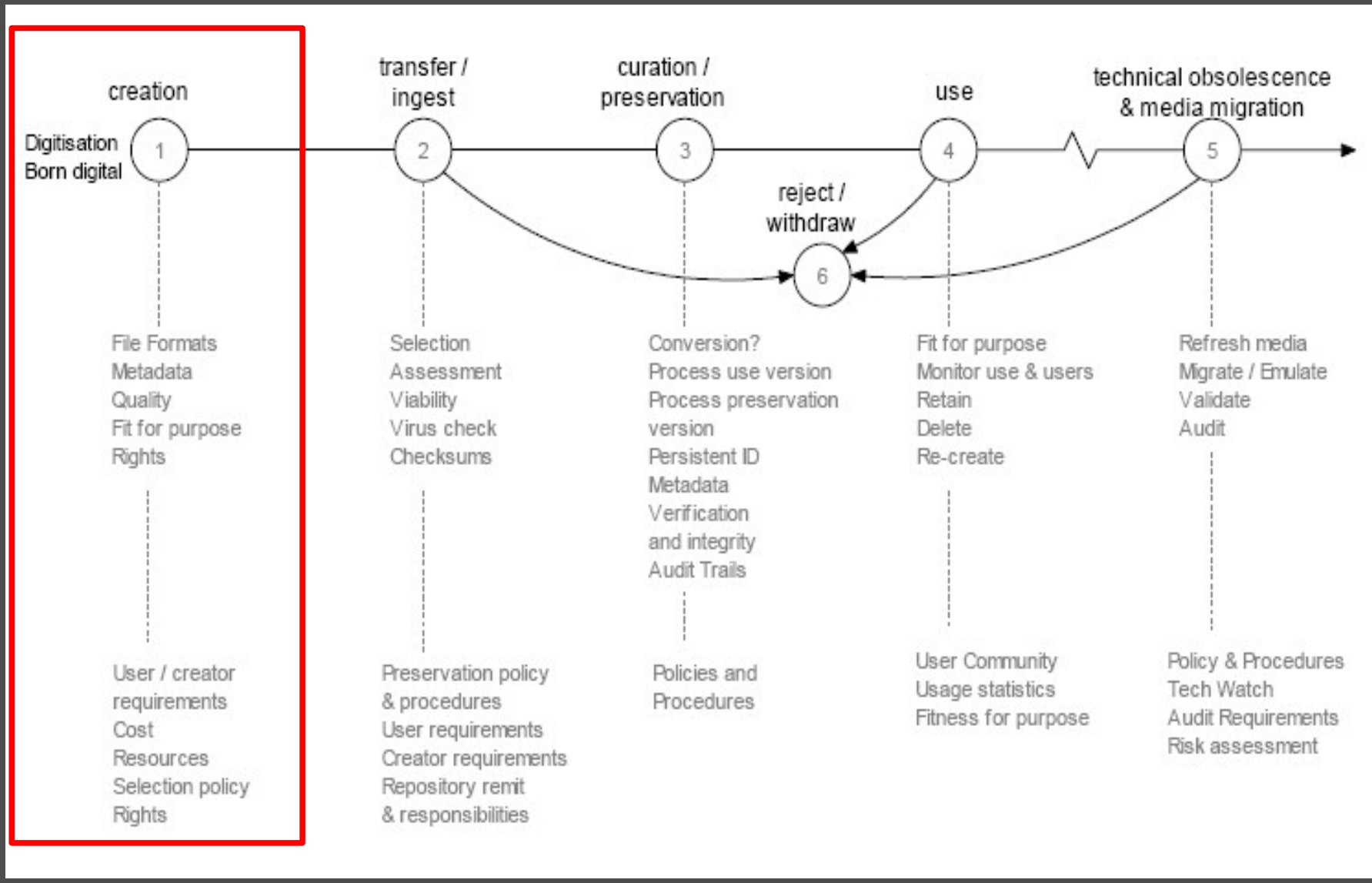
Andrea Kosavic

Digital Initiatives Librarian, York University

Introduction to Digitization

- Digitization in context
- Why digitize?
- Digitization challenges
- Digitization of images
- Digitization of audio
- Digitization of moving images
- Metadata
- The Inuit through Moravian Eyes

Digitization in Context



http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Why Digitize?

- Obsolescence of source devices (for audio and moving images)
- Content unlocked from a fragile storage and delivery format
  - More convenient to deliver
  - More easily accessible to users
  - Do not depend on source device for access
- Media has a limited life span
- Digitization limits the use and handling of originals

Why Digitize?

- Digitized items more easy to handle and manipulate
- Digital content can be copied without loss
  - Analog formats degrade with each use and lose quality when copied
- Can be delivered to a far reaching audience over internet
- Can add metadata, ie. MPEG7 allows enhanced searching

Digitization challenges

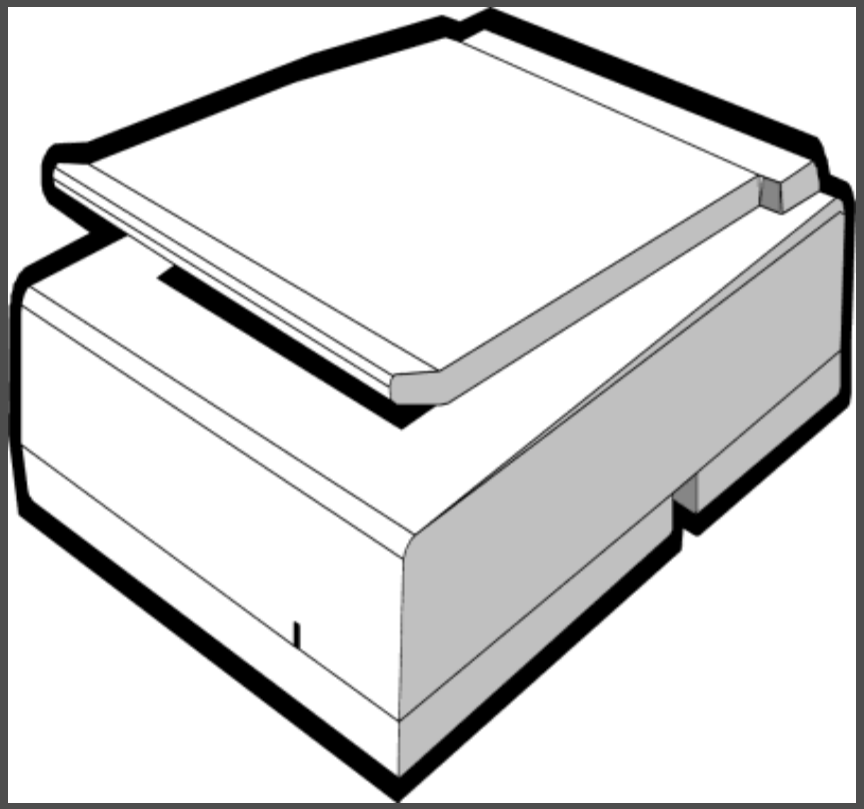
- Multiple formats to choose from
- Can’t match quality to that of the source
- Analog version is still considered the preservation master copy
- Expensive
  - Digitization equipment
  - Storage
  - Staff time

Digitization challenges

- Storage…we're talking TBs!
  - CD quality audio is 520 MB per hour
  - DVD-quality video = 13 GB per hour
  - Broadcast quality video = 75 GB per hour (ITU-R BT.601)
- Technical limitations
  - Compression algorithms still evolving
  - High bandwidth required for transfer
    - For an audio file recorded at preservation standards, it takes 5x the duration of the file to transfer over T1 network

Digitization of Images

- Introduction to various materials
- The Digitization Process
- Common Image Formats



http://www.wpclipart.com/computer/hardware/scanner.png

Multiple format types

- Maps
- Plans
- Manuscripts
- Plain Text
- Drawings
- Paintings
- Photographs
- Negatives
- Microfilm
- Transparencies
- Slides
- Charts & graphs

Flatbed Scanner

- Good for smaller plans / maps, photographs, plain text
- Auto Sheet Feeder attachments allow for fast digitization of single sheets
- Scans a variety of resolutions 200 dpi – 9600+ dpi
- Scans at 1 bit (black and white), 8 bit (grayscale), and 24 or 48 bit (colour)

Image: http://content.answers.com/main/content/img/CDE/_CREOSCN.JPG

Flatbed Scanner Tips

- Scan plain black and white text at 1 bit, this avoids grey background
- Scan black and white drawings with shading at 8 bit, or 1 bit with half-toning
- Scanning colour images with text is difficult, if scanning at 24 bit, text quality will suffer, will have to play with settings or scan separately

Digital Camera

Images:
http://www.digital-photography.org/CruseGmbHdigitalscannersystem/Cruse_repro-stand_copystand.htm

Digital Camera – Book Cradle

- Can be used with a book cradle
- Book cradle keeps pages flat without damaging book
- Book cradle necessary for rare manuscripts
- Ideal for maps, plans, manuscripts, drawings paintings

Image: http://www.i2s-bookscanner.com/visualisationMiniature.asp?image=upload/produits/gammes/acc_BC1590.gif

Specialized Scanner Types

- Microfilm scanner
  - Specialized for microfilm
- Slide/Negative scanner
  - Higher resolution capture
  - Come with specialized cartridges to hold different sizes of film
- Photo scanner
  - Higher resolution capture

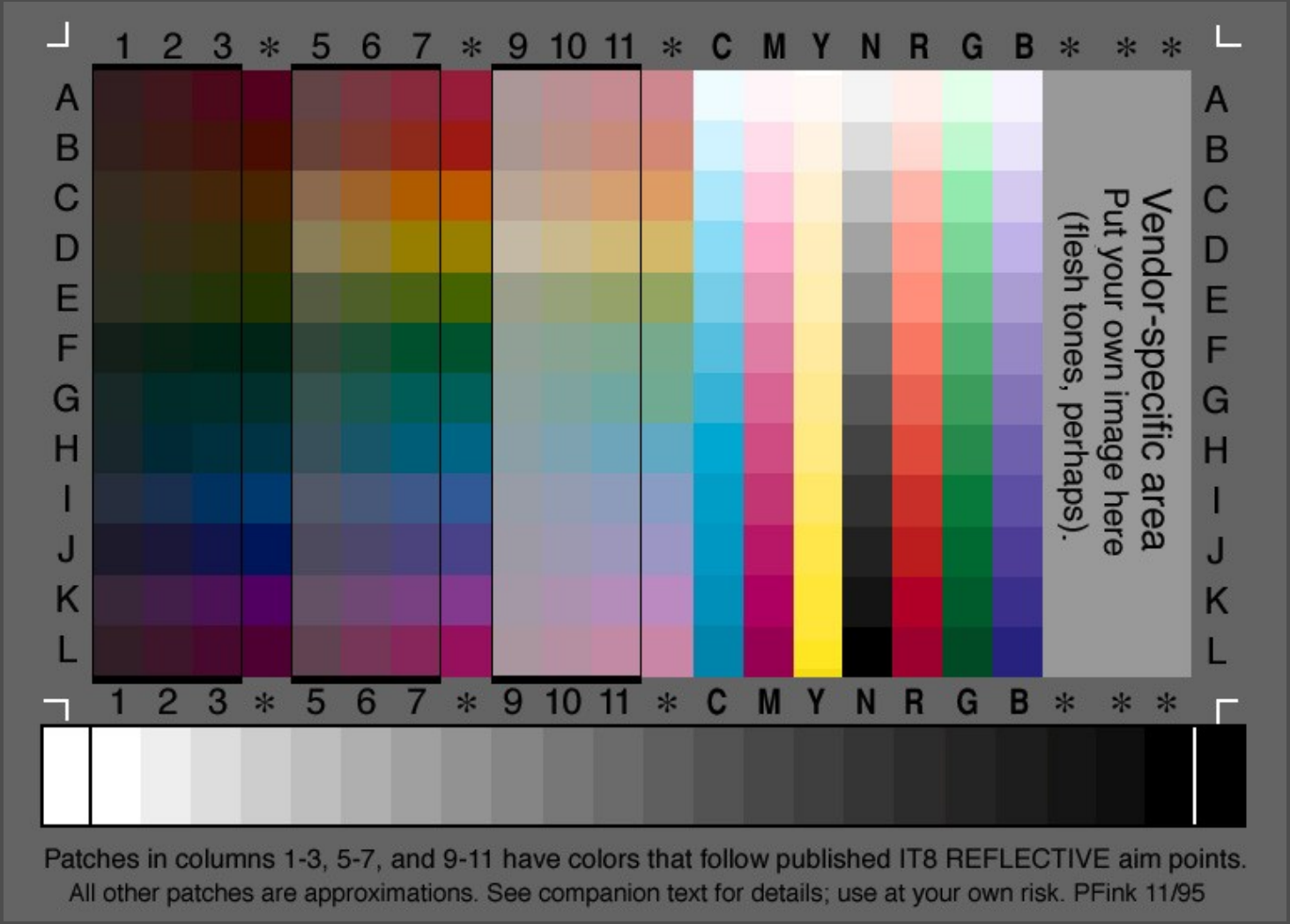
Images:
http://www.solar-imaging.com/digital-universal.html (top)
http://www.bearclover.net/epson-scanner/silverfast.html (right)

Automated Book Scanner

- 1200 pages per hour
- Must be supervised
- Used by Google and Internet Archive projects for books
- Not suitable for rare or fragile materials
- Does not create preservation grade images (JPEGs)

Image: http://www.kirtas-tech.com/uploads/images/APTFrontPage.jpg

Targets for scanning



http://www.imagequality.com/dtp/images/elec.it8.refl.jpg

Targets for scanning

- Many different sizes and types available
- Scanned with image
- Help to calibrate colour balance for scan
- Use scanning software to create white and black calibration with target for each scan
- Saved with archival digital master
- Derivatives are usually made with the target cropped out

Image: http://www-rcf.usc.edu/~gainer/impa/imaging/kodak_q_60_example.jpg

Image Processing

- De-skew
- De-speckle
- Reduce background
- Rotation
- Register

Warning

- *Only de-speckle and reduce background on images if absolutely necessary*
- *Processing often results in image quality loss*

OCR Notes and Recommendations

- Do not compress TIFFs, incompatible with some OCR programs
- Adjust brightness and contrast so that text is as dark as possible and background is as light as possible (using a copy of original)
- Skew in text will compromise OCR
- OCR tends to be less reliable with headings
- OCR tends to not be corrected

OCR Notes and Recommendations

- Require special ‘zoning’ algorithms for text in column format, ie. magazines
- Some OCR programs have a maximum pixel width of file
- OCR will not recognize handwritten script
- Special OCR programs are available for Gothic script ie. ABBYY FineReader7

Sample Imaging Requirements

Table 1: Digital Master Image Files— Recommended Imaging Requirements

Document Type	Resolution	Bit Depth	Enhancements Allowed	File Format	Compression
Printed Text[2]	600 dpi	bitonal	Sharpening, descreening, cropping, deskewing, and despeckling	TIFF 5 & 6	Lossless compression (e.g., ITU-G4)
Rare/damaged printed text	400 dpi	8-gray or 24-color	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Book Illustrations	400 dpi[3] 600 dpi with enhancement	8-gray or 24-color --- bitonal	Contrast stretching Minimal adjustments for tone and color --- Descreen/rescreen, sharpen	TIFF 5 & 6	Uncompressed or lossless compression (e.g., ITU-G4, LZW)
Manuscripts	300-500 dpi	8-gray or 24-color, if color present in the original	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Maps & other oversized items	300-400 dpi	8-gray or 24-color	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Graphic Art	400-600 dpi	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Photographic Prints	400 dpi	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)

http://www.library.cornell.edu/imls/image%20deposit%20guidelines.pdf

Sample Imaging Requirements cont'd

Works of art on paper	400 dpi	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Transparencies	4000-5000 on long end or 400 dpi on output > 8" x 10"	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	uncompressed or lossless compression; (e.g., LZW)
Microfilm	600 dpi blown back to original size ---- 300-400 dpi blown back to original size	Bitonal ---- 8-bit gray	Sharpening, descreening; cropping deskewing, and despeckling	TIFF 5 & 6	Uncompressed or lossless compression (e.g., ITU-G4, LZW)

[2] Although 600 dpi 1-bit is a defacto standard for printed text, a comparable or richer text file may be produced in grayscale at 400 dpi.

[3] Random or irregular halftones and those produces in color may be imaged at lower resolution, e.g., 300 because there is a lower incidence of moiré. It is recommended that high quality book illustrations, such as aquatints, collotypes, and engravings, especially those produced as separate plates, be retained for their artifactual value.

http://www.library.cornell.edu/imls/image%20deposit%20guidelines.pdf

Scanning Formats

Digital Master

- TIFF format
- Resolution of 600 dpi/ppi widely adopted for most materials
- Lower resolutions may be used to keep file sizes down for materials such as maps
- Bit depth depends on type of material

Web Delivery

- JPEG, JPEG 2000
- GIF only captures 256 colours

Digitization of Audio

- Introduction to various media types
- The Digitization Process
- Audio Formats

Image: http://www.addclasses.com/file.php/1/1earphone5-med.jpg

Wax or Celluloid Cylinders

- 1890s & 1900s, up to 5” diameter, 2-4 minutes playing time
- Source device is the phonograph
- See www.tinfoil.com for details of the digitization process

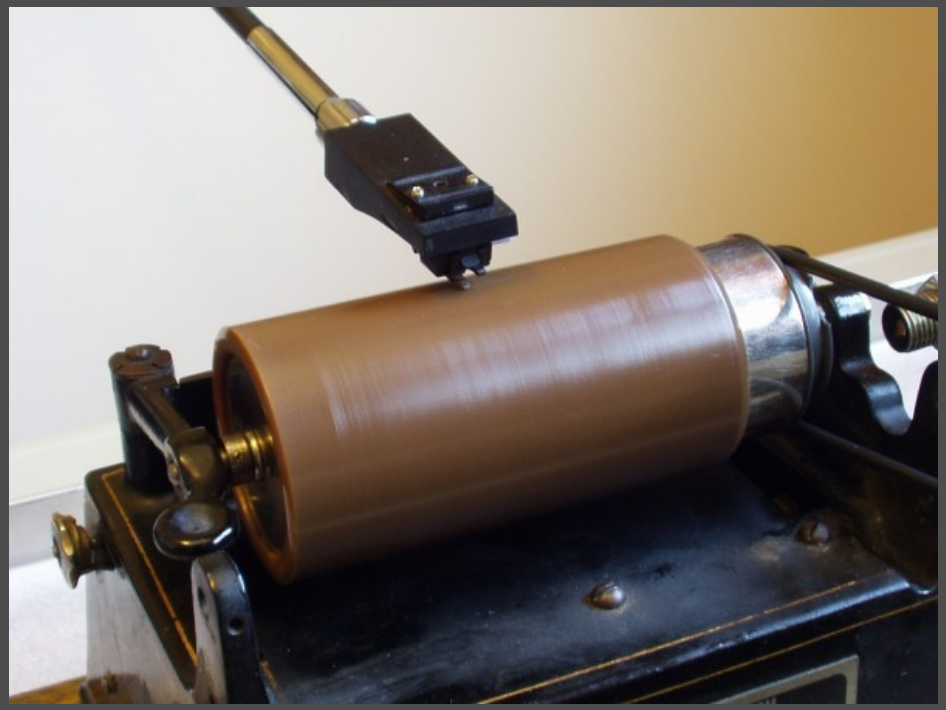


Image:
http://www.tinfoil.com/xferpics/600x/Xfer%20cyl%20front.jpg

Wire

- Magnetic coated wire drums or reels. Invented 1898. Widely used by the US military in WWII. Eclipsed by magnetic tape by the mid 1950s.
- Source device is the wire recorder.

Images:
http://www.videointerchange.com/wire_recorder1.htm
http://en.wikipedia.org/wiki/Image:Peirce_wire_recorder.jpg

78 rpm shellac resin disks

- 1898 to late 1950s
- 10” (25cm) and 12” (30 cm) most common sizes
- Source device is the wind up gramophone or Hi-Fi with a 78 rpm turntable and 78 rpm stylus
- Replace needle after each side or record played
- consistency of a fragile china plate - thick, heavy and highly breakable
- cracked and chipped easily

Images:
http://bcyesteryear.com/fulltext.php?article=67

48 rpm and 33 rpm vinyl discs

- 7” (20 cm) single and 12” long play (30 cm)
- Introduced in 1948
- Stereo recordings in 1958
- Source required is the Hi-Fi with 45 and 33 rpm turntable speeds

Images:
http://en.wikipedia.org/wiki/Image:Vynil_record.jpg
http://www.sonicperfectionists.com/Equipment.htm

Reel to Reel magnetic tape

- ½” to ¼” magnetic tape
- BASF and AEG developed 6.5 mm ferric tape and Magnetophone player in Germany from 1935
- Post-war development in USA by Ampex and 3M
- Stereo capability from 1949
- Requires Reel to Reel player for appropriate width of tape

Image:
http://www.analogstereo.com/images/r2r/b77_main.jpg

Compact Cassette

- Magnetic polyester tape
- Introduced by Philips in 1963
- Requires compact cassette player

Image:
http://content.answers.com/main/content/img/CDE/_AUDCASS.JPG

Cartridge

- ¼” magnetic tape
- Fidelipac (4-track, devised 1956, released 1962) cartridge system
- Lear (8-track, 1956) cartridge system
- 4 and 8 track cartridges are not compatible and require separate players
- Predominantly used for in-car audio

Image: http://www.videointerchange.com/audio_history.htm

Capture Devices

Choose your capture device

- Internal computer sound card
  - prone to electrostatic interference from computer circuitry
  - Often built from inferior quality components

Image: http://www.techexcess.net/images/products/other/sb0200_medium.jpg

Capture Devices

Choose your capture device…

- External analog to digital device
  - Provides superior results to sound cards

Image: http://www.synthman.com/midiman/117212.html

Connect to ADC

Cassette players and **hi-fi systems** are still available can be connected to an analog to digital converter for digitization…

Direct sound output to ADC

Wire recorders, **cartridge players** and **reel-to-reel** players often have an analogue signal-out connection or can be modified by sound engineers to produce a direct sound output…

Microphone to ADC

For **wax cylinders** and other **older formats**, an external microphone can record the sound which can then be digitized…

Recommendations for digital sound preservation

- Higher sampling rate preferred, eg. 96kHz
- 24-bit sample word-length preferred
- Linear PCM preferred over compressed
- Higher data rate (128 kbps) preferred
- AAC compression preferred over MP3
- Encoding in stereo preferred over surround sound (unless essential to creator's intent)

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Sampling Rate & Precision

- sampling rate = how many samples of sound are taken per second
  - at 96 kHz, sound is sampled 96,000 times per second
- precision is calculated in bits
  - the more bits a sample contains, the better the sound quality
  - 24 bit sample: 010011111100111001001101

Table of standard audio formats

Wrapper Formats
Advanced Authoring Format (AAF)
Advanced Systems Format (.asf)
Audio Interchange File Format (.aif; .aiff) – preservation standard
Audio/Video Interleaved (.avi)
Broadcast Wave Format (.bwf) – preservation standard
Jpeg 2000 (JP2)
MPEG-4
MPEG-7
MPEG-21
Material Exchange Format (MXF)
OGG format (.ogg)
Quicktime (.mov, .moov, qt)
Real Media (.rm)

File Formats	
Compressed	Uncompressed/Lossless Compression
Advanced Audio Coding (.aac; .m4a)	Compact Disc Audio (CDDA)
Digital Audio Compression (AC-3; Dolby Digital)	Linear Pulse Code Modulated Audio (LCPM) – preservation standard
MPEG-1 Layer-3 (.mp3)	Real Audio (.ra; .rm; .ram)
Real Audio (.ra; .rm; .ram)	Standard Musical Instrument Digital Interface (MIDI) File (.smf; .mid)
Windows Media Audio format (.wma)	Wave (.wav)
	Extensible Media Format (.xmf)

TABLE DATA FROM: http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

WAV vs BWF

- WAV files contain an info portion that is not governed by standards
- Broadcast Wave Format is a European standard created to append standardised metadata to the WAV audio file format
- BWF work on WAV players
- For more information on BWF: http://www.ebu.ch/en/technical/trev/trev_274-chalmers.pdf

Audio Preservation Standards

Sampling rate: 96 kHz

Precision: 24 bit

Format: broadcast wave format or AIFF

Encoding: LCPM

Notes:

- IASA (International Association of Sound and Audiovisual Archives) minimum recommendation for analogue originals is 48 kHz/24 bit
- DVD quality is 96 kHz/24 bit
- CD quality is 44.1 kHz/16 bit

http://www.iasa-web.org/IASA_TC03/TC03_English.pdf

Audio use and access copy

- Need expensive proprietary software to play preservation master copies (96 kHz/24 Bit files)
  - Create CD with 44.1kHz/16 Bit file in .wav or .bwf format
- Web Accessible Copy
  - MP3
  - RealAudio, Quick Time (for streaming)

Use and Access Copy

- Original remains untouched
  - “Imperfections” may be significant to historians
- Copies may be enhanced by filtering and noise reduction techniques
  - Remove hiss, clicks and pops
  - Adjust calibration and EQ curves to approximate signal characteristics of original

Digitizing Moving Images

- Introduction to various media types
- The Digitization Process
- Moving Image Standard Formats

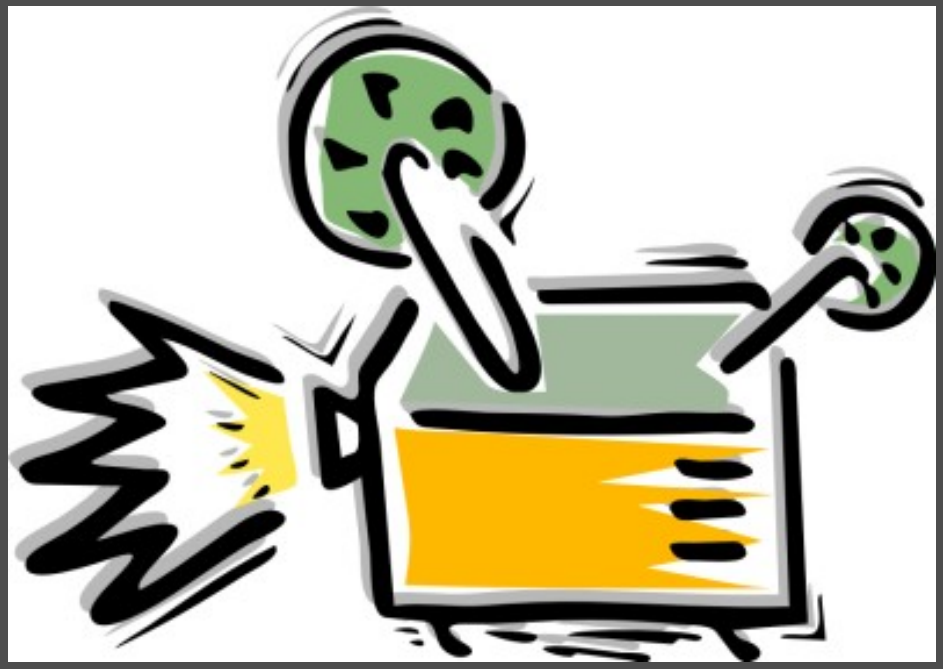
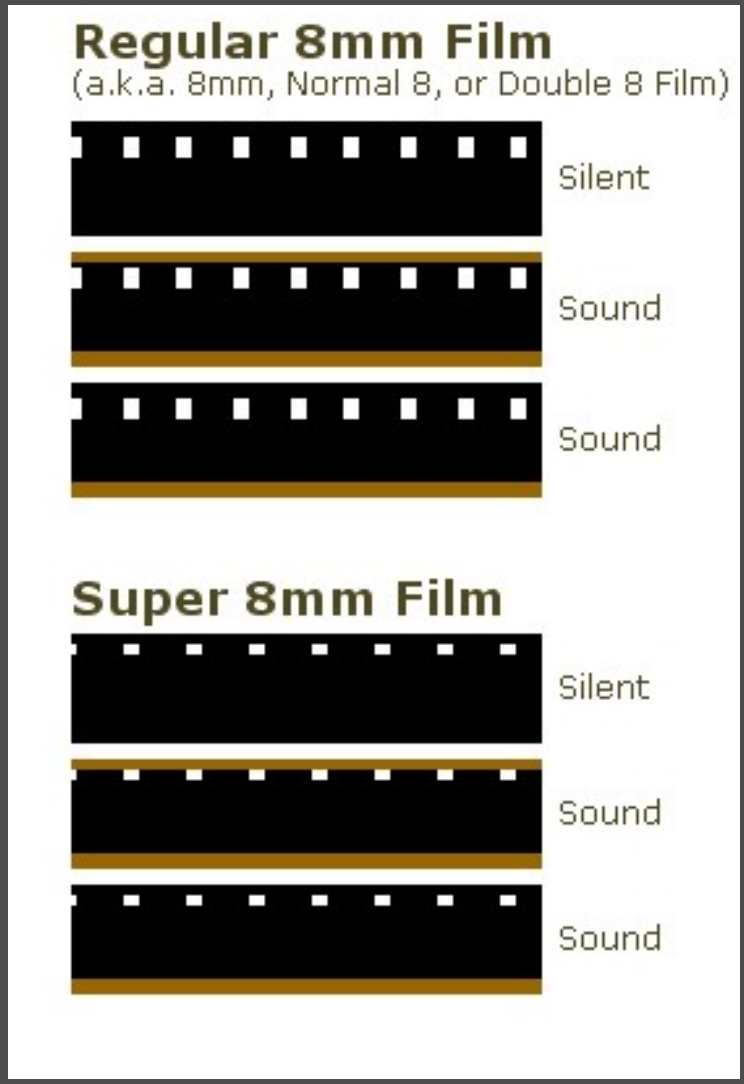


Image: www.wpclipart.com/camera/movie_projector.png

8mm & Super 8 Film

- Determining frame rate for digitization can be problematic
- Both are 8 mm wide and require their own projectors (dual duty available, but not recommended)
- The holes in Regular 8mm film are larger and almost square, whereas the holes in Super 8mm are elongated.



16 mm and 35 mm film

- Both 16 mm and 35 mm film are very common film formats
- 16 mm requires a 16 mm film projector
- 35 mm requires a 35 mm film projector

¼” and ½” Reel to Reel Video Tape

- ¼” can be confused with audio tape – 10” reels are audio only
- For ¼” tape, 7” and 5” reels can be video as well as audio
- Each require their own videotape recorders
- ½” video recorder maintenance and parts very difficult

http://www.randallareed.com/Reel_to_Reel_Tape.htm
http://www.fondation-langlois.org/html/e/media.php?NumObjet=10340&NumPage=471

1” and 2” Reel to Reel Video Tape

- 2” used in TV from late 1950’s to 1970’s
- 2” Reel to Reel tape player increasingly rare
- 1” requires its own 1” Reel to Reel tape player

Image:
http://www.lyrec.dk/images/tr532rcu.jpg

8mm Video Cassette

- 8mm video comes in 8mm and Hi-8
- 8mm has 240 horizontal line resolution, while Hi-8 is rated at 400
- Hi-8 player can play standard 8mm, but not vice versa

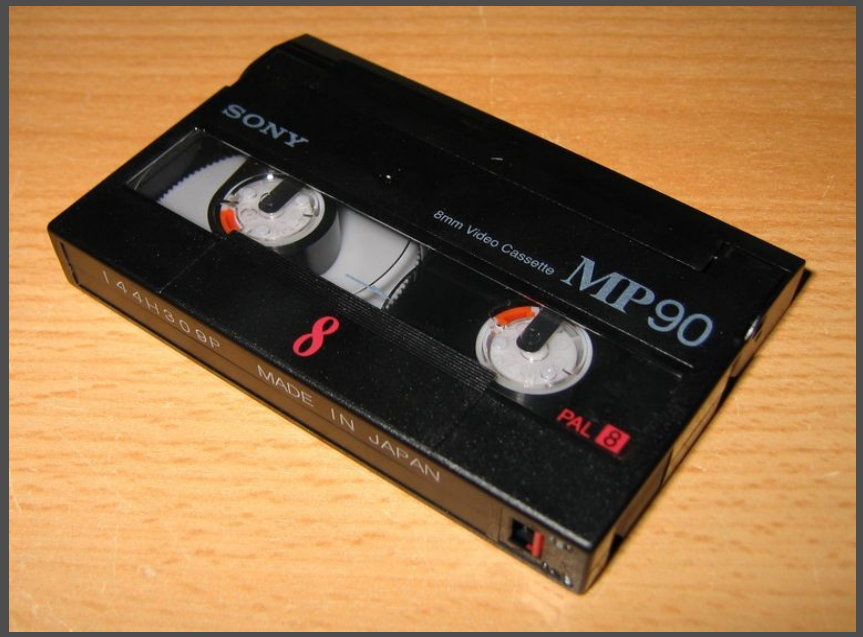
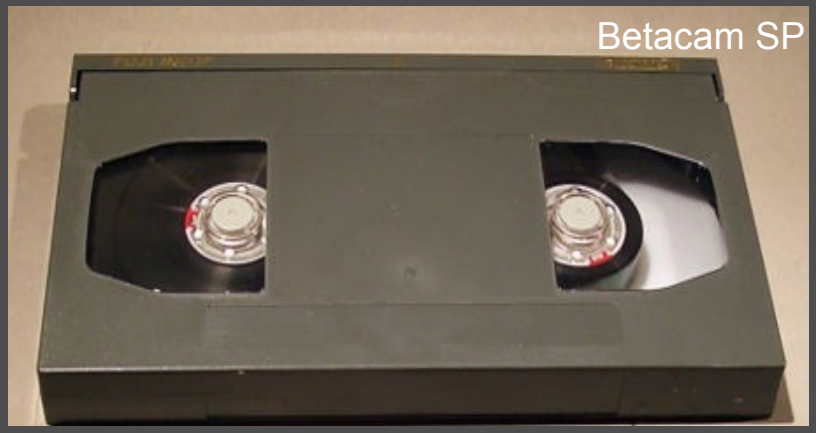


Image:
http://upload.wikimedia.org/wikipedia/commons/4/49/Video_8_cassette.jpg

¼” (12.5 mm) Video Tape Cassette

- Betacam SP, MII and S-VHS (also obsolete Video 2000 and Beta)
- S-VHS players will play VHS but not vice versa
- Betacam SP and MII require compatible players



Images:
http://en.wikipedia.org/wiki/Image:Beta_tape_sizes_2.jpg
www.russellvideo.com/images/Formats/mii.jpg

Digitizing via the Transfer Box Method

- Requires projector at one end and video camera at the other
- Rear image projection screen in the middle
- Film is projected into a box with a mirror and onto a rear image projection screen
- Video camera on the other side captures video from projection screen
- The video is then digitized
- Results in generational loss of quality

Transfer Box Method

- Project movie onto 3 inch screen on the side of the box
- Mirrors inside the box send the image to a port designed for a video camera

Image: http://www.brienposey.com/kb/film_to_dvd.asp

Digitizing via Multiplexer (Telecine)

http://www.nfsa.afc.gov.au/glossary.nsf/Pages/Telecine?OpenDocument

Multiplexer (Telecine)

- Requires projector, camera, lens and mirrors
- Image projected via lens and mirrors directly into camera
- Image recorded to a common video tape format

Multiplexer (Telecine)

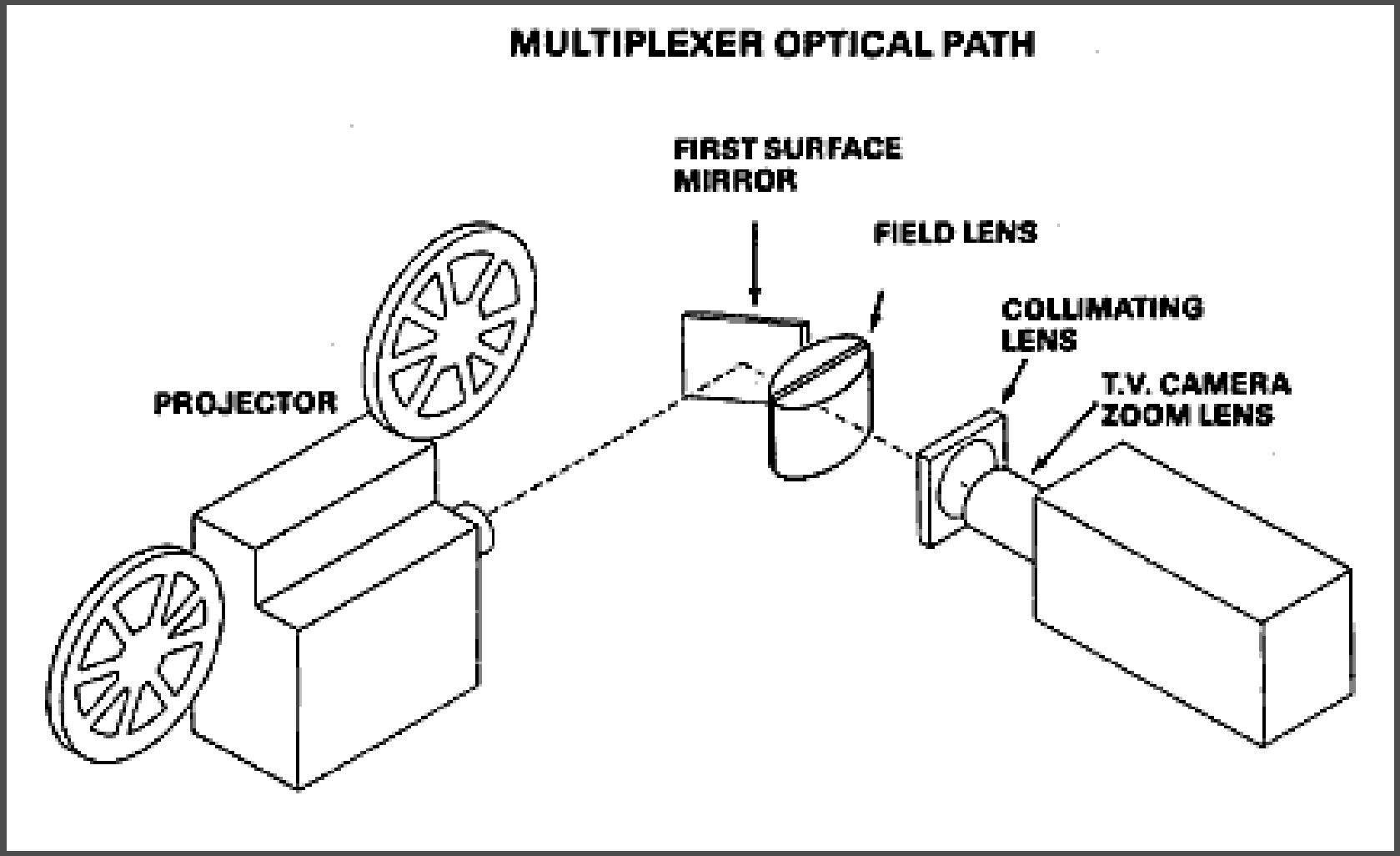
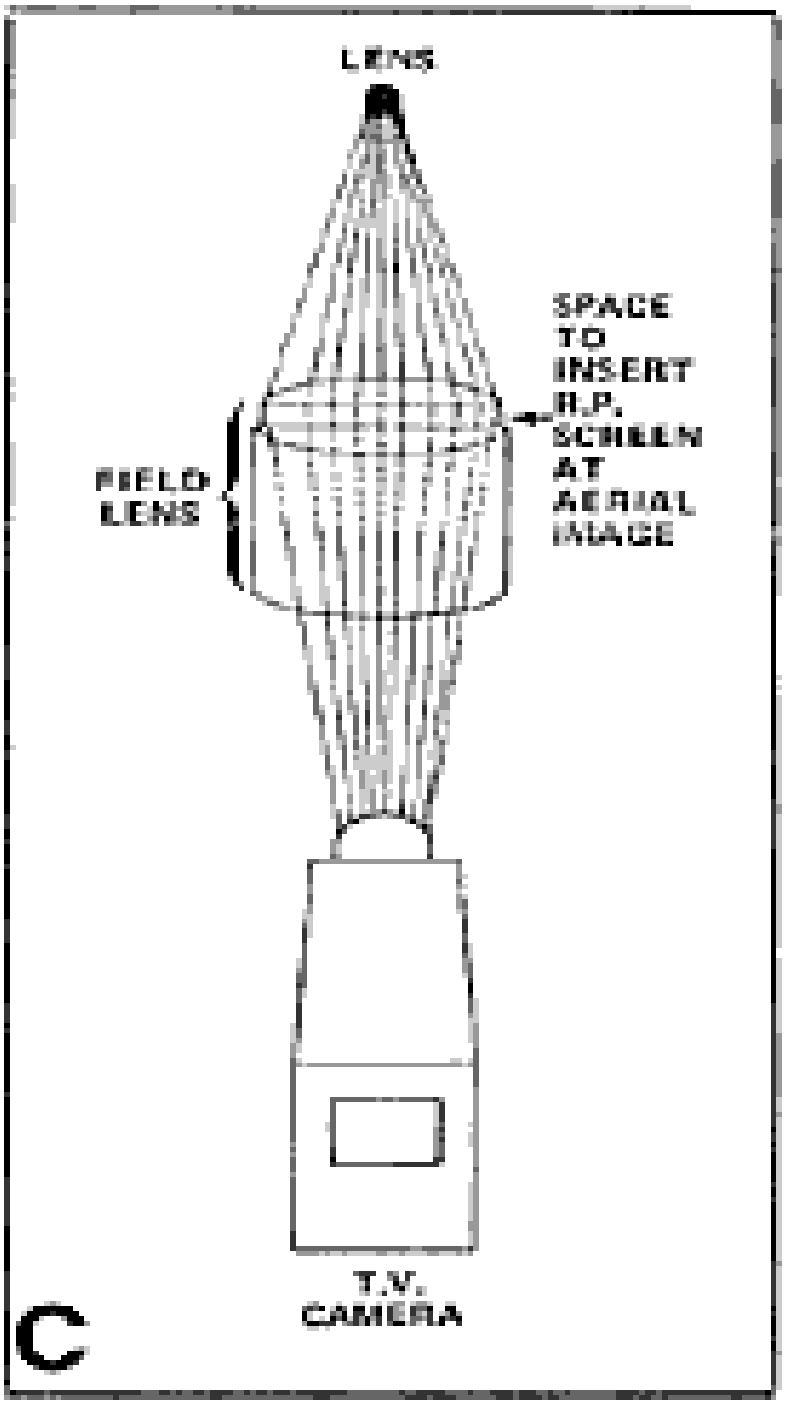


Image: http://www.toddvideo.com/transfers/images/multi.gif



Field Lens

- Directs light from all parts of the field lens into a small circle
- When the camera lens is placed at this circle, the entire field is illuminated
- Produces the highest quality image

Multiplexer (Telecine)

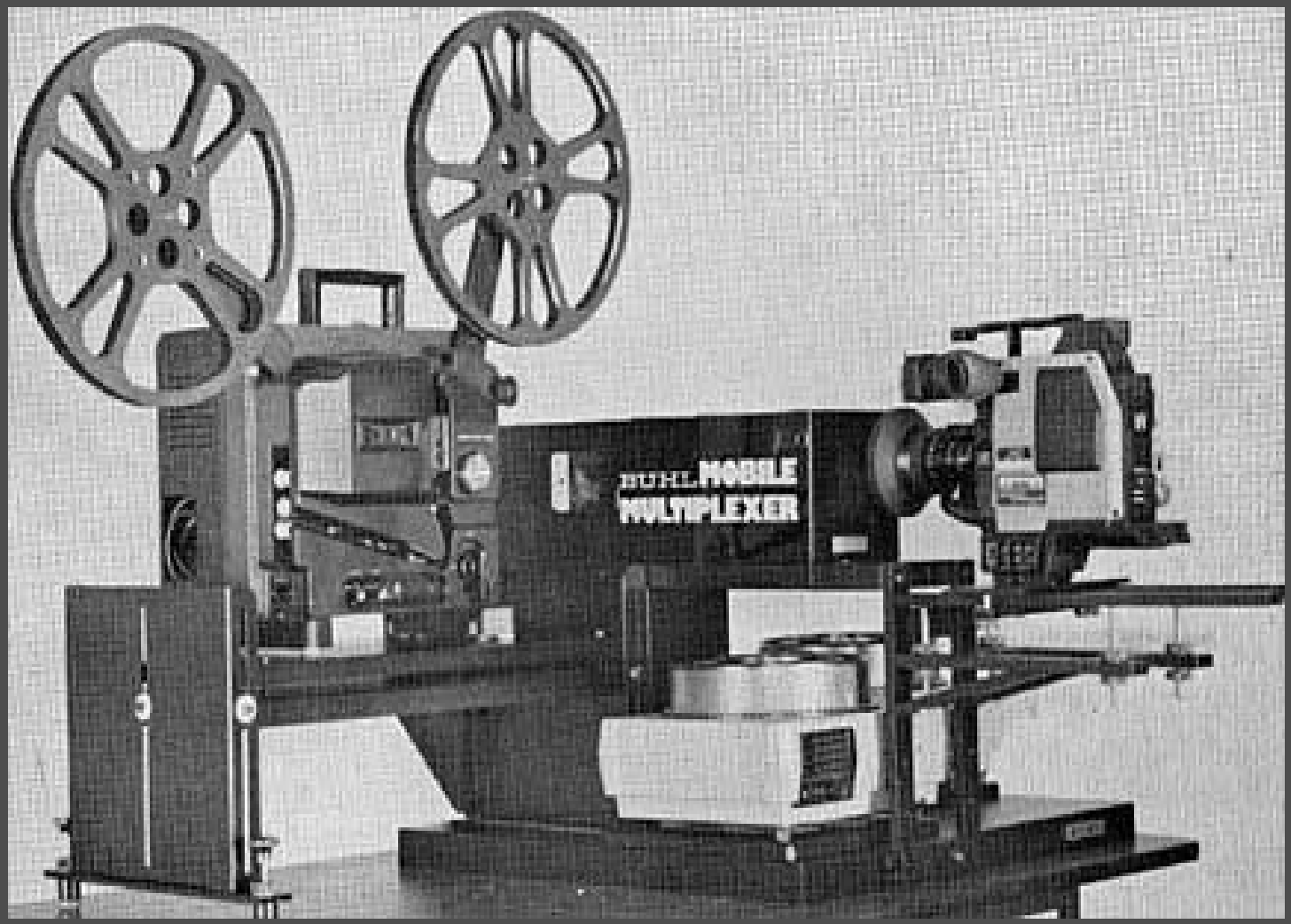


Image: http://www.toddvideo.com/transfers/film_chain.html

Multiplexer (Telecine)

- Better image quality than transfer box method
- Quality still suffers generational loss
- Generally used for film to videotape transfer or for television broadcasting of films
- Popular due to acceptable quality and affordability

http://www.nfsa.afc.gov.au/glossary.nsf/Pages/Telecine?OpenDocument

Telecine Transfer Price List

- Mini DV $370.00 / hour
- Digital Betacam $455.00 / hour
- DV Cam $370.00 / hour
- Betacam or BCSP $400.00 / hour
- 3/4" or 3/4 " SP $335.00 / hour
- VHS or SVHS $335.00 / hour

Monaco Digital Film Labs, San Francisco
http://www.monacosf.com/

Chain Film Scanner

- Digitize directly from 8, 16, or 35 mm
- Scans the film and digitizes at the scanner
- Passes the digital signal to the computer
- Digital conversion is done at the camera instead of computer
- Less opportunity for noise
- Extremely expensive to acquire hardware

Digital Film (Chain) Scanners

Images:
www.visinst.com/1635Photo2.gif (top)
http://uk.gizmodo.com/flashscan8.jpg (right)

Recommendations for digital master preservation

- Larger picture size preferred
- High definition content preferred (assuming picture size is equal or greater)
- Encodings that maintain frame integrity preferred over temporal compression
- Uncompressed or lossless compressed preferred over lossy compressed

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Recommendations for digital master preservation cont'd

- Higher bit rate (mb/s) preferred over lower for same lossy compression scheme
- Extended dynamic range (brightness) preferred over “normal” dynamic range (for scanned motion picture film and Digital Cinema)
- Stereo and monoaural sound preferred over surround sound (surround sound only necessary if essential to creator's intent)

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Common moving image wrapper and file formats

Wrapper Formats	File Formats	
	Compressed	**Uncompressed/Lossless compression**
Advanced Authoring Format (AAF)	MPEG-1	Digital Cinema Initiative Distribution Master (DCDM)
Advanced Systems Format (.asf)	MPEG-2	Motion JPEG (mj2, mjp2) – preferred by Library of Congress
Audio Interchange File Format (.aif; .aiff)	MPEG-4	Animation codec (Quicktime)
Audio/Video Interleaved (.avi)	Real Video (.ram, .rm)	SheerVideo
Jpeg 2000 (JP2) – preferred by Library of Congress	Windows Media Video format (.wmv)	
MPEG-4	DivX (.divx)	
MPEG-7	Digital Video formats (DV, DVCAM, DVCPRO	
MPEG-21		
Material Exchange Format (MXF)		
OGG format (.ogg)		
Quicktime (.mov, .moov, qt)		
Real Media (.rm)		

TABLE DATA FROM: http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Format Size Comparison

Format	1 min video	1 hour video
MPEG1	10.4 MB	624 MB
WMV	12.4 MB	744 MB
AVI	214 MB	12 000 MB (12 GB)

Source: http://linguistlist.emeld/school/classroom/video/archive.html

Format recommendations for digital masters

Digital moving images (general case):

- .mjp or .jp2 inside a JPEG2000 wrapper

Digital video converted from analog tapes:

- MPEG-2 at a minimum data rate of 1 Mb/s
- MPEG-4 at a minimum rate of 0.5Mb/s

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Format recommendations for digital masters cont'd

High quality video (professional videotape):

- JPEG2000 uncompressed

Commercial movies:

- DCDM

Digital broadcase television streams:

- Inconclusive, industry is in a state of flux

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Format recommendations for digital masters cont'd

- Note: Other preferred wrapper formats are AVI, QuickTime or WMV as long as audio and video bitstreams are uncompressed or use loseless compression

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Popular use and access formats

Streaming:

- Real Media Video
- Windows Media Video
- Quicktime
- MPEG-4 (multimedia)

Video CD:

- MPEG-1

DVD:

- MPEG-4

http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

Metadata

- Why create metadata?
- Types of metadata
- Systems & Schemas

```
<titlStmt>
<titl>Survey for Rural Economies, 1998</titl>
<IDNo>7856</IDNo>
</titlStmt>
<fundAg>Countryside Agency</fundAg>
<fundAg>Department of the Environment, Transport and the
Regions</fundAg>
<copyright> Social Research Centre</copyright>
<grantNo>3289460</grantNo>
<distStmt>
<distrbtr abbr="UKDA" affiliation="University of Essex, Wivenhoe Park,
Colchester, Essex, England, CO4 3SQ">UK Data Archive</distrbtr>
<depositr>National Centre</depositr>
<depDate date="2000-05-08"/>
<distDate date="2000-06-08"/>
</distStmt>
<keyword>ACCESS TO COUNTRYSIDE</keyword>
<keyword>AGRICULTURAL DEVELOPMENT</keyword>
<keyword>AGRICULTURAL PRODUCTION</keyword>
<keyword>AIR POLLUTION</keyword>
<keyword>COUNTRYSIDE</keyword>
<keyword>ENVIRONMENTAL CONSERVATION</keyword>
<nation>Great Britain national</nation>
<geogCover>GREAT BRITAIN</geogCover>
<geogUnit>(A)Wards; (B)Standard Regions; (C)Postcode Sectors;
(D)Parliamentary Constituencies; (E)Local Authority Districts; (F)Counties;
(G)Scottish Regional Councils</geogUnit>
```

Why do we need metadata?

- Digital identification
  - Used to differentiate one object from another
  - Used to identify sets of data

  Examples:
    - ISBN
    - file name
    - URL
    - persistent identifiers, e.g., PURL (Persistent URL); DOI (Digital Object Identifier)

Why do we need metadata?

- Resource discovery
  - Allowing resources to be found by relevant criteria
  - Identifying resources
  - Bringing similar resources together
  - Distinguishing dissimilar resources
- Organizing e-resources
  - Organizing links to resources based on audience or topic
  - Building these pages dynamically from metadata stored in databases

Why do we need metadata?

- Facilitating interoperability
  - Federated searching across collections
  - Allows for sharing and transfer of data
  - How?
    - Use defined metadata schemas
    - Share transfer protocols and crosswalks
    - Example: OAI protocol for Metadata harvesting

Why do we need metadata?

- Archiving and preservation
  - Digital information is fragile and can be corrupted or altered
  - It may become unusable as storage technologies change
  - Metadata is key to ensuring that resources will survive and continue to be accessible into the future:
    - track lineage
    - detail its physical characteristics
    - document its behavior in order to emulate it in future technologies

Types of Metadata

- Descriptive
  - Describes a resource for purposes such as discovery and identification
  - Can include elements such as title, abstract, author, and keywords

http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Types of Metadata

- Structural
  - Indicates how compound objects are put together
  - Example:
    - Show relationships between digital object and page number of book
    - The first scanned page of a book is rarely marked as page #1 of the book itself

http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Types of Metadata

Administrative

- Provides information to help manage a resource such as:
  - when and how it was created, file type and other technical information, and who can access it
- Subsets of administrative data:
  - Terms and Conditions
    - deals with intellectual property rights
  - Preservation Metadata
    - contains information needed to archive and preserve a resource

http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Dublin Core

- Comes in a simple (15 elements) and a larger qualified set
- All elements are optional and repeatable
- Minimum standard for describing digital objects

Simple Dublin Core Set:

Title	Source	Contributor
Creator	Language	Date
Subject	Relation	Type
Description	Coverage	Format
Publisher	Rights	Identifier

Dublin Core Example

Title=”Metadata Demystified”
Creator=”Brand, Amy”
Creator=”Daly, Frank”
Creator=”Meyers, Barbara”
Subject=”metadata”
Description=”Presents an overview of metadata conventions in publishing.”
Publisher=”NISO Press”
Publisher=”The Sheridan Press”
Date=”2003-07"
Type=”Text”
Format=”application/pdf”
Identifier=”http://www.niso.org/standards/resources/Metadata_Demystified.pdf”
Language=”en”

METS

- Metadata Exchange and Transmission Standard
- Created for describing complex digital library objects
- Encoded in XML format
- Components of a METS File:
  - METS Header
  - Descriptive Metadata – MODS, MARC, MARCXML etc.
  - Administrative Metadata – provenance and copyright
  - Structural Map – hierarchy and links to digital objects
  - Structural Links
  - Behavior

MARC, MARCXML, MODS

- MARC - Machine Readable Cataloguing Record
- Can easily be transform MARC21 into MARCXML with software programs
- MODS is a subset of MARCXML elements
  - Can easily transform MARCXML into MODS
  - MODS is embedded in METS records for item level descriptive metadata

MODS Example

```
<mods>
    <titleInfo>
        <title>Metadata demystified</title>
    </titleInfo>
    <name type="personal">
        <namePart type="family">Brand</namePart>
        <namePart type="given">Amy</namePart>
        <role>
            <roleTerm authority="marcrelator" type="text">author</roleTerm>
        </role>
    </name>
    <typeOfResource>text</typeOfResource>
    <originInfo>
        <dateIssued>2003</dateIssued>
        <place>
            <placeTerm type="text">Bethesda, MD</placeTerm>
        </place>
        <publisher>NISO Press</publisher>
    </originInfo>
    <identifier type="isbn">1-880124-59-9</identifier>
</mods>
```

Extension Schemas for AV Material

- METS allows the embedding of extension schemas to further describe digital objects
- MIX – Metadata for Images in XML - is used to provide additional technical information about images
- There is a version of MIX for Audio Visual Materials
- AudioMD, VideoMD, ImageMD – technical extension schemas

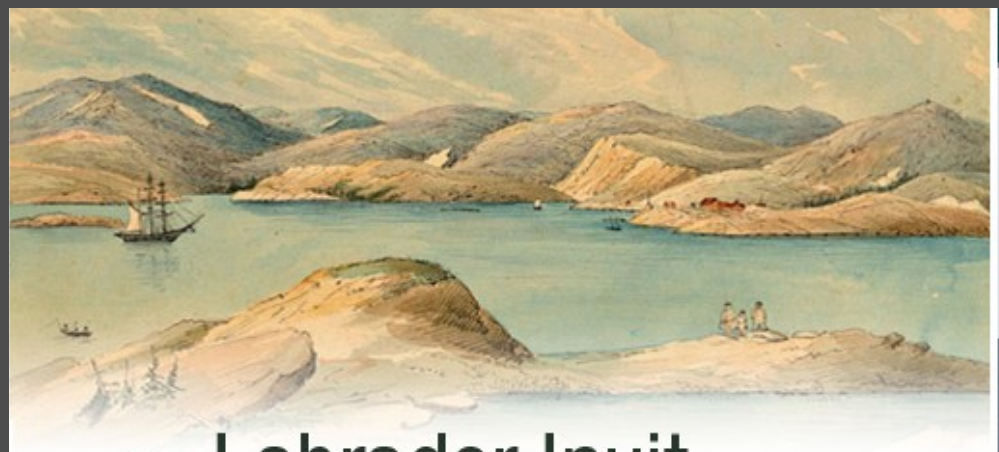
For more information: http://www.loc.gov/rr/mopic/avprot/metsmenu.html

Recommended mimium metadata set for archiving moving image and sound resources:

Combines elements from Dublin Core, PREMIS, AudioMD, VideoMD, TVAnytime, MPEG-7

See pages 82 through 89:
http://www.jisc.ac.uk/media/documents/programmes/preservation/moving_images_and_sound_archiving_study1.pdf

The Labrador Inuit Through Moravian Eyes



The Labrador Inuit Through Moravian Eyes

This site provides information on the 250-year relationship between Moravian missionaries and the Inuit of Labrador. This interaction led to the establishment of settlements for a formerly nomadic people, their conversion to Christianity and exposure to aspects of North American culture. The information has been gathered from a variety of sources that shed light upon this unique adventure. Read more >>

Interactive Timeline

Interactive Map

Educational Toolkit

Français Site Info Help Contact

BROWSE QUICK SEARCH ADVANCED SEARCH

Browse the Collection

Author

View highlights from the Collection

START/STOP

Black and white photograph of two Inuit children, circa 1927.

About the Project

- Canada Culture Online grant for 400,000+
- Collaboration between University of Toronto Libraries, Memorial University Libraries and the Bibliothèque de l'Université Laval
- Memorial University of Newfoundland provided source materials and description
- U of T responsible for digitization and interface
- Université Laval responsible for French translation

Types of Media

- Video
- Audio
- Photographs
- Drawings/Paintings
- Plans/Maps
- Manuscripts
- Published Texts

Additional Metadata for Browsing

Digitization Standards

- Photographs, Manuscripts, Plans/Maps, Drawings/Paintings captured as 600 dpi 24 bit TIFFs, Published Texts as 600 dpi 1 bit TIFFs.
  - Delivered online as 3 sizes of JPEG
    - Thumbnail: 75 pixels across
    - Small: 500 pixels across
    - Large: 775 pixels across (to neatly fit inside borders of website)

Zooming Capabilities

- For Plans/Maps, we wanted to be able to show more detail
- The Zoomify program was used
- Zoomify takes an image and creates nested directories of tiles, only retrieving the tiles of interest
- The result is slick and smooth zooming
- This works like the zooming feature of JPEG 2000

Scotiabank Information Commons

New Media Suites

- For use by UofT community
- Must complete free certification course
- Course teaches you how to use the equipment (about 2-3 h)
- Have facilities for digitizing audio and video, scanners available as well
- Rent rooms for 3 hour time blocks

http://www.utoronto.ca/ic/newmedia/equipment.html

New Media Suites

A/V Equipment in the Suites:

- Tascam 102 MK2 audio cassette recorder
- Pioneer DV-525 DVD player
- Panasonic 5710 SVHS video tape recorder
- JVC BR-DV3000 professional DV recorder

Software in the Suites:

- Avid Xpress Pro
- Adobe Photoshop
- Sorenson Squeeze
- Ulead DVD MovieFactory

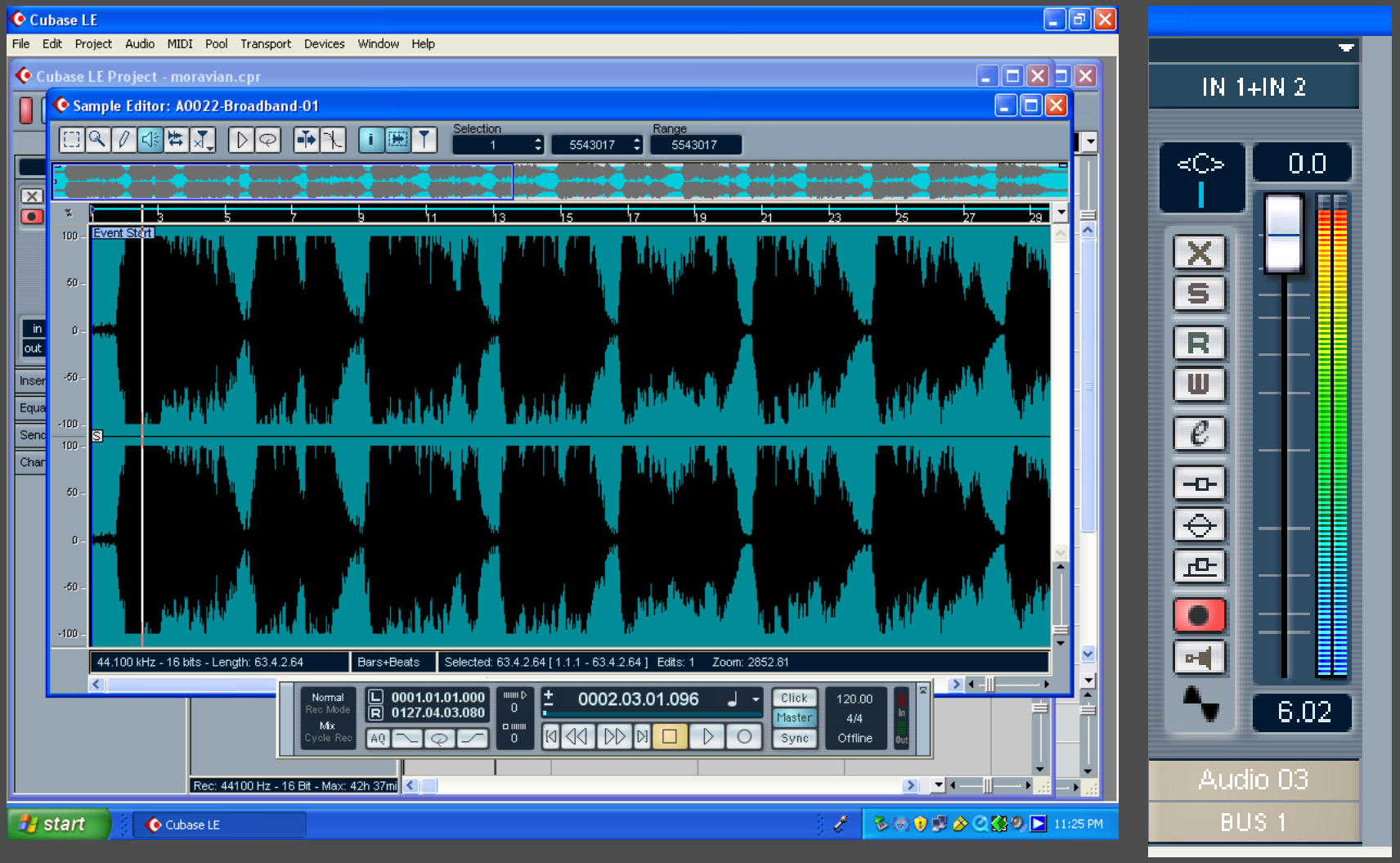
Audio Items

- Digitized from audio cassettes at Scotiabank Information Commons in New Media Suites
- Digitized at 44.1 kHz, 16 Bit
- Used Avid Express Pro to capture and edit
  - Tape Player > ADC > Computer
- Pro Tools was used to boost gain where capture was not adequate

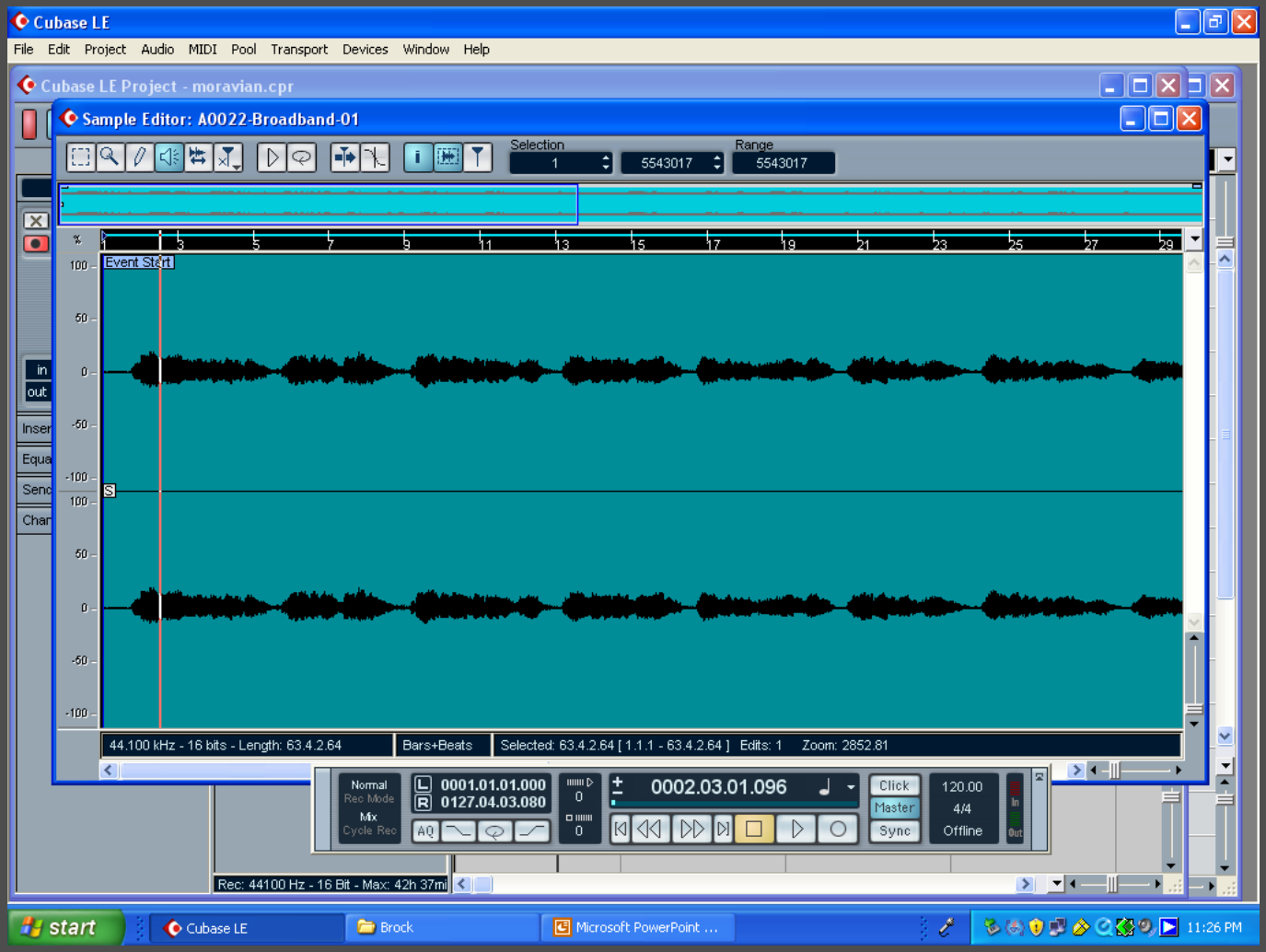
Basic Sound Recording Principles

- Must control input levels so that captured sound is not:
  - Too loud, otherwise clipping will occur
  - Too soft, otherwise you will have to process it to be louder
- We captured files too quietly, had to go back and boost levels

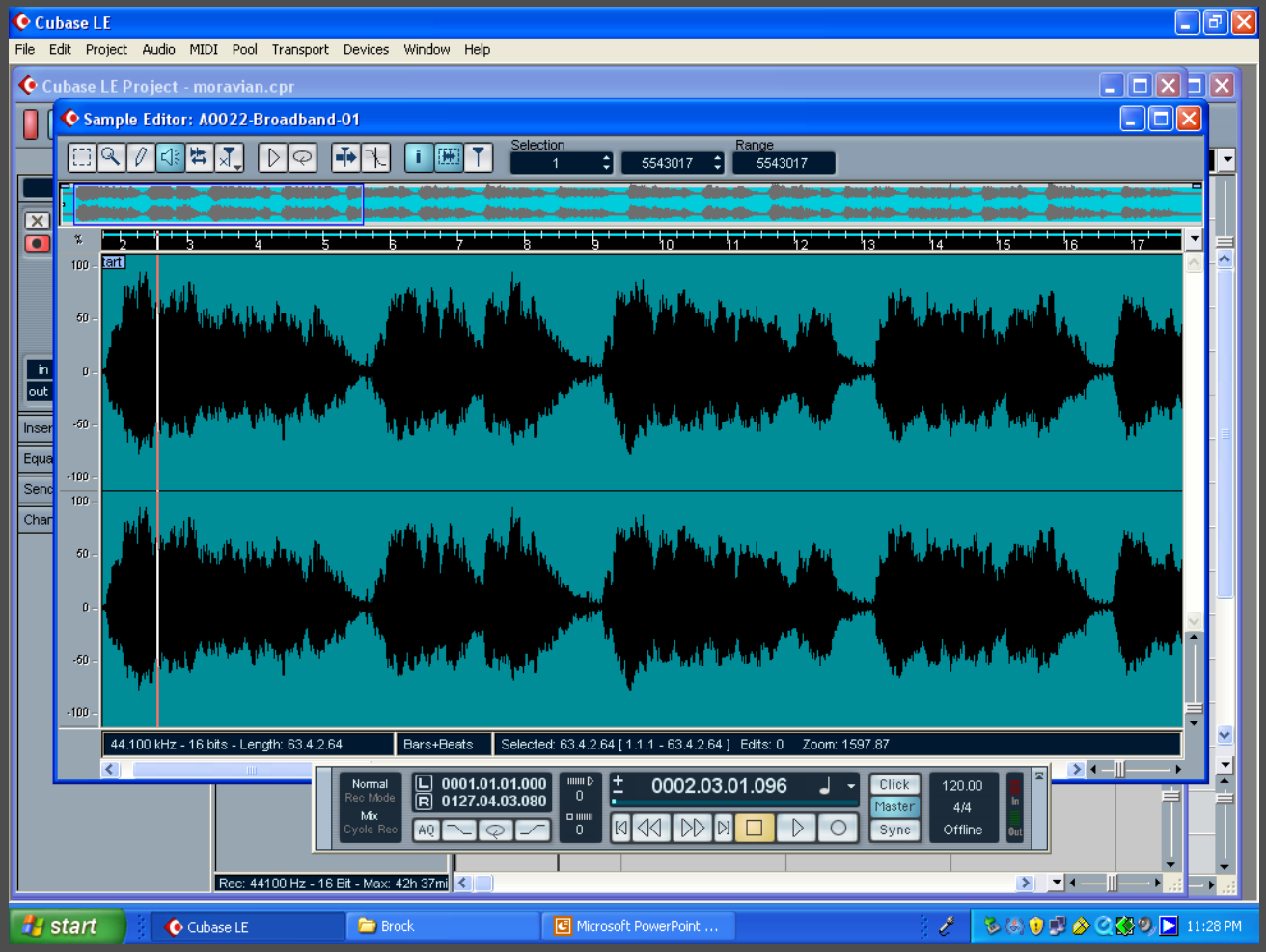
Example of a clipped wave

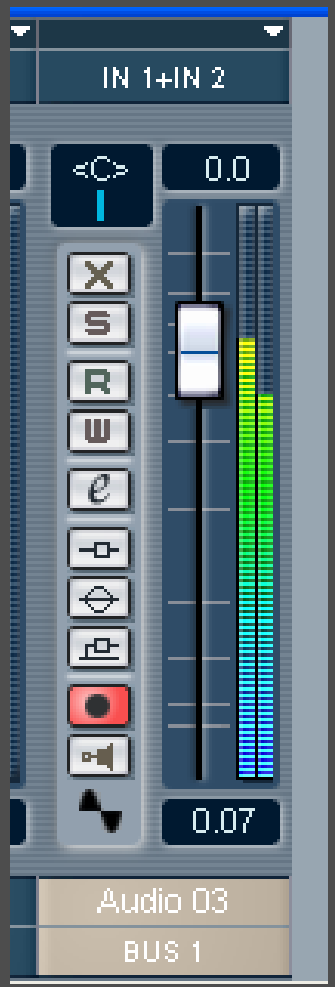


Example of a wave that needs boosting



Acceptable audio wave





Vendors

When money, time, equipment or expertise is short…

- Outsource to a trusted, recommended vendor
- This is often the most affordable and desirable option, especially for older formats
- Talk to your network of colleagues for recommendations
- Try to find a local vendor if possible

Video Items

- Super 8 mm reels with sound
- Digitized to DVD (MPEG2) by trusted, local vendor
- Vendor recommended by Thomas Fisher Rare Book Library
- Digitization cost about $150 / reel
- Transferred from DVD into Avid environment for editing

The Real Work Begins

To ensure that capture was successful:

- Listened to each entire tape
- Watched each DVD

- Selected excepts from digitized audio and video for web
- Used Sorensen Squeeze to create derivative formats
- Digital masters saved in MPEG2 format

Web Delivery Formats

Video

- Quick Time and Windows Media
  - 256Kbps (56 Kbps was too blurry)
  - 512Kbps
  - 1Mbps

Audio

- Quick Time Audio and Windows Media Audio
  - 56Kbps
  - Broadband (128 Kbps)

Questions?