Managing Digitization Projects

17 June 2016

Anna St.Onge

Archivist, Digital Projects & Outreach

Clara Thomas Archives & Special Collections, York University

What We Will Cover Today

- Why Bother?
- Managing a Digitization Project
- Digitization of Text and Images
- AV Preservation: The Paradigm Shift
- Digitization of Audio
- Digitizing Moving Images
- Asking for help
- Metadata
- Opportunities for Collaboration
- Islandora at York University

What I Will Gloss Over, Ignore, Avoid

- Text scanning for the most part
- The Obstructionist Tendencies of Copyright and DRM
- Emulation
- Nostalgia that keeps us chained to obsolete and deteriorating mediums

Why bother?

Why Digitize?

- Obsolescence of source devices (for audio and moving images)
- Media has a limited life span
- Digitization limits the use and handling of originals
- 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

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 (enhances preservation, searching, etc.)

Questions to ask yourself

- What purpose would these digital objects serve?
- If I digitize these analog materials, can I / my institution commit to their preservation in perpetuity (i.e. migration forward, fixity checks, secure repository etc.)
- The records may belong with us, but is our institution best suited to manage the infrastructure necessary to ensure their preservation, access and use?

What is preservation?

- One simple definition for AV materials:
- "Preservation is the totality of the steps necessary to ensure the permanent accessibility – forever - of an audiovisual document with the maximum integrity."
- http://www.bbcarchive.org.uk/pmwiki/pmwiki.php?n=Main.WhatIsPreservation

"Access drives preservation."

Brewster Kahle
Founder, Digital Librarian, Internet Archives, United States
27 September 2012, UNESCO Memory of the World Conference

"universal access to all knowledge can be one of our greatest achievements" … we have the money, we're just not doing it well…

Brewster Kahle,
27 September 2012, UNESCO Memory of the World Conference

Digitization challenges

- Multiple formats to choose from
  - Formats constantly evolving
- Can't match quality to that of the source
- Preservation challenges
  - Analog version must be kept, but the digital copy is the preservation copy
- Loss of knowledge & expertise
  - Technicians/engineers/manufacturers of older tech are retiring, dying, closing
- Expensive
  - Equipment, storage, metadata, staff time, long term preservation

Digitization challenges

- Storage
  - CD quality audio = 520 MB per hour
  - DVD-quality video = 13 GB per hour
  - Broadcast quality video = 75 GB per hour (ITU-R BT.601)
- Time
  - Many formats must be digitized in real time with supervision; hard to automate
- Technical limitations
  - Compression algorithms still evolving
  - High bandwidth required for transfer
    - At preservation standards, it takes 5x the duration of an audio file to transfer over T1 network

Managing a Digitization Project

Slides 13-22 adapted from: Learning Lessons from Other Digitisation Projects, http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/learning-lessons-from-other-digitisation-projects/ . See JISC Digital Media's Infokits here: http://www.jiscdigitalmedia.ac.uk/infokits/ .

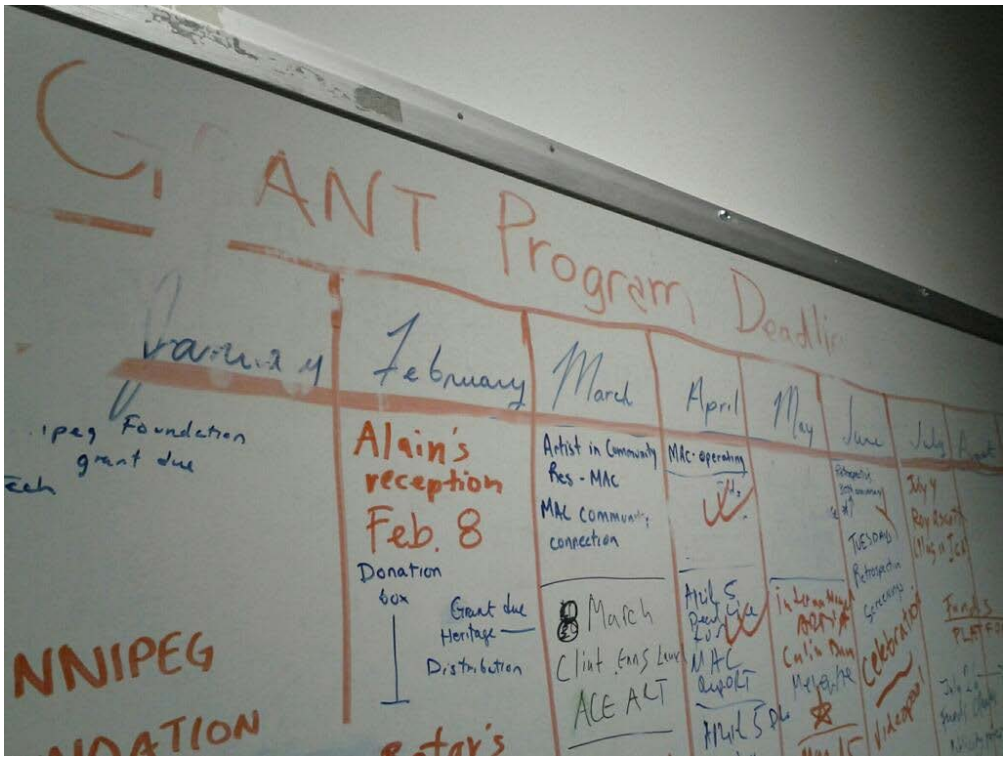
Know your collection

- What do you want to scan?
- Will you be selecting specific items, if so, what's your criteria?
  - Condition of originals
  - Rights status (copyright, donor permissions, moral rights etc.)
  - Items in high demand
  - Subject knowledge of selector
  - Preservation
- Need estimated numbers

Project Planning

- What are your aims and needs?
- What do your users need? Try to integrate their feedback at all stages.
- What does administration want? Does this mesh with their aims?
- Distinguish between these needs, prioritize them, and create a plan.

Managing Projects and Grants



Schedule for Winnipeg's Video Pool activities based on grant deadlines, goals and reporting requirements, 2013. Photo: Anna St.Onge.

- Many projects are grant based and will require realistic scheduling, scaling and goals
- How can you fit your objectives into a grant's set of deliverables?
- Managing grants, budgets, reporting obligations can be just as time consuming as the digitization itself

Making the Case for Digitization

- Digital Preservation Business Case Toolkit http://wiki.dpconline.org/index.php?title=Digital_Preservation_Business_Case_Toolkit (Includes CC-ANC 3.0 illustrations)
- Endangered Archives Programme (British Library)
- Maja Kominko, ed., *From Dust To Digital: Ten Years of the Endangered Archives Programme*, 2015.
- NARA, *Strategy for Digitizing Archival Materials for Public Access, 2015-2024*, December 2014, accessed February 25, 2015, http://www.archives.gov/digitization/strategy.html
- Canadian Council of Archives Preservation Committee, *Digitization and Archives*, 2002, accessed February 25, 2015, www.cdncouncil**archives**.ca/**digitization**_en.pdf.

http://wiki.dpconline.org/index.php?title=File:Format_cassette.png and http://wiki.dpconline.org/images/5/59/Risks_octopus_web.png

Minimize duplication of effort

- Check to see whether the items you wish to digitize have already been digitized
- Places to check:
  - WorldCat
    - Special instructions to search the Registry of Digital Masters here: http://www.oclc.org/services/collection/default.htm
  - Internet Archive
  - Early Canadiana Online
  - DPLA
  - ArchivesCanada.ca and Library Archives Canada
  - Google Books, Hathi Trust, Seren-dip-o-matic, etc.

Digitization is a team effort

- Ensure you have the required support (departments, administration) and resources
- Collection knowledge is just as important as technical knowledge
- Plan for staff recruitment, training and attrition
- Plan for time required to complete all steps.
- Budget appropriately (see calculator here)
- Appropriately compensate your workers for their labour/knowledge
- Keep channels of communication open
  - Problem solving has to be timely

Digital capture

- Establish file naming conventions and directory structure
- Conduct a small pilot study to test your workflow and settings, establish iterative process that will allow adjustment to your activities
- Identify special handling requirements for materials and put in place appropriate guidelines and training
- Document the workflow and encourage team feedback
- Establish quality assurance measures

Metadata

- Establish how and where metadata will be captured
- Metadata is time consuming
  - Determine quality benchmarks
  - Can be an iterative process
- Determine how you want your collection to be searched and displayed
- Adopt controlled vocabularies
- ***When adapting formal metadata standards, ensure that you are not sacrificing interoperability***

Outsourcing

- Get a trusted referral if possible
- You need to know technical details and standards to ensure that you get what you need
- Don’t forget about metadata
- Clarify what the price covers and how it breaks down
- Your agreement should include timelines and penalty clauses, quality assurance standards and procedures, and reporting requirements

Quality Assurance (QA)

- Establish clear criteria and well-documented quality assurance procedures
- Be realistic
- Allow adequate time to undertake QA and any corrective work
- Enable your users to alert you to any errors and provide you with evaluative feedback
- Evaluate as you go along and integrate what you learn into your project

Collection delivery

- Think about your interface at the beginning to ensure adequate digital and metadata capture
- Note that your content/metadata will need to outlive any current management system
- Involve your users in interface design and testing
- Address issues of usability and accessibility
- Support standards for dissemination and interoperability

Preservation and Maintenance

- Investigate yourself, or talk to your IT support people about file storage and software upgrades
- Put in place a strategy for preservation, identifying how often your collection should be backed-up, checked, and migrated
- Fully document the project to ensure understanding of all aspects: digitization and metadata standards, copyright status, system architecture
- Make this a pillar of institutional decision making
  - i.e. review every two years, review when you're migrating/replacing old servers etc.
  - Make it a central activity, not an afterthought.

Digitization of Text and Images

Digitization hardware

The Digitization Process

Common Image Formats

Scanners are format specific

- Take inventory of what needs to be scanned, and the composition of your collections
- Don't forget to assess whether it makes more sense to contract out!
- Are there consortial solutions? Can you work with other institutions to work out gift-in-kind/labour exchange
- Are things in a uniform format? Can you create batch workflows?

- Choose the scanner that best suits the largest volume of your materials:
  - Maps
  - Plans
  - Manuscripts
  - Plain Text
  - Drawings
  - Paintings
  - Photographs
  - Negatives
  - Microfilm
  - Transparencies
  - Slides
  - Charts & graphs

Digitization hardware

- Flatbed
  - Smaller maps, drawings, plain text, etc.
- Digital Camera
  - Maps, plans, rare books (book cradle)
- Microfilm scanner
- Slide/Negative scanner
  - Higher resolution capture, specialized cartridges to hold different sizes of film
- Photo scanner
  - Higher resolution capture

Automated Book Scanner

- Hundreds of pages per hour
- Must be supervised
- Used for large book scanning projects
- Not suitable for rare or fragile materials
- May not create preservation grade images

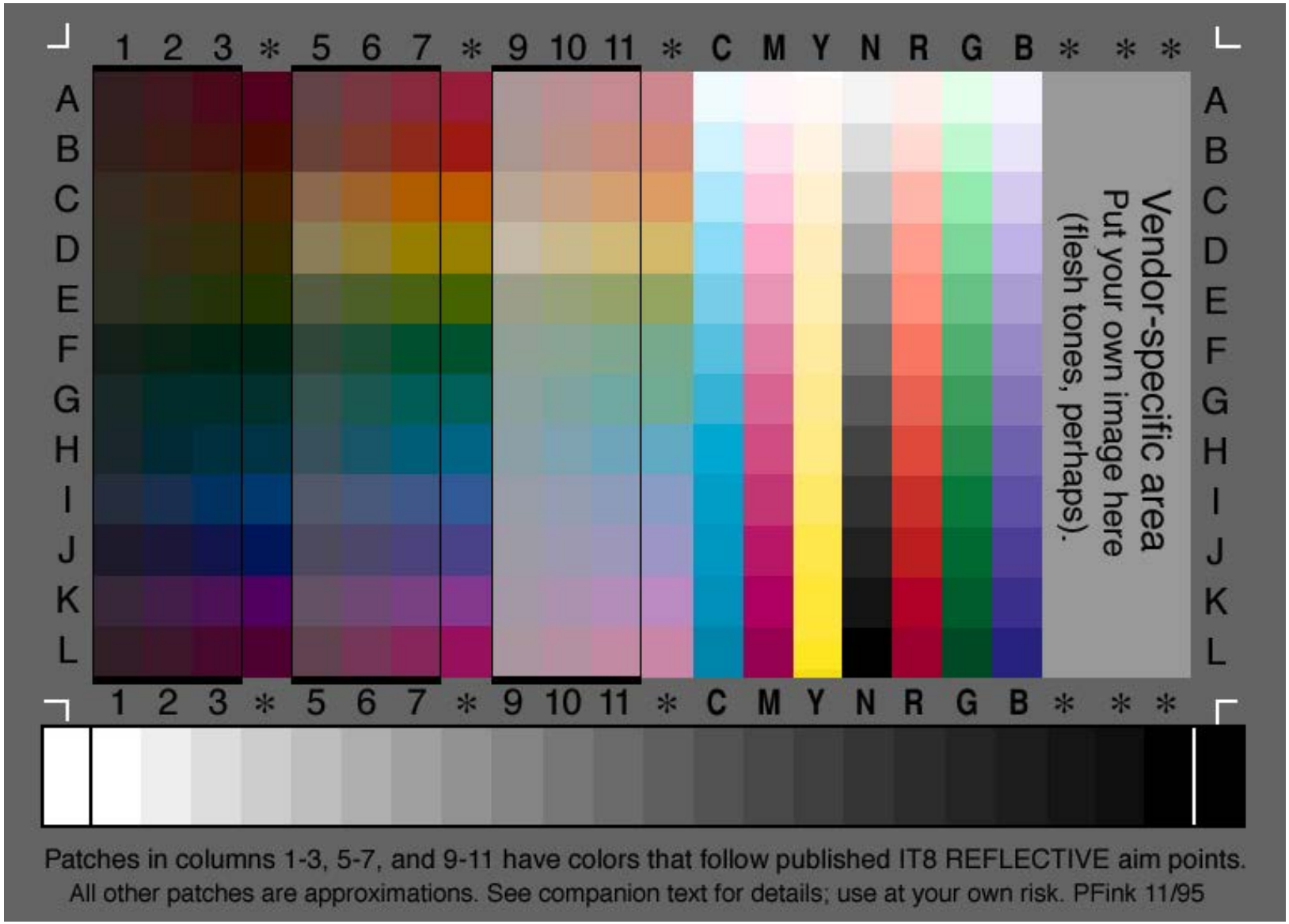
Image http://www.kirtas-tech.com/uploads/images/APTFrontPage.jpg

Targets for scanning

- Many different sizes and types available
- Scanned with image or used to set colour profile
- Help to calibrate colour balance for scan
- Saved with archival digital master
- Derivatives are usually made with the target cropped out

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

Targets for scanning



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

Optical Character Recognition Notes and Recommendations

- Do not compress TIFFs, this can lead to incompatibilities
- 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

LAC internal imaging standards

	Master Copy	Access Copy
Books	300 dpi 8-bit colour (sometimes bitonal or greyscale) JPG	300 dpi bitonal (with some elements in greyscale or colour) PDF with OCR
Serials	300 dpi 24-bit Adobe RGB (1998) TIFF	72 dpi colour or greyscale JPG or PDF with OCR
Government publications and reports	300 dpi 24-bit Adobe RGB (1998) TIFF	72 dpi colour or greyscale JPG or PDF with OCR (when possible)
Newspapers (microfilm)	150 dpi 8-bit greyscale TIFF	150 dpi greyscale JPG or PDF with OCR
Theses (microfiche)	150 dpi 8-bit greyscale TIFF	150 dpi greyscale JPG or PDF with OCR

http://www.collectionscanada.gc.ca/digital-initiatives/012018-1200-e.html

LAC internal imaging standards

	Master Copy	Access Copy
Maps	6000 to 8000 pixels (longside) 24-bit Adobe RGB (1998) TIFF	600-1000 pixels (longside) 72 dpi JPG, PDF or JPG 2000
Documentary art, photography, portraiture	Varies depending on format and scanner, ranges 6000 to 8000 pixels (longside), between 300 dpi to 4000 dpi 24-bit Adobe RGB (1998) TIFF	600-1000 pixels (longside) 72 dpi or 150 dpi colour or greyscale JPG

http://www.collectionscanada.gc.ca/digital-initiatives/012018-1200-e.html

York publically posts our in-house standards here:
http://www.library.yorku.ca/web/collections/digitalscholarship/digitization-best-practices/

Scanning Formats

- Digital Master
  - TIFF format
  - Resolution of 300 or 600 dpi/ppi widely adopted
  - 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 (scalable)
  - GIF only captures 256 colours

Don’t boutique when you can streamline

- Uniform format? Standard set of fields?
- Create a workflow that streamlines the process, batch processes and allows for presets.
- Example: Smithsonian’s Herbarium Digitization Project. See video here.
  - They now digitize in one day what took them 3 weeks to do previously. Works out to be about $1.50 per scan. Generating 1 TB of content per day.

AV Preservation: The Paradigm Shift

AV media: the landscape

- AV documents are the documents of modernity. Without them, we would have no understanding of contemporary history, no adequate record of orally transmitted cultures, or linguistic and cultural diversity, no intercultural discourse.
- 80% of historical film heritage is already lost.
- 80% of the world's AV is outside proper archival custody.

Deitrich Schuller, UNESCO Memory of the World Conference, September 28, 2012

AV media: the landscape

- AV documents are machine readable formats. Need playback devices, and technical developments lead to a shorter life cycle of formats. The industry has also swiftly withdrawn from spare part and service support.
- While text is chemically and physically highly stable, AV documents are generally lower in chemical stability.
- We need digitization for access but also for long term preservation!

Preservation of AV documents

Change of preservation paradigm:

- Carrier preservation is ultimately in vain!
- We should concentrate on content preservation and fully extract content

Archival Principles

- Adequate digital resolution is imperative
- Signal extraction must be complete and unmodified
- No signal restoration for archival files (unethical)
- No compression allowed! Compression deletes data irreversibly - it is not archival
- Must use openly defined file formats

Digitization of Audio

- Key Documents
- The Digitization Process
- Audio Formats

Key documents

- International Association of Sound and Audiovisual Archives (Technical Committee), The safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy, ed. by Dietrich Schüller. Version 3, 2005 (= Standards, Recommended Practices and Strategies, IASA-TC 03). International Association of Sound and Audiovisual Archives. www.iasa-web.org/tc03/ethics-principles-preservation-strategy
- IASA Technical Committee, Guidelines on the Production and Preservation of Digital Audio Objects, ed. by Kevin Bradley. Second edition 2009. (= Standards, Recommended Practices and Strategies, IASA-TC 04). www.iasa-web.org/tc04/audio-preservation

IASA TC-03 (key points)

Digitization projects must seek a compromise between urgency factors:

- need for access
- carrier decay
- format and equipment obsolescence
- retarding factors
- present high cost, lack of budget
- technical improvements

IASA TC-03 (key points)

Safeguarding the information:

- By preservation of the carrier and equipment
- By copying the information (can only be achieved by subsequent lossless copying for one information carrier to the next)
- Unmodified extraction of the entire document, the intended signal and then unintended and undesired artefacts. RETRIEVE AS IS
- Keep whole recording from beginning to the end

IASA TC-03 (key points)

Safeguarding the information (cont'd)

- Aesthetic improvements must only be made in a second process on the basis of an objective archival master
- Signal extraction from analogue carriers determines quality of document for the rest of its life
- Quality expectation is increasing
- KEEP THE ORIGINALS
- Transfer is expensive, unlikely to do it again
- Digital long term archiving means permanent migration every 3-5 years

IASA TC-04 (key points)

Audiovisual documents contain primary and secondary information

- Primary: content, signal, "essence"
- Secondary: associated materials and information, metadata, technical representation
- All information is part of the document and must be presented.
- Note that some technical information is lost in conventional transfer procedures.

IASA TC-04 (key points)

Key principles:

- Standalone A/D converter
- Encoding: Linear PCM
- Minimum 48 KHz 24 bit
- BWF
- No data reduction (compression) for analogue or linear digital originals - keep it as simple as possible

IASA TC-04 (key points)

Signal extraction from originals:

- Use original for transfer
- Cleaning, physical/chemical restoration
- Choice of playback equipment
- Choice of playback parameters (speed, equalization)
- Correction for misaligned recording equipment
- Removal of storage artefacts
  - Example: rewind tape a few times to remove magnetism

Audio Digitization Setup

- Playback device
  - With audio out (ideal)
  - OR professional microphone (only as last resort)
- Analog to digital converter
  - This is your stand-alone capture device
- Computer with audio digitization software
- Headphones

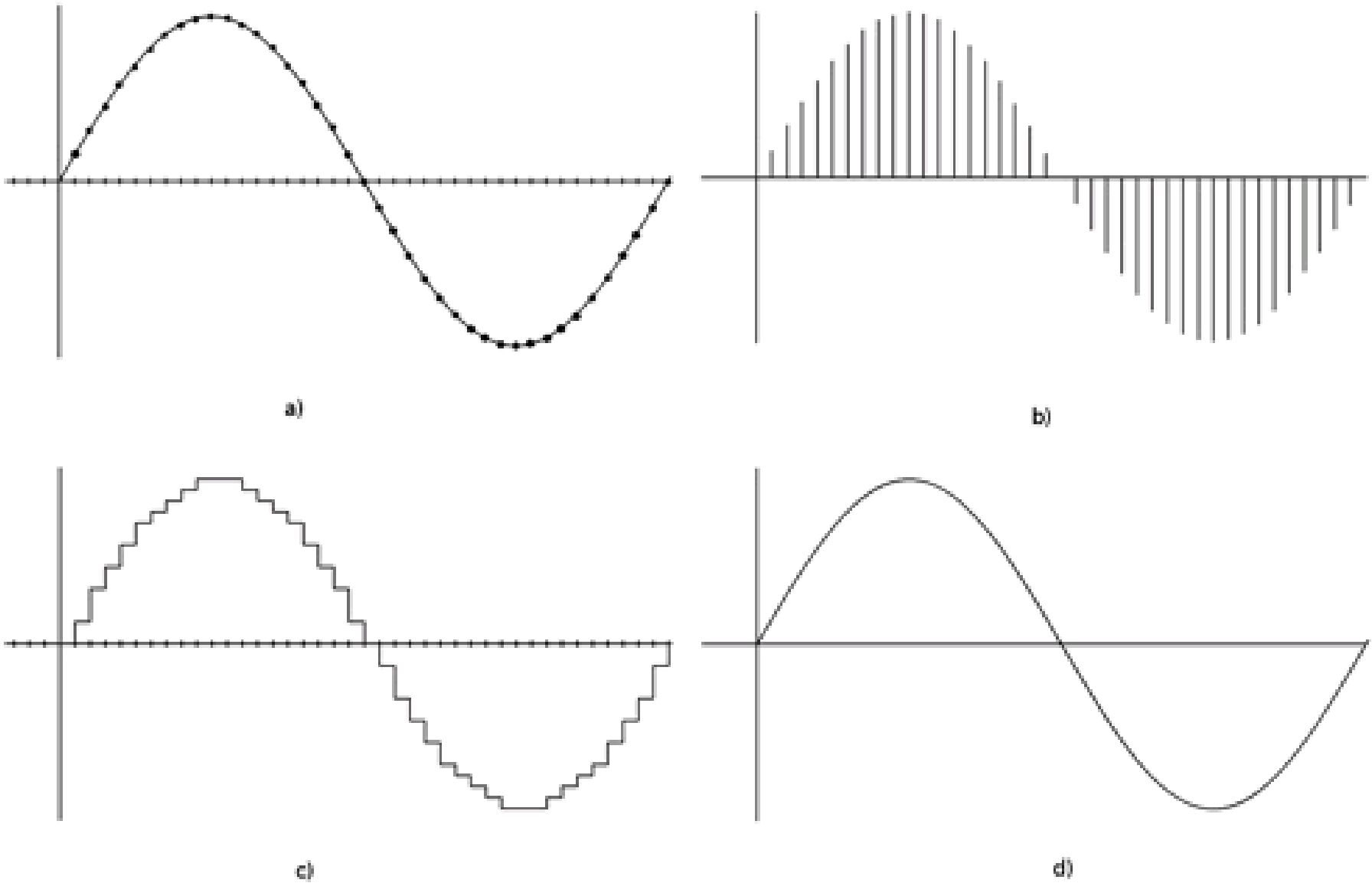
Analogue to digital converters

- IASA TC-04 has strict guidelines
  - **Internal computer sound card not adequate!**
  - Stand-alone A/D must be connected via fireware
  - USD, AES/EBU or S/PDIF interface.
  - Specifications must be in accordance with the Audio Engineering Society standards
    - AES 17-1998 (r2004)
    - IEC 61506-3

http://www.iasa-web.org/tc04/key-digital-principles

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: 010011111110011100100 1101



http://www.jiscdigitalmedia.ac.uk/audio/advice/an-introduction-to-digital-audio/

Audio Preservation Standards

- Sampling rate: 96 kHz ideal, but 48kHz acceptable
- Precision: 24 bit
- Encoding: Linear Pulse Code Modulation (LPCM) (uncompressed)
- Wrapper: Broadcast Wave Format (.bwf) or AIFF
- Stereo encoding 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

More on standards…

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

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

Use and access copy

- Need 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
- BE CLEAR ABOUT YOUR INTENTION: PRESERVATION OR RESTORATION?

LAC Standards for Audio

	Master Copy	Access Copy
Music recordings	96 kHz 24-bit BWF	MP3
Spoken word recordings	96 kHz 24-bit BWF	MP3

http://www.collectionscanada.gc.ca/digital-initiatives/012018-1200-e.html

Digitizing Moving Images

Thou shalt not compress video!

5 categories of formats

Moving Image Standard Formats

Association of Moving Image Archivists
Annual Conference 2011, Austin, TX

"What Should We Do Today: Toward an Interim-Master for the Preservation of Digital Audiovisual Materials"

These slides illustrate why moving image compression is incompatible with preservation. See pages 42-63 of the presentation for the slides and a detailed transcript.
http://www.georgeblood.com/Resources.html#AMIA

Let's have a look at the examples

The Muppet Show - The best of Swedish Chef

http://www.youtube.com/watch?v=9OfsABOGw3c

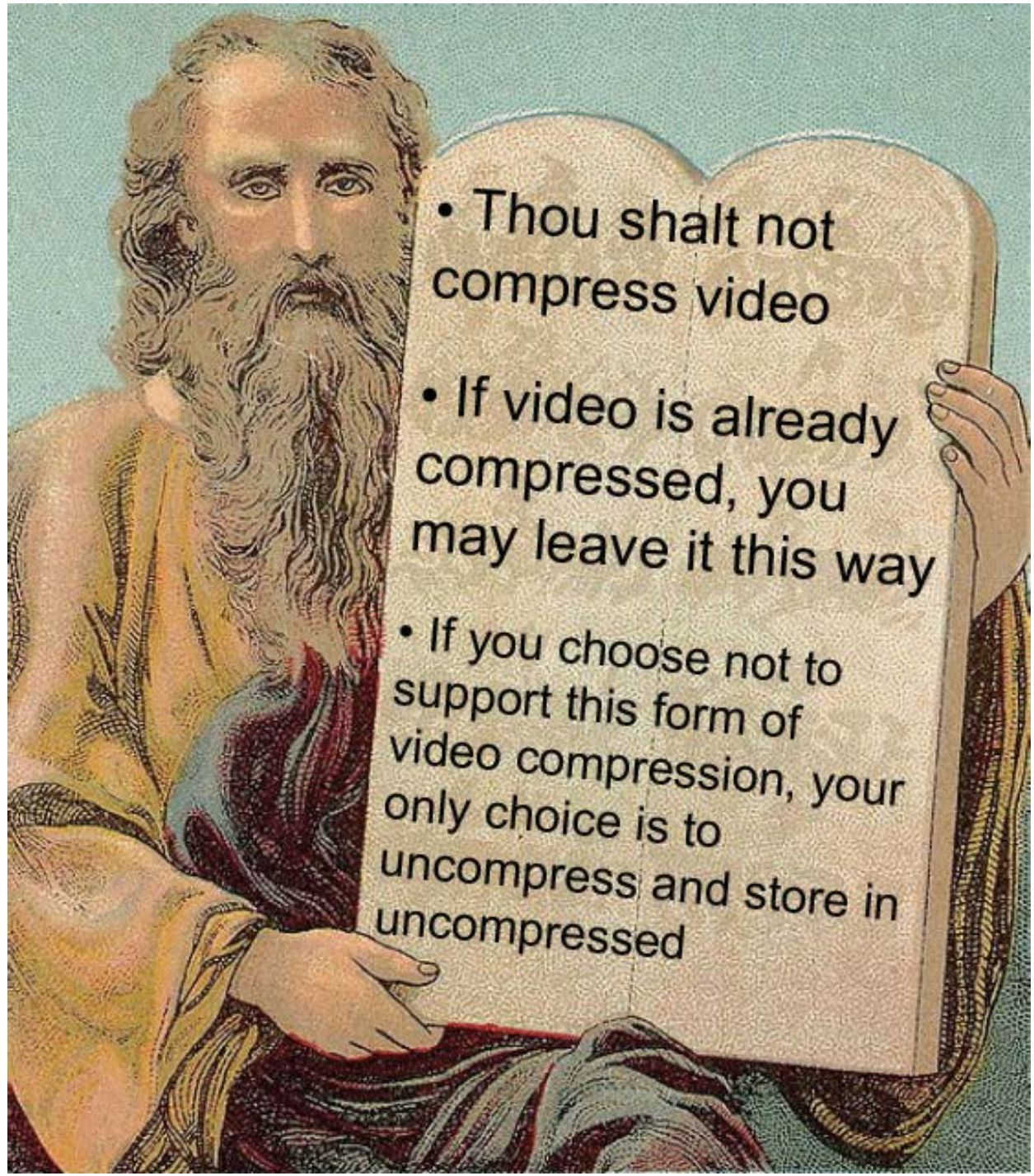
Generation Loss

http://vimeo.com/3750507

I Am Sitting in a Video Room

By ontologist (Patrick Liddell)

http://youtu.be/icruGcSsPpo


• Thou shalt not compress video
• If video is already compressed, you may leave it this way
• If you choose not to support this form of video compression, your only choice is to uncompress and store in uncompressed

Preservation Self-Assessment Program (University of Illionis)

https://psap.library.illinois.edu/format-id-guide#audiovisual

IASA TC-06

This will eventually be the companion document for video. But it is still in development…

Library of Congress Specification for Moving Images

- Files are encoded with lossless JPEG 2000 and wrapped in XMF
- Detailed specifications are still under development for JPEG 2000/MXF
- The JPEG 2000/MXF specification is called AS-AP (MXF Application Specification for Archiving and Preservation)
  - http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html
- Common desktops have limited ability to access either JPEG 2000 or MXF

In the interim…

Refining Conversion Contract Specifications: Determining Suitable Digital Video Formats for Medium-term Storage.

George Blood, principal author*

http://dl.dropbox.com/u/81562888/George%20Blood%20LIbrary%20of%20Congress%20IntrmMastVidFormatRecs_20111114.pdf

AV is separated into 5 categories

- Category 1: All analogue sources
- Category 2: Digital sources on tape with no transcode transfer required
- Category 3: Digital sources on tape with transcode transfer required
- Category 4: Born digital, non-tape formats
- Category 5: Optical discs

Category 1: All analogue sources

- Digitization specifications:
  - Playback the file, capture SDI or HDSDI output stream decompressed
  - Wrapper: .mov (QuickTime) or .avi
  - Bit Depth: 10 bit
  - Frame Size: 720x486
- Examples:
  - BetacamSP, Betamax, 2" quadruplex, etc.
  - Details: see pages 7-8

Category 2: Digital sources on Tape (non-transcode transfer possible)

- Digitization specifications:
  - Move data from one carrier to another
  - Wrapper: Native (.dv, imx, .mpeg, mp4, etc.) or .mov (QuickTime) or .avi
  - Bit Depth: Native, 8 bit or 10 bit
  - Frame size: Native
  - No reason to decompress if you support the codec.
- Notes:
  - Formats that allow access to the 1s and 0s on the tape (including metadata and error correction info)
  - i.e. DV, HDV, IMX in PD
  - Details: see pages 9-12

Category 3: Digital sources on Tape (transcode transfer required)

- Digitization specifications:
  - Playback machine must decode the video, then output is captured uncompressed
  - Wrapper: .mov (QuickTime) or .avi
  - Decompress to SMPTE 259M (10 bit, 720x486) if NTSC
- Notes:
  - Formats that do not allow access to the 1s and 0s on the tape
  - i.e. DigiBeta, D-1, etc.
  - Details: see pages 13-16

Category 4: Born Digital, Non-tape formats

- Digitization specifications:
  - Migrate from acquisition carrier
  - Must decide whether to leave compressed or uncompressed
  - Wrapper: Native (.dv, .imx, .mpeg, .mp4, etc.) .mov (QuickTime) or .avi
  - Bit Depth: Native, 8 bit or 10 bit
  - Frame size: 720x486 for NTSC
- Notes:
  - All file-based formats
  - i.e. P2 flash cards, iPhones, etc.
  - No reason to decompress if you decide to support the file, such as MPEG4.
  - Details: see pages 17-20

Category 5: Optical Discs

- Digitization specifications:
  - Create an ISO Disc image
  - Wrapper: ISO Image native (.img)
  - Bit Depth: Native, 8 bit
  - Frame size for SD: 720x480
  - Frame size for HD: Native
- Notes:
  - ISO Disc images are not playable easily
  - Representing discs as linear video would lose menus, navigation, subtitles
  - Need to make accessible surrogates
  - DVDs, BluRay
  - Details: see pages 21-24

Challenges

- Correctly identifying the material
- Understanding how the material was meant to be played back (eg. frame rate)
- Finding a compatible play back device:
  - In good working order
  - Within budget
  - With service professionals available
  - With extra parts available
- Obsolescence of digitization software used for AV
  - See Dave Rice's post here.

Reality check

- Local moving image digitization will likely be VHS centric
- Playback devices are difficult to obtain and maintain
  - Need specialists to properly run and maintain the devices
  - Need parts for upkeep
  - Need to pass on working knowledge to next generation of av archivists
  - Need to collaborate on shared services
- For older formats, vendors will be a more viable option

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!
- However, a non-proprietary lossless video codec FFV1 is being used at some institutions as a preservation format that will allow for lossless compression.

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 (10 bit instead of 8 bit)
- 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

Choosing a digital file type

- Consists of a “wrapper” and a “codec”
- Wrapper is like an envelope that holds metadata video and audio together
  - i.e. AVI, Ogg, Real Media, MP4, MJ2
- The video is encoded using a codec (compression-decompression algorithm)
  - i.e. JPEG 2000, WMV, MP4, H264
- Note: not all wrappers work with all codecs

http://www.jiscdigitalmedia.ac.uk/movingimages/advice/choosing-a-digital-video-file-type

File types for digital masters

- Choose ‘open’ over ‘proprietary’ formats
  - Open formats are maintained by a community of users, proprietary by a company
- Do not choose ‘lossy’ codecs, as image quality suffers (MPEG-4, real)
  - You will likely make a lossy copy of your master for a use and access copy
- A big factor will be file size. Test different lossless formats for quality and affordability.

http://www.jiscdigitalmedia.ac.uk/movingimages/advice/choosing-a-digital-video-file-type

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

Popular use and access formats

- Streaming:
  - Real Media Video
  - Windows Media Video
  - Quicktime
  - MPEG-4 (multimedia)
- Video CD:
  - MPEG-1
- DVD:
  - H.264/MPEG-4

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

See also…

- Vtape (Toronto)
  - Vtape is a vibrant distribution organization that represents an international collection of contemporary and historical video art and media works by artists. Tape runs the most complete video duplication and tape restoration facility in Canada.
- Video Pool (Winnipeg)
  - Video Pool Media Arts Centre is a non-profit, artist-run centre dedicated to the creation, exhibition and promotion of independent media art.
- Western Front (Vancouver)
  - The Western Front was founded in 1973 by eight artists (Martin Bartlett, Mo van Nostrand, Kate Craig, Henry Greenhow, Glenn Lewis, Eric Metcalfe, Michael Morris, Vincent Trasov) who wanted to create a space for the exploration and creation of new art forms.

Where to go for help with digitization questions…

Get involved, share, contribute back

Resources…

The AABC Archivist's Toolkit: Automation and Digitization: http://aabc.ca/resources/archivists-toolkit/automation/

Our Digital World: http://ourdigitalworld.org/services/resources/

ACA Sound and Moving Images Special Interest Section:

http://archivists.ca/content/sound-and-moving-images-sis-smisis

OCUL Digital Curation Interest Group:
https://spotdocs.scholarsportal.info/display/ODCC/OCUL+Digital+Curation+Community+Home

International Organization of Sound and Moving Image Archivists: http://www.iasa-web.org/

George Blood Audio and Video: http://www.georgeblood.com/Resources.html

JISC Digital Media Website: http://www.jiscdigitalmedia.ac.uk/

Digital Formats at the Library of Congress: http://www.digitalpreservation.gov/formats/intro/intro.shtml

Archives Association of Ontario: Archives Advisors: archivesadvisor@aao-archivists.ca (likely need to be a member)

Indigitization: Toolkit for the Digitization of First Nations Knowledge: http://www.indigitization.ca/

Preservation Self-Assessment Program (University of Illinois):

https://psap.library.illinois.edu/format-id-guide#audiovisual

Digital Library Federation: http://diglib.org

Mailing Lists…

- Digipres mailing list: digipres@ala.org
- JISC Digital Media mailing list: http://www.jiscdigitalmedia.ac.uk/mailing-list/
- Association of Moving Image Archivists (AMIA) discussion list: http://www.amianet.org/participate/listserv.php
- International Association of Sound and Audio Visual Archives mailing list: http://www.iasa-web.org/listserv.asp
- Web4lib mailing list: web4lib@listserv.nd.edu
- ALCTS Preservation Administrators Interest Group: padg@ala.org

PRONOM technical registry

- Holds information about file formats, and the software products which can process them
- Supports preservation efforts
- Search by file format, extension, vendor, software, lifecycle, migration pathway
- http://www.nationalarchives.gov.uk/aboutapps/PRONOM/tools.htm

Welcome | About GDFR | UDFR | Documentation | Wiki | Download | Links

Welcome to the Global Digital Format Registry!

The GDFR is meant to be a distributed and replicated registry of format information populated and vetted by experts and enthusiasts world-wide.

Formation of the Unified Digital Formats Registry (UDFR)

In April 2009 the GDFR initiative joined forces with the UK National Archives' PRONOM registry initiative under a new name - the Unified Digital Formats Registry (UDFR). The UDFR will support the requirements and use cases compiled for GDFR and will be seeded with PRONOM's software and formats database.

For more information about the UDFR, plase see the UDFR website.

The GDFR is a collaborative project of the Harvard University Library, NARA and OCLC with funding generously provided by the Andrew W. Mellon Foundation.

The Andrew W. Mellon Foundation

http://www.gdfr.info/

Metadata

Why create metadata?

Types of metadata

Systems & Schemas

Why do we need metadata?

- Digital identification
  - Used to differentiate one object from another
  - Used to identify sets of data
- Organizing e-resources
  - Organizing links to resources based on audience or topic
  - Building these pages dynamically from metadata stored in database

Why do we need metadata?

- Resource discovery
  - Allowing resources to be found by relevant criteria
  - Identifying resources
  - Bringing similar resources together
  - Distinguishing dissimilar resources

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/provenance
    - detail its physical characteristics and behaviour 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, subject, 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 and Technical
  - 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

Wrapper Formats

- Wrapper formats tie together many different types of metadata
- Ideal for preservation
- MPEG-21 and METS support moving images
- XML based

MPEG-21

- Specialized for preservation of moving images
- Allows detailed capture of intellectual rights info
- Very complex and hence only adopted by specialized archives

Agnew, G., Kniesner, D., & Weber, M. B. (2007). Integrating MPEG-7 into the moving image collections portal. *Journal of the American Society for Information Science and Technology*, 58(9), 1357-1363.

METS

- Metadata Exchange and Transmission Standard
- Created for describing complex digital library objects
- Components of a METS File:
  - METS Header
  - Descriptive Metadata – MODS, MARC, MARCXML etc.
  - Extension schemas – format specific info, provenance and copyright
  - Structural Map – hierarchy and links to digital objects
  - Structural Links and behaviours

RAD, MARC, MARCXML, MODS

- RAD (Rules of Archival Description)
- EAD (Encoded Archival Description)
- MARC (Machine Readable Cataloguing Record)
- Can easily transform:
  - MARC21 > MARCXML >MODS
- MODS is a subset of MARCXML elements
  - MODS is embedded in METS records for item level descriptive metadata
  - RAD-EAD-MARC-DC-MODS Crosswalk (courtesy of Creighton Barrett, Dalhousie University Archives)

Sample Extension Schemas

- Audio
  - AudioMD, specific to audio e.g., channel or track specifications, sampling frequency.
- Video
  - VideoMD, specific to video files, e.g., bit rate, compression codec.
  - MIX, specific to images, e.g., bits per pixel, color space
- Images
  - ImageMD, specific to images e.g., type or condition
  - MIX, specific to images, e.g., bits per pixel, color space

http://www.loc.gov/rr/mopic/avprot/metsmenu2.html

Sample Extension Schemas cont'd…

- Other
  - RightsMD: Rights, restrictions, and/or other categorizing information that can be used to support rights-management and/or access-management systems.
  - ProvenanceMD: About the events/steps/processes that occurred in reformatting or migrating entities.
  - PREMIS: Captures core preservation metadata needed to support the long-term preservation of digital materials
  - PBCore: metadata standard for audiovisual media developed by the public broadcasting community

http://www.loc.gov/rr/mopic/avprot/metsmenu2.html

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

Dissemination

Platforms

Collaborative opportunities

Sample Digital Collections Platforms

- Content DM , MINISIS, PastPerfect (vendor)
- Greenstone, Kete, Omeka, Scriblio (open source)
- California Digital Library's eXtensible Text Framework (XTF) (open source)
- Repository platforms: DSpace, Islandora (Fedora) (open source)

SORT BY: TITLE | Date

FILTER: ALL | Tags | Newber

All 270 items

Page: 1 2 3

.577 caliber percussion rifle-carbine, 1862

Royal Small Arms Factory, Enfield, England
Tags: soldiers

"$100,000 reward!" 1865

George F. Nesbitt & Co., New York
Tags: assassination, broadsides, conspirators

"10 Likely and Valuable Slaves at Auction," 1823

Tags: broadsides, slavery

"100 Dollars Reward," July 6, 1857

Tags: broadsides, slavery

2,000,000 Acres of Illinois Central R.R. Lands, Illinois Central Rai 1856

New York: John W. Amerman
Tags: Illinois, free labor, pamphlets, railroads

Omeka http://omeka.org/

Island LIVES

TITLES PEOPLE PLACES ORGANIZATIONS IMAGES CONTRIBUTE FAQ

Book Search

Text

and

Text

search

Repository Advanced Search

Total Hits = **2312,** Number of Hits/page = 50

You may not have sufficient privileges to view any or all of the items found. The objects you have rights to view v

Next >

1. **ilives:90627-z_004-img1p**
Score:(9.3404255)
Photograph 1 - Back Matter 4 - It happened in Iona

2. **ilives:90627-p_116-img1i**
Score:(9.3404255)
Illustration 1 - Page 116 - It happened in Iona

3. **ilives:90627-p_109-img1i**
Score:(9.3404255)
Illustration 1 - Page 109 - It happened in Iona

4. **ilives:90627-p_105-img1i**
Score:(9.3404255)
Illustration 1 - Page 105 - It happened in Iona

5. **ilives:90627-p_099-img1i**
Score:(9.3404255)
Illustration 1 - Page 99 - It happened in Iona

Islandora http://islandora.ca/

North Carolina MAPS

ncmaps home | browse | about | search selected collections | help | my favorites | all collections home

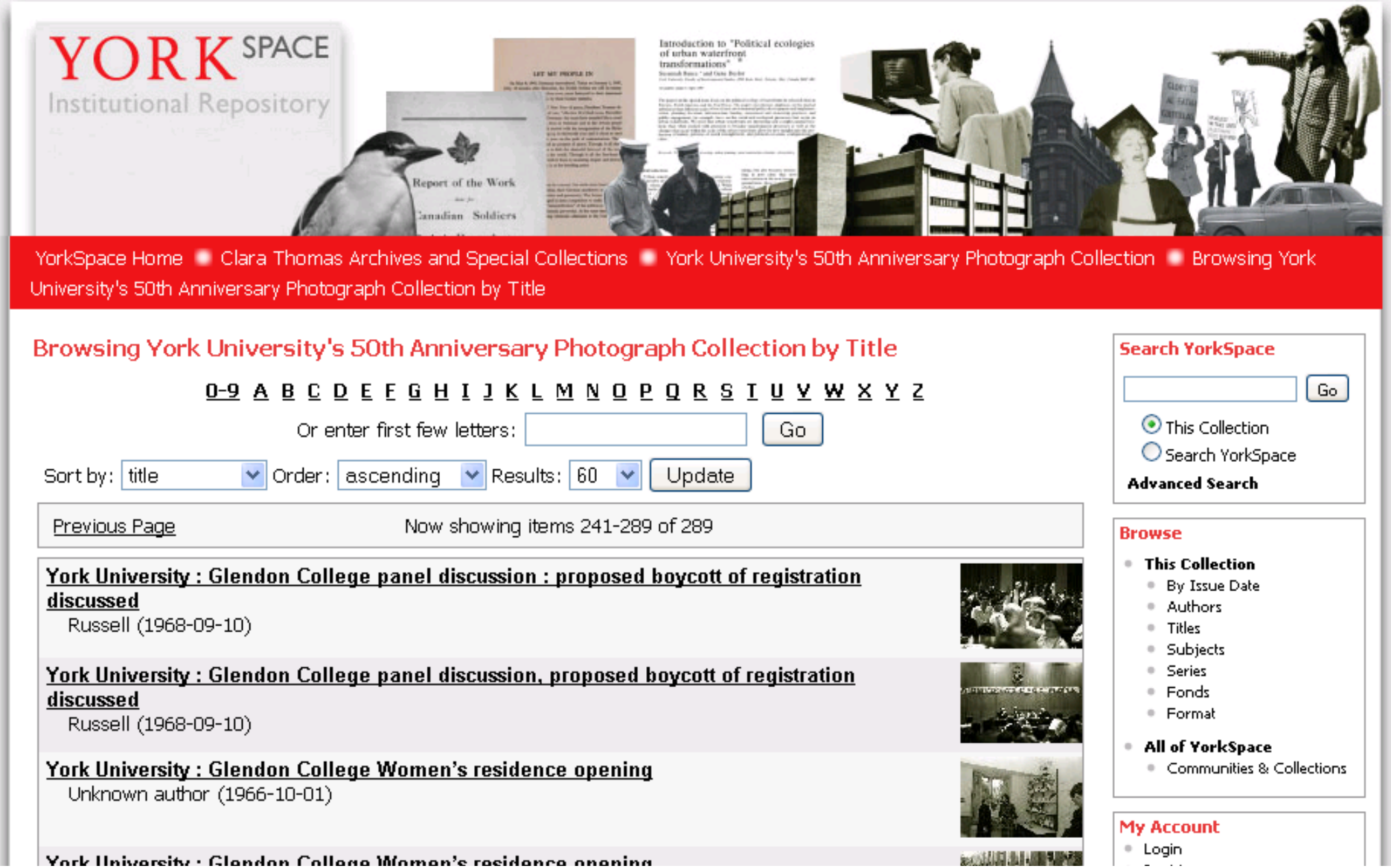
Search results: 149 item(s) for: **"Atlas"** search again

select all : clear all : add to favorites

page 1 of 5 : (<< 1 2 3 4 5 >>) :: previous : next

	Image:	Title:	Date Published:	Publisher:	Repository:
☐ 1.		Carte qui contient une desciption des iles & terres que les Anglois possedent dans l'Amerique Septentrionale	circa 1719	chez l'Honore & Chatelain Libraires	North Carolina Collection
☐ 2.		Carolina	1729		North Carolina Collection
☐ 3.		Carte de la Caroline méridionale et septentrionale et de la Virginie	circa 1770s		North Carolina Collection
☐ 4.		Carte de la Caroline et Georgie. Pour servir a l'Hist. des Etablissemens Europeens. Tiree des Auteurs Anglois par M. B. Ing. de la Marine. A. v. Krevelt, Sculpsit Amsteldam 1773.	1773	Raynal, Guilaume Thomas Francois	North Carolina State Archives
☐ 5.		Carte de la Caroline et Georgie pour servir à l'Histoire générale des voyages	circa 1780		North Carolina Collection
☐ 6.		North Carolina	1795		North Carolina State Archives

Content DM http://www.contentdm.com/



Dspace http://dspace.org/

Opportunities for collaboration…

Economies of scale

Share, contribute back, support

Take advantage of consortial communities with a higher profile

vita digital toolkit

Tour | Pricing | Showcase | About | Contact

Multimedia | Newspapers | Features | Sign Up | Terms of Service

Pricing Plans

Something for every project idea

Multimedia Collections

Basic

For small organizations

1 site
2 management accounts
1 GB storage
$349/year

- ✓ Upload images, text, 3D objects
- ✓ Customizable display
- ✓ Social media & interactive

Standard

Our most popular plan!

2 sites
8 management accounts
10 GB storage
$899/year

VITA Basic +...

- ✓ Upload audio & video
- ✓ Multilingual site option

Plus

For large institutions

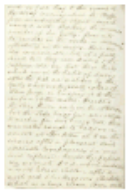
8 sites
Infinite management accounts
50 GB storage
$1299/year

VITA Standard +

- ✓ Upload streaming video
- ✓ Fan mail

http://vitatoolkit.ca/

Spotlight Item



Excursion to Scotland, 1832
Seven commercially produced plates of prominent buildings and views are inserted. In addition there are a number of small hand-drawn pictures and a hand-drawn map included in the text

About the Internet Archive

Background

Frequently Asked Questions

Contributors

Alex Aitken
Alex White
Andrea Mills
Carole Moore
Dean Seeman
Gabe Juszel
Graham Stewart
John McArdle
Jonathan Bengtson
Katie Lawson
Kent Weaver
Lajolla Young

Welcome to Canadian Libraries 184,722 items

Historical Texts

A dictionary of the English language By Samuel Johnson(details | pdf)
A dictionary of the English language in which the words are deduced from their originals, and illustrated in their different significations by examples from the best writers : to which are prefixed, a history of the language, and an English grammar (1785)

- A general history of all voyages and travels throughout the old and new world (1708) (details | read it)
- The history of the devil, as well ancient as modern (1727) (details | read it)
- Pastorals, epistles, odes, and other original poems, with translations from Pindar, Anacreon, and Sappho (1748) (details | read it)
- Esprit, maximes et principes de m. Jean-Jacques Rousseau, de Geneve (1764) (details | read it)
- Twenty stories from Grimm (1896) (details | read it)

Photography

- Photography in the studio and in the field (1887) (details | read it)
- Beginner's guide to photography (1888) (details | read it)
- Uncle Alberts manual of practical photography (1890) (details | read it)
- Studio light ... a magazine (1910) (details | read it)
- The evolution of photography (1890) (details | read it)

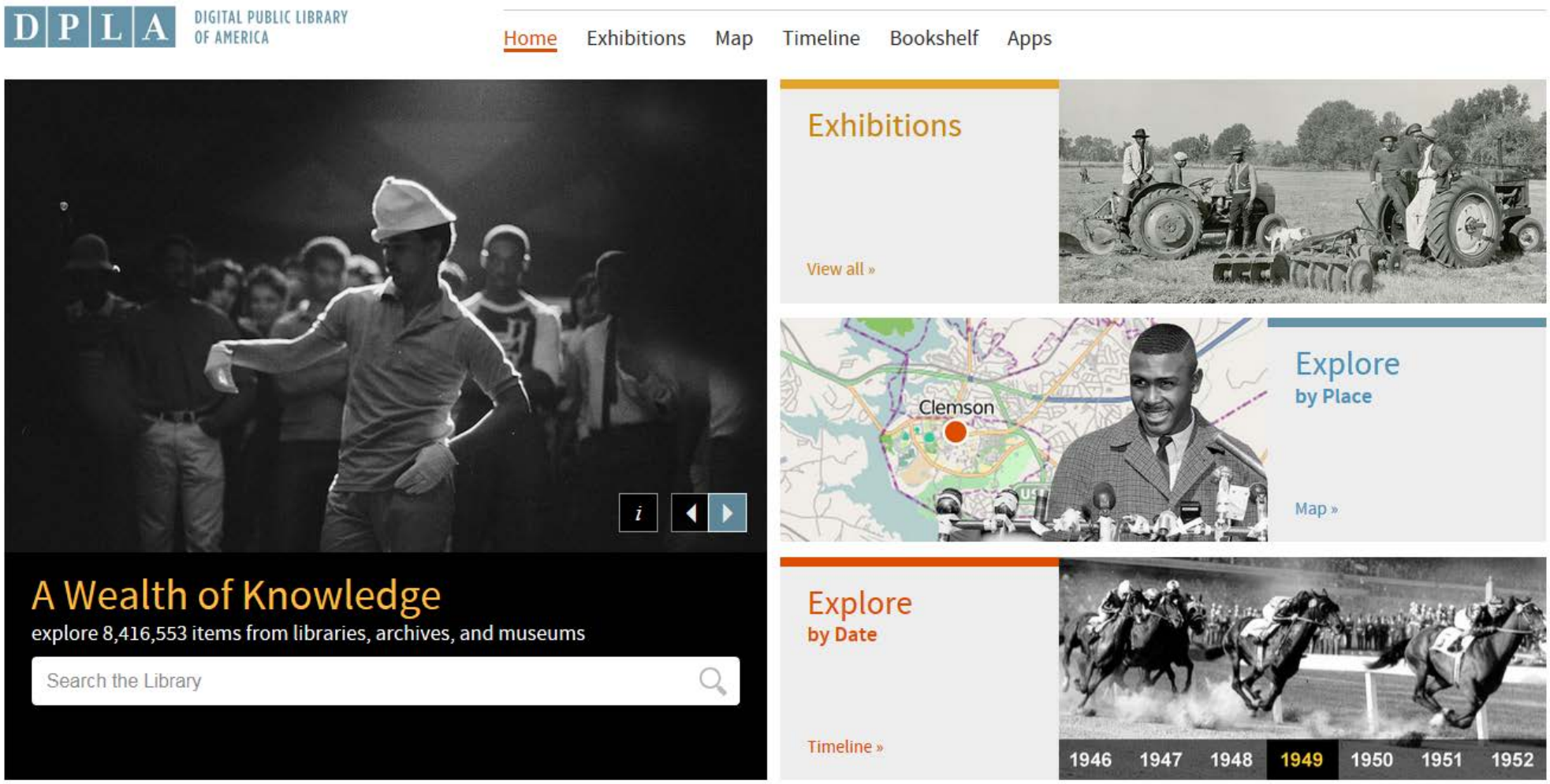
Ephemeral Canadiana

- Tremendous Toronto (1914) (details | read it)
- Se-ke and the beaver (1920) (details | read it)
- Grimsby Park, historical and descriptive: with biographical sketches of the late president Noah Phelps and others (1900)

Most Downloaded Items Last Week more

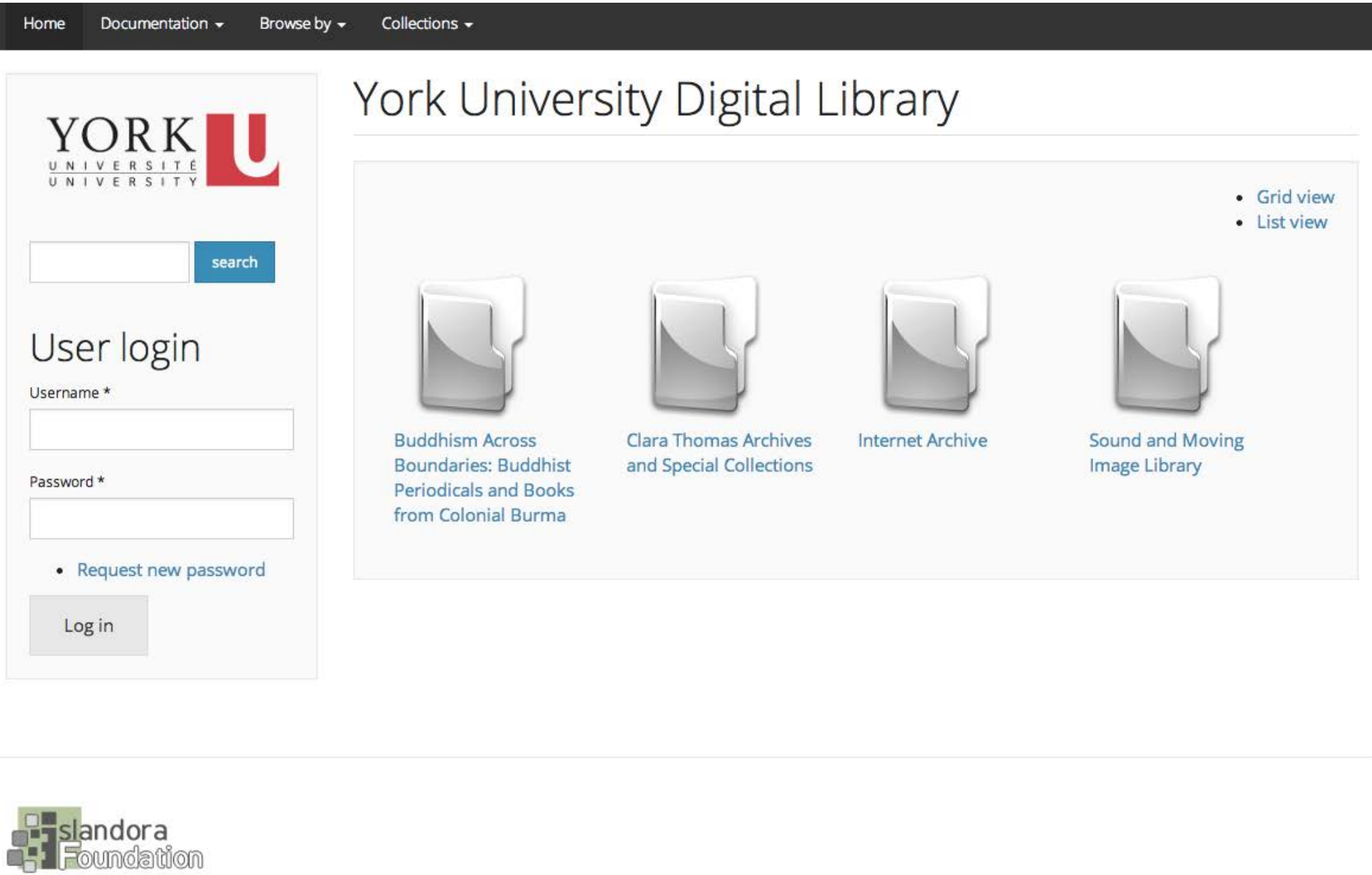
1. Who's who in China; containing the pictures and biographies of China's best known political, financial, business and professional men
1,746 downloads
2. A general history of all voyages and travels throughout the old and new world, from the first ages to this present time, illustrating both the ancient and modern geography, containing an accurate description of each country, its natural history and...
1,658 downloads
3. A dictionary of the English language : in which the words are deduced from their originals, and illustrated in their different significations by examples from the best writers : to which are prefixed, a

http://www.archive.org/details/toronto



http://www.youtube.com/watch?feature=player_embedded&v=mongLBa4ewM

Islandora at York University


Home
Documentation
Browse by
Collections
YORK UNIVERSITÉ UNIVERSITY
search
User login
Username *
Password *
Request new password
Log in
York University Digital Library
Grid view
List view
Buddhism Across Boundaries: Buddhist Periodicals and Books from Colonial Burma
Clara Thomas Archives and Special Collections
Internet Archive
Sound and Moving Image Library
Islandora Foundation
http://digital.library.yorku.ca/

Current Collections

- Include audio, video, images, text
- Modularity of forms allows assignment to various collections, can include instructions to guide digitization staff
- Video metadata sample:
- http://digital.library.yorku.ca/yul-130766/restorative-justice-law/view_premis#overlay-context=yul-130766/restorative-justice-law

Statistics as of Feb 2014

Collection	PID	Fedora Objects	Datastreams	Space used (G)
)Golhaye Rangarang(نبرنامه گل‌های رنگارنگ	yul:68504	406	2434	17.706
)Taraneha ve Khatereha(ترانه‌ها و خاطره‌ها	yul:61286	16	126	0.624764
)Barg-e-Shabz(پوچه نبرنامه برگ سبز	yul:67908	298	1786	11.04
)Golhaye Tazeh(ل‌های تازه	yul:71789	165	1152	5.989
)Golhaye Sahrai(گل‌های صحرایی	yul:71744	45	311	1.085
)Yek Shakhe Gol(یک شاخه گل	yul:68206	419	2512	12.489
AIDS Committee of Toronto	yul:72118	2	37	5.294
Internet Archive	yul:ia	133	304183	324.5560838
Allan Robb Fleming	yul:F0529	74	589	5.539
Barbara Godard	yul:F0236	2	53	1.866
Clara Thomas Archives collection	yul:F0486	12	96	0.338609375
Birth Control & VC Info Centre	yul:85483	13	280	59.075
Domingos Marques	yul:F0573	28	221	1.539
Edgar Wardwell McInnis fonds	yul:f0353	6	85	5.128
Excalibur Publications Inc.	yul:F0502	9	68	0.0343007813
James Tenney fonds	yul:F0428	20	792	19.841
John Warkentin fonds	yul:F0184	140	1116	5.363
Lou Wise fonds	yul:F0539	6237	50001	201.005
Mariposa Folk Foundation	yul:F0511	245	13056	35.468
Michael Posluns fonds	yul:F0382	41	1540	20.526
Murray George Ross fonds	yul:F0398	5	37	0.1186210938
The Music Gallery	yul:F0119	30	1155	26.957
Toronto Telegram	yul:F0433	10572	86221	179.653
Varpu Lindström fonds	yul:F0558	123	982	2.208
YFile	yul:yfile	427	4242	208.107
York University Archives Calendar Collection	yul:F0158	28	295	11.329
York University Computing and Network Services	yul:F0477	116	924	2.259
York University Department of Communications	yul:F0047	9	68	0.2374521484
York University Libraries	yul:F0066	3	20	0.2088105469
York University Photograph collection	yul:F0091	35	276	1.681
Cephalonica-Ithaca Association of Toronto	yul:F0646	10	132	1.037
Law & Social Change: Restorative Justice	yul:LW2750F03	7	92	1.742
Totals				
32		19676	474882	1170.0446417454

Preservation needs & dissemination needs

Collection	Items	Described	Size	GRAP
AIDS Committee of Toronto VHS collection	927	Yes	32T*	Yes
Music Gallery Concert Series Collection	338	Yes	235G	Yes
Iranian Music Collection	1759	Yes	40G	No
Canadian Independent Film and Video collection	450	Yes	31T**	Yes
World War II projectionist collection	52	Yes	3.7T**	Maybe
Dance in Canada: Jean A. Chalmers Choreographic collection	122	Yes	22.7T**	Yes
Labatt's Breweries of Canada sports collection	262	Yes	48.7T***	No
National Film Board of Canada collection	475	Yes	88.4T***	No
Toronto Reference Library	211	Yes	39.3T***	No
On-demand digitization	???	???	???	Yes

* Assuming average of 30 minutes per VHS * 1.18G/min (VHS/NTSC)
** Assuming average of 60 minutes per VHS * 1.18G/min (VHS/NTSC)
*** Assuming average of 60 minutes per 16mm * 3.1G/min (1280x720)

UIT Hosting*

High availability model - $140/month + storage costs

- 12GB RAM
- 2 virtual CPUs (one included, additional $30 per)
- Storage (No local storage)
    - Tier 1: SAS/FC - $0.24/GB/month
    - Tier 2: SATA - $0.16/GB/month
- Managed (included) - automatic restart of server on another host if the primary fails
- Backup
    - 2x Storage + $0.06/GB/month
- Yearly cost

Base system	$110.00	
Additional processor	$30.00	$30.00/processor
Storage (1TB)	$163.84	$0.16 / GB / mth
Backup	$389.12	2 x Storage cost + $0.06 / GB / mth
Systems management	$0.00	
Total	$8,315.52	

Challenge of Costs

- 1 TB = 1024 GB
- AIDS committee VHS collection: 32TB
- Cost to host:
  - 32 TB x 1024 = 32,768 GB
  - 32,768 GB x $0.24/GB per month = $7864.32 per month to host just that one collection!

OLRC: Ontario Libraries Research Cloud

- The Ontario Library Research Cloud (OLRC) project is a collaboration of Ontario's university libraries to build a high capacity, geographically distributed storage and computing network using proven and scalable open source cloud technologies. The OLRC will be designed to house large volumes of digital content to allow for cost effective and sustainable long-term preservation and to support data and text mining using innovative research tools.
- https://spotdocs.scholarsportal.info/display/ODLRC/About+the+OLRC
- Cost savings but requires investment, commitment
- Closed to academic libraries

Thanks for your time

Special thanks to Andrea Kosavic, digital initiatives librarian at York University Libraries, who developed this lecture and has presented it for the past several years. Thanks also to Nick Ruest, digital assets librarian for statistics regarding YUDL and Crieghton Barrett, for providing online access to research and findings generated by Dalhousie's implementation of ArchivesSpace and ICA-AtoM.