DHPS | NY

Documentary Heritage & Preservation Services for New York

Thanks for joining us! Today's presentation will begin shortly.

Please check your audio and video settings.
You should currently hear music in the background.

If you have questions or want to report any technical issues,
contact us at info@dhpsny.org or (215) 545-0613.

DIGITAL PRESERVATION FOR SMALL REPOSITORIES

Bonnie Weddle
New York State Archives

March 14, 2018

Welcome!

Bonnie Weddle
Coordinator, Electronic Records
New York State Archives

What we'll do today

- Explore digital challenges
- Define key terms
- Detail how to set up a processing computer
- Discuss core digital preservation activities
- Identify next steps and helpful resources

What we won't do today

- Identify any one-size-fits-all answers
  - They don't exist
- Give in to despair
  - We can figure this out
  - We're running a relay race
- Delve into advanced topics
  - Digital preservation theory
  - File format migration

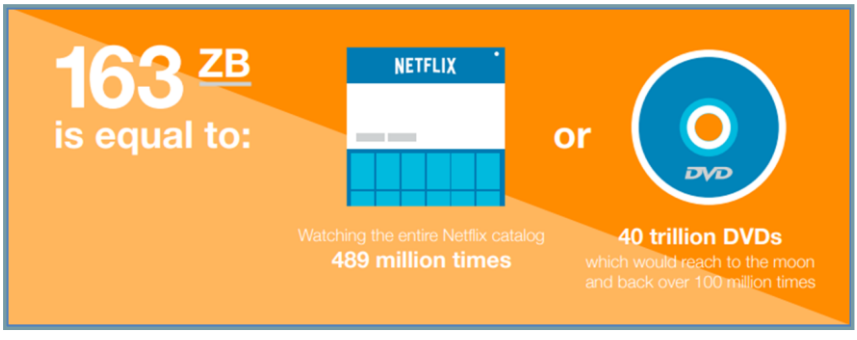
Digital challenges

Technology dependence

- Digital files can be accessed only via hardware and software
  - Some files can be accessed in multiple environments
  - Some files require a very specific combination of hardware and software

Ever-increasing volume

- IDC: **global total** of **163 zettabytes** of data by **2025**



Report: IDC, *Data Age 2025: The Evolution of Data to Life-Critical* (https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf)

Graphic: IDC, "The Evolution of Data Through 2025" [excerpt] (https://www.seagate.com/files/www-content/our-story/trends/files/data-age-2025-infographic-2017.pdf)

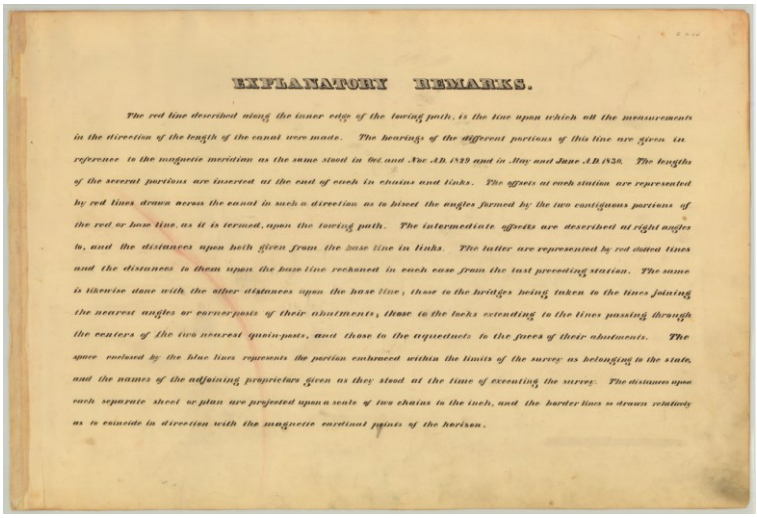
Obsolescence, instability, vulnerability

- Technology changes rapidly
  - Hardware, software, storage media
- Every form of computer storage media
  - Is inherently fragile
  - Has an unpredictable lifespan
- Ever-smaller devices and media
  - Hold ever-larger quantities of data
  - May be lost, stolen, or destroyed
- Internet connectivity introduces risk
  - Viruses and other malware
  - Hacking

Decentralization and inconsistency

- Data may be spread across
  - Network servers
  - Desktops and laptops
  - Smartphones
  - Tablets
  - “Smart” devices
  - Portable media
  - “The cloud”
- Creators typically given little training or guidance
  - Individual file naming and organizing idiosyncrasies prevail
  - Duplicate and draft copies proliferate
  - Organizational records may be stored on personal devices and in personal e-mail accounts

Key terms



EXPLANATORY REMARKS.

The red line described along the inner edge of the towing path, is the line upon which all the measurements in the direction of the length of the canal were made. The bearings of the different portions of this line are given in reference to the magnetic meridian as the same stood in Oct and Nov A.D. 1829 and in May and June A.D. 1830. The lengths of the several portions are inserted at the end of each in chains and links. The offsets at each station are represented by red lines drawn across the canal in such a direction as to bisect the angles formed by the two contiguous portions of the red or base line, as it is termed, upon the towing path. The intermediate offsets are described at right angles to, and the distances upon both given from the base line in links. The latter are represented by red dotted lines and the distances to them upon the base line reckoned in each case from the last preceding station. The same is likewise done with the other distances upon the base line; those to the bridges being taken to the lines joining the nearest angles or corner posts of their abutments, those to the locks extending to the lines passing through the centers of the two nearest quoin-posts, and those to the aqueducts to the faces of their abutments. The space enclosed by the blue lines represents the portion embraced within the limits of the survey as belonging to the state, and the names of the adjoining proprietors given as they stood at the time of executing the survey. The distances upon each separate sheet or plan are projected upon a scale of two chains to the inch, and the border lines so drawn relatively as to coincide in direction with the magnetic cardinal points of the horizon.

"Explanatory Remarks" page, Erie Canal Survey, 1830.
New York State Archives.

http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/43128)

Digital preservation

- **Policies**, **strategies**, and **actions** that ensure long-term access to digital content
- **Accurate** rendering of **authenticated** content **over time**

American Library Association, Association for Library Collections and Technical Services, Working Group on Defining Digital Preservation, "Definitions of Digital Preservation" (http://www.ala.org/alcts/resources/preserv/defdigpres0408)

Digitized (born analog)

- Paper, film, or analog tape-based materials that have been **converted** to digital format

Society of American Archivists, *A Glossary of Archival and Records Terminology*, s.v. "Born analog" (https://www2.archivists.org/glossary/terms/b/born-analog)
Society of American Archivists, *A Glossary of Archival and Records Terminology*, s.v. "Digitization" (https://www2.archivists.org/glossary/terms/d/digitization)

Born digital

- Materials **created digitally**
  - May be converted to analog format
    - Paper printouts
    - Computer-output microfilm
  - May be maintained digitally
    - For convenience
    - Analog conversion simply not an option

Society of American Archivists, *A Glossary of Archival and Records Terminology*, s.v. "Born digital" (https://www2.archivists.org/glossary/terms/b/born-digital)

Digital content

- Any form of digitized or born-digital information
  - Records
  - Personal papers
  - Images
  - Publications
  - Audio
  - Video
  - Databases and datasets
  - Artworks

Hardware

- Physical, mechanical, and electrical components of a computer or computer system
  - Monitors
  - Keyboards
  - Mice and trackpads
  - Floppy and CD/DVD drives
  - Storage devices
  - Processors
  - Circuitry

Society of American Archivists, *A Glossary of Archival and Records Terminology*, s.v. "Hardware" (https://www2.archivists.org/glossary/terms/h/hardware)

Software

- Instructions that govern hardware operations
  - System—governs operation of specific hardware
    - Operating systems (Windows, macOS, Linux, iOS, Android, Chrome OS)
    - Device drivers
  - Application—performs specific tasks
    - Word processors (Word, Notepad, Pages)
    - Media players (Windows Media Player, iTunes)
    - Web browsers (Explorer, Firefox, Chrome)
    - Photo editors (Photoshop, iPhoto)
    - Games

Society of American Archivists, *A Glossary of Archival and Records Terminology*, s.v. "Software" (https://www2.archivists.org/glossary/terms/s/software)

File formats

- Conventions for encoding human-readable data into binary form and back again
- Enable software to interpret and display data correctly
- May be open/free
  - Full technical specification freely available
- May be proprietary
  - Owned and controlled by a corporation

Linux Information Project, "File Formats: A Brief Introduction" (http://www.linfo.org/file_format.html)

"The cloud"

- Storing and accessing software and files on the Internet instead of your computer or local server or storage array
  - GoogleDocs, Google Drive, and most other Google services
  - Facebook, Twitter, Instagram, and other social media services
  - Netflix, Hulu, Instagram, and others use Amazon Web Services cloud data centers
- Offers both promise and peril

Eric Griffith, "What is Cloud Computing?" *PC Magazine*, May 3, 2016 (https://www.pcmag.com/article2/0,2817,2372163,00.asp)

Metadata

- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource

National Information Standards Organization, *Understanding Metadata*, 2004 edition (https://web.archive.org/web/20080916055242/http:/www.niso.org/publications/press/UnderstandingMetadata.pdf)

- Descriptive
  - Creator
  - Scope and content
- Structural
  - Pages, chapters
- Administrative
  - Rights management
- Preservation
  - Fixity
  - When and how created
  - File format(s)

Setting up a processing computer

Why a processing computer?

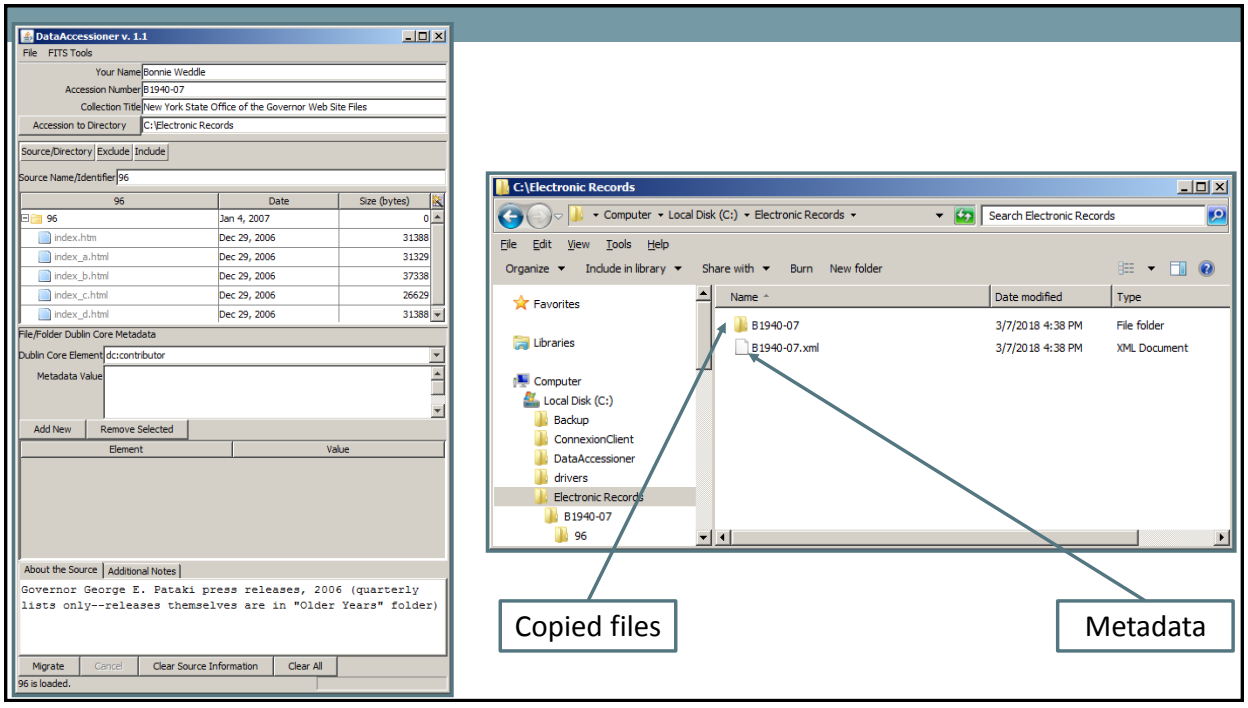
- Helps with a variety of tasks
  - Inventorying born digital and digitized files
    - Especially those housed on portable media
  - Selecting files for preservation
    - Review of content
  - Storage for content warranting preservation
    - Preferably short term, may be longer term depending on resources

Key requirements

- Reasonably new laptop or desktop
  - Slightly older with floppy disk drive is better than brand-new without
- Ability to read portable media
  - May need to purchase drives or accessories
- Software and the ability to install it
  - Applications needed to access files
  - Data Accessioner
- If possible, devote computer exclusively to processing
  - Reduces risk of exposure to viruses, etc.

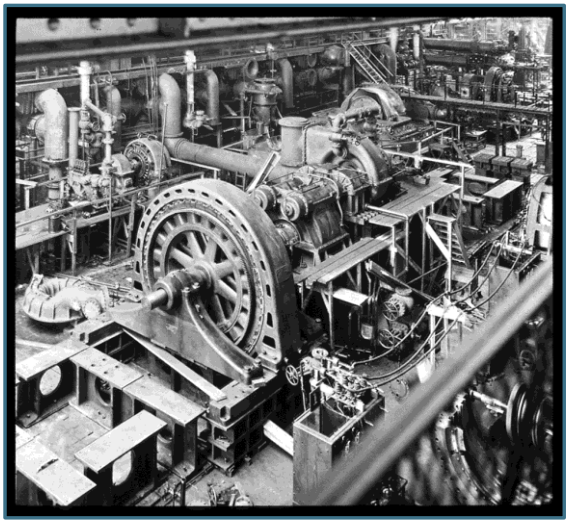
Data Accessioner

- Facilitates copying of data
- Extracts and creates metadata that supports preservation
  - Fixity (checksums)
    - Error detection
  - File information
    - Format and version
    - Date of last modification
  - Processing information
    - Archivist who copied data
    - Date and time of copying
- Supports Dublin Core descriptive metadata (optional)




```
- <collection name="New York State Office of the Governor Web Site Files">
  - <accession number="B1940-07">
    - <ingest_note>
        New York State Office of the Governor Web Site Files transferred by Bonnie Weddle on Wed Mar 07 16:38:51 EST 2018
      </ingest_note>
      <ingest_time>00:00:02.2978</ingest_time>
    - <source_note>
        Governor George E. Pataki press releases, 2006 (quarterly lists only--releases themselves are in "Older Years" folder)
      </source_note>
    - <folder name="96" last_modified="2007-01-04T16:49:55.000">
      - <file name="index.htm" last_modified="2006-12-29T12:07:16.000" size="31388" MD5="c6c006f551470e27ee8c728e780a9180">
        - <premis:object xsi:type="premis:file">
          - <premis:objectIdentifier>
              <premis:objectIdentifierType>uuid</premis:objectIdentifierType>
              <premis:objectIdentifierValue>c0046472-e543-4c0e-b35c-3b99abb28ea2</premis:objectIdentifierValue>
            </premis:objectIdentifier>
          - <premis:objectCharacteristics>
              <premis:compositionLevel>0</premis:compositionLevel>
            - <premis:fixity>
                <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
                <premis:messageDigest>c6c006f551470e27ee8c728e780a9180</premis:messageDigest>
                <premis:messageDigestOriginator>OIS File Information</premis:messageDigestOriginator>
              </premis:fixity>
              <premis:size>31388</premis:size>
            - <premis:format>
              - <premis:formatDesignation>
                  <premis:formatName>Hypertext Markup Language</premis:formatName>
                  <premis:formatVersion>4.01</premis:formatVersion>
                </premis:formatDesignation>
              - <premis:formatRegistry>
                  <premis:formatRegistryName>http://www.nationalarchives.gov.uk/pronom</premis:formatRegistryName>
                  <premis:formatRegistryKey>fmt/100</premis:formatRegistryKey>
                </premis:formatRegistry>
                <premis:formatNote>text/html</premis:formatNote>
                <premis:formatNote>DROID Signature File Version: 88</premis:formatNote>
                <premis:formatNote>Identified by: Droid v6.1.5</premis:formatNote>
                <premis:formatNote>Identified by: Jhove v1.11</premis:formatNote>
```

Core digital preservation activities



Marine turbine testing room, General Electric Works, Schenectady, N.Y., 1917. New York State Archives.

(http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/3319)

Overlapping and ongoing

- **Identify** digital content you have (or will have)
- **Select** content that warrants preservation
- **Store** content that warrants preservation
- **Protect** content from loss and harm
- **Manage** content over the long term

Library of Congress Digital Preservation Outreach & Education, http://www.digitalpreservation.gov/education/

Identify

Scientist taking part in a biological survey of Great Sacandaga Lake, Fulton County, N.Y., July 28, 1931. New York State Archives. (http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/4482)

Why identify?

- If you don't know what you have (or will likely have in the future), you can't plan for its care
  - Digital preservation requires ongoing commitment of resources
  - *Not all content warrants preservation*

Inventory what you have

- Include both collections and content your organization creates
  - If your organization is large, may need to pilot test and work unit by unit
- Aim for consistency and conciseness
- Keep in mind that your inventory will grow and change over time
  - Create a spreadsheet or database

Information you should capture

- Physical location
  - Collier Papers, box 3, "Family History" folder; now in Portable Media box 1
  - 18571-07, behind folders in box 74
  - J:\Executive Board\Meeting Minutes
  - Executive Director e-mail account (HistSocExecDir@gmail.com)
- Type of media
  - 3.5" floppy disk
  - CD-R disc
  - USB external hard drive
- Label on media
  - "Work stuff"
  - "Travel expenses, 2009"
- For computers or external drives
  - Manufacturer
  - Serial number
- Other useful information
  - Visible damage
  - Mentioned in access tools?
  - "Creator used a Mac"

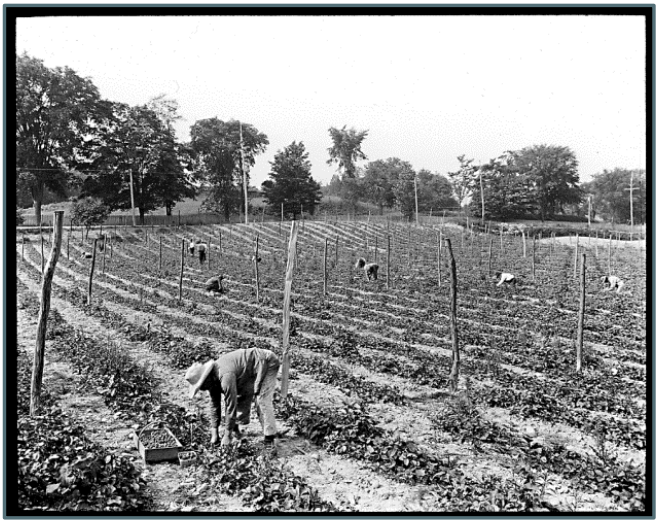
If you've set up a processing computer

- File format types
  - Text
  - Images
  - Audio
  - Video
  - Maps/geospatial data
  - Web content
  - Spreadsheets
  - Databases
- Size and extent
  - Number of files
  - Total size of files (KB, MB, GB, TB)
- Dates
  - Last modified
  - Transferred to archives
- Copying/transfer information
  - Is media readable?
  - Files copied to hard drive or other storage via DataAccessioner?
- For organizational records
  - Creator
  - Person or unit currently responsible
  - Estimated future growth

Analyze your results

- What do we have that we didn't know about?
- What should we have that we have now?
- What will we likely acquire in the future?
- What are we **required** to keep?
- What do we need to review?

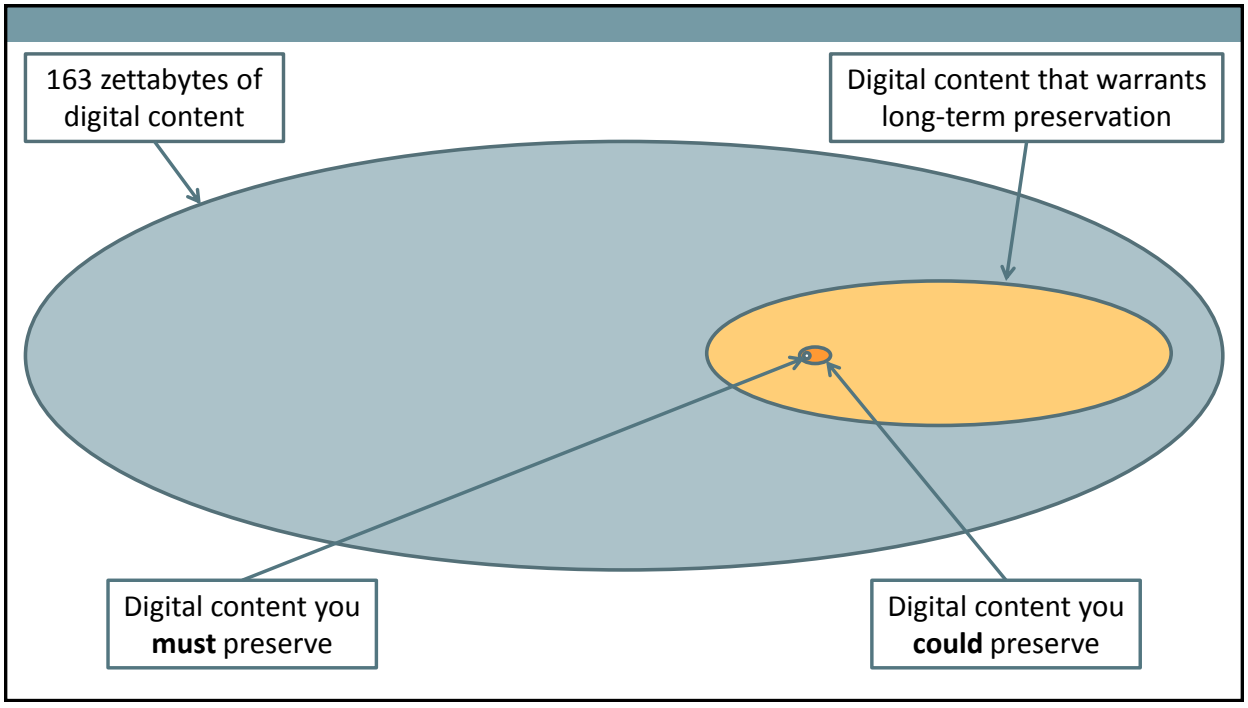
Select



Picking strawberries in New Hope, Orange County, N.Y., 1913. New York State Archives. (http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/3036)

Why not preserve everything?

- Storage is cheap, but management is not
  - One hard drive may cost $100
  - One hour of IT time may also cost $100
  - Hardware, software, and storage media must be replaced periodically
- *Not all content warrants long-term preservation*
  - We're archivists—we appraise!



Content should match mission

The mission of the Boynton Beach City Archives and Local History Collection is to identify, collect, preserve, and make accessible records, resources, and personal papers of enduring historical, legal, administrative, and fiscal value that document the social, economic and political development of Boynton Beach, Palm Beach County, and the state of Florida.

Boynton Beach City Library, *Boynton Beach City Library Archives and Local History Collection Development Policy* (http://www.flamuseums.org/site/assets/files/1277/boyntoncollection_development_policy.pdf)

Other criteria

- Appeal and interest
  - Value to researchers or other constituencies
- Born digital vs. digitized
  - Born digital, can't be made analog?
  - Digitized copy, original is unstable?
- Uniqueness
  - If it's not unique, why are you preserving it?
- Completeness
  - Is this the most comprehensive content?
- Manageability
  - Do you have the staff, equipment, and budget needed to care for it?
- Access
  - If you can't make it accessible, why are you preserving it?
- Legal obligations
  - Collections
  - Organizational records
- Consequences of loss
  - Impact on your operations
  - Damage to your reputation

Selection criteria and decisions

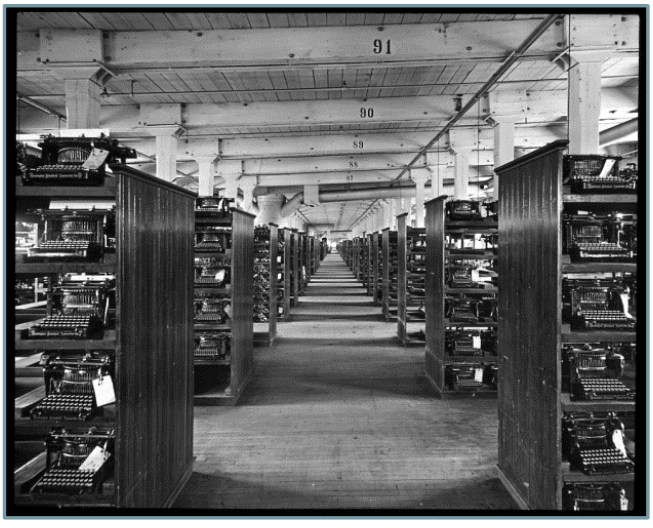
- Document in writing
  - Collection development policy
- Implement
  - Review inventory of existing materials
  - Assess digital content likely to come your way
  - Dispose of content that doesn't meet your criteria

The Boynton Beach City Library Archives and Local History Collection will acquire materials in a variety of formats.

For the Floridiana Collection hard cover books are the preferred format, however trade paperbacks and mass market paperbacks will be purchased when they are the only viable option. Books of all sizes and shapes are collected, including (but not limited to): oversized (gazetteers, art books, local government documents), manuscripts, pamphlets, homemade books, and directories. All are catalogued and shelved with the collection. Textbooks are generally not collected. **DVDs that depict local or Florida history are collected; commercially available movies or documentaries are generally not collected.**

The Archives and Local History Collection consists of **a variety of formats, including (but not limited to)**: manuscripts, documents, photographs, negatives, slides, maps, microfilm, 16mm film, VHS video tapes, audio cassettes, vinyl records (all sizes and speeds), current and historic local city newspapers, periodicals, bound plat books, scrapbooks, and flags/banners. County, regional, and national newspapers are not collected. Vertical files of newspaper clippings and other ephemera as they apply to Boynton Beach are also kept with this collection.

Store



Typewriters stored in machine assembly room, Remington Typewriter Works, Ilion, Herkimer County, N.Y., 1911. New York State Archives. (http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/3045)

Key considerations

- **How** are you going to organize it?
- **What** are you going to store it on?
- **Where** are you going to store it?
- **How many copies** do you need?

File naming conventions

- Keep file and folder names short and descriptive
- Avoid repetitive words
- Use consistent patterns and date conventions
- Clearly label drafts and revisions
- Don't use special characters (<>/\!:@*&)

- Good
  - 2018-03-14_BoardMins_final.docx
  - Ltr_JasmineBates_03-14-18.docx
  - 2018-OpsBudget-draft3.xslx
  - NYSA_A3045-78_8678.tiff
- Bad
  - MarchMins.docx
  - DearJasmine.docx
  - Bob'sBudgetDraft.xslx
  - TypewritersOnShelves.tiff

File organization and management

- Store related content together
  - If you have network drives, encourage their use
- Keep folder hierarchy as "flat" as possible
- Get rid of easy-to-purge items
  - Rescued or recovered documents
  - Empty file folders
  - ~.tmp files
- Make decisions about what you're not keeping
  - Draft and duplicate copies
  - Same item, different file format

File formats

- When creating files, use preservation-friendly formats
  - Free/open
    - Technical documentation is readily available
  - Non-proprietary
  - Widely used
  - Can be read and edited by multiple software applications

Document and guide

- Document decisions in writing
  - File naming conventions
  - Folder organization
  - Acceptable formats
- Train your colleagues
  - Guidance documents
  - In-person training sessions
  - One-on-one advice
  - Online guidance

Archival storage vs. backup

- Archival storage keeps content accessible for future users
  - May house any kind of digital content
  - Houses “information objects” (files plus metadata)
  - Reliable, long-term “bit-level” preservation
  - At least two copies in two separate places
- Backups keep your computer(s) working
  - May be difficult to retrieve individual files
  - Overwriting is common
  - Metadata may be scant

Archival storage options

- Local
  - You manage all hardware and software
  - May be very simple
    - C drive of processing computer (or network drive space) plus second copy offsite
  - May be very sophisticated
    - Array of servers and networked storage housing terabytes of data in multiple locations
- Institutional partnership
  - MetaArchive Cooperative
- Cloud
  - Storage
    - Amazon Web Services
    - Microsoft Azure
    - Others
  - Digital preservation services
    - ArchivesDirect
    - Preservica

Factors that shape storage decisions

- Immediate costs
  - Number of files you have
  - Number of copies you need
  - Storage media
- Available resources
  - Budget
  - Staff knowledge and skills
- Institutional constraints
  - Legal restrictions
  - Organizational bylaws

Create multiple copies

- Minimum: two copies
  - Primary copy kept onsite
  - Secondary copy kept offsite
- Additional copies in additional locations mitigate risk
  - The greater the distance, the lower the risk of catastrophic loss

Keep in mind

- Pick the best option available to you now
- You will need to revisit your decisions every 3-10 years
  - Your circumstances will change
    - Collections grow
    - Resource availability evolves
  - Technology will change
    - Hardware, software, and storage media must be replaced periodically
    - New services and best practices will emerge

Protect

Posted notice explaining New York State laws and regulations relating to forest fires, circa 1905. New York State Archives. (http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/3746)

What are you protecting?

- Hardware and other equipment
- Digital content
- Information within digital content
  - May include "sensitive" or restricted information
- Your investment of time and money
- Your reputation

Identify and evaluate risks

- Risks to digital content
  - Fragility and unpredictability of hardware and storage media
  - Inadequate security practices
  - Technological obsolescence
  - Human error
  - Non-compliance with organizational policy
  - Physical disasters
  - Natural disasters
- Incorporate findings into your disaster response plan

Everyday protection

- Keep multiple copies
  - In multiple locations
- Know where your content is located
  - Onsite, offsite, online, offline
- Know who can access your content
  - Archives staff, IT staff, others?
  - If possible, establish permissions
- Know who can access sensitive/restricted information
  - Staff, creators, users
- Keep hardware and portable media away from known hazards
  - Basements and attics
  - Leaky pipes
  - Windows
- Identify priority records
  - Loss would greatly affect operations
  - Loss would damage reputation
- Train staff
  - Policies and procedures

Manage

Staff of the Auditing and Accounting Department, Division of Lands and Forests, New York State Conservation Commission, Albany, N.Y., April 20, 1924. New York State Archives. (http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/8728)

Developing a sense of direction

- Assessment
  - Where are we now?
- Planning
  - Where should we be going?
- Need to focus on
  - Staff
  - Program
  - Technology

Staff

- What skills do we (I) have?
- What skills do we (I) need to develop?
- How do we (I) develop these skills?

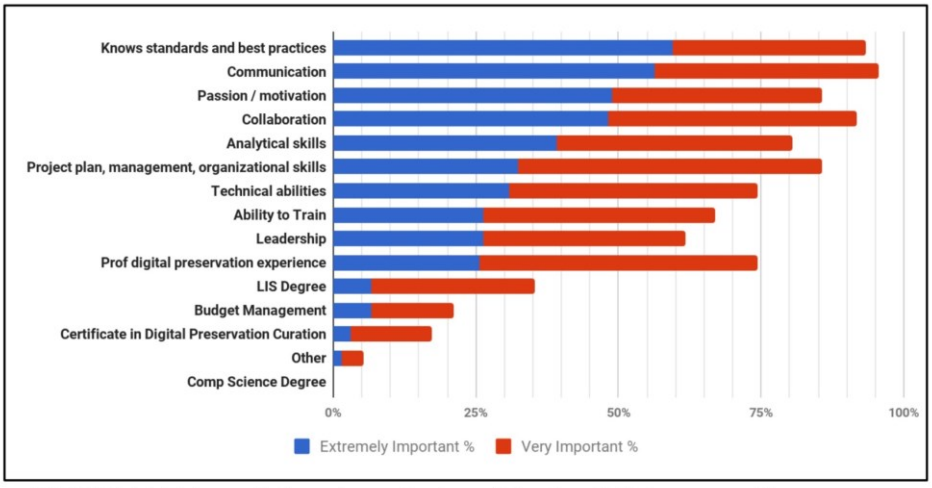


FIGURE 16: Extremely/very important qualities for digital preservation manager

National Digital Stewardship Alliance, Staffing for Effective Digital Preservation 2017 (http://ndsa.org/documents/Report_2017DigitalPreservationStaffingSurvey.pdf)

If you don't have money . . .

- Can you spare some time?
  - Key theoretical readings
  - Free webinars and online tutorials
- If you're a supervisor or manager
  - Make ongoing staff development an expectation
  - Support staff as much as you can

Program

- Digital content incorporated into
  - Collection development policy
  - Disaster response plan
  - Organizational records management policy
- Senior management made aware
  - Importance of preserving digital content
  - Resource implications
- Next steps
  - National Digital Stewardship Alliance, Levels of Preservation

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies

National Digital Stewardship Alliance, "Levels of Digital Preservation," http://ndsa.org/activities/levels-of-digital-preservation/

Technology

- Storage needs
  - How much do we currently have?
  - How much do we anticipate creating or acquiring?
- IT expertise
  - How much support do we have?
  - How much support can we expect?
- Budget
  - How much money do we have?
  - How much money do we expect to have?
  - How do we make the best use of our funding?

Wrap up

- **Identify** digital content you have (or will have)
- **Select** content that warrants preservation
- **Store** content that warrants preservation
- **Protect** content from loss and harm
- **Manage** content over the long term

Questions?

Administration of Regents examination questions and answers, Albany, N.Y., 1915. New York State Archives.

http://digitalcollections.archives.nysed.gov/index.php/Detail/Object/Show/object_id/545

Contact

Bonnie Weddle
Coordinator, Electronic Records
New York State Archives
9D64 Cultural Education Center
Albany, NY 12230
518-473-4258
bonita.weddle@nysed.gov

DHPS | NY
Documentary Heritage & Preservation Services
for New York

Questions?

DHPSNY staff is available to answer your questions.
Contact us at **info@dhpsny.org** or **(215) 545-0613**.

Connect with us!

facebook.com/dhpsny
@dhpsny
@dhpsny