Preservation in a Digital World

Tools and best practices for digital preservation, born digital records, and web archiving
Processing Born-Digital Archival Records

Brandon Pieczko
Processing and Digital Archivist
Richard B. Russell Library for Political Research and Studies
University of Georgia Libraries
General Principles for Processing Digital Records

- Don’t wait to transfer digital records to the archives
- Do no harm
- Transfer content off storage media asap
- Implement scalable procedures using tools that can be easily replaced
- Digital archival records are still records
- Document everything
Documentation (for humans and computers)

<table>
<thead>
<tr>
<th>Collection</th>
<th>Accession</th>
<th>Date</th>
<th>Media Identifier</th>
<th>Action</th>
<th>Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-02-22</td>
<td>RBRL.462.F.001</td>
<td>Files received as email attachments from donor. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-06</td>
<td>RBRL.462.CD.001</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-06</td>
<td>RBRL.462.CD.001</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.002</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.002</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.004</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.005</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.005</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.006</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.007</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
<tr>
<td>RBRL.462</td>
<td>2019-07-ER</td>
<td>2019-03-07</td>
<td>RBRL.462.CD.008</td>
<td>Virus check with Trend Micro OfficeScan. No security risks were detected. Brandon Pieczko</td>
<td></td>
</tr>
</tbody>
</table>

Contents

1.0 Introduction ........................................ 1
1.1 Purpose ............................................. 1
1.2 Approach ........................................... 1

2.0 Accessioning & Ingest Procedures ................. 3
2.1 Create Accession Record ................................ 3
2.2 Ensure that Data is Not Written to Transfer Media ........................................ 3
2.3 Transfer Files to Archives Working Space ................. 4
2.4 Create Inventory of Accessioned Files .............. 5
2.5 Transfer Files from Optical Media ............... 5
2.6 Create Disk Image of Storage Media ............... 6
2.7 Link Ingested Files and Metadata to Accession Record ............... 8

3.0 Arrangement & Description Procedures ........... 8
3.1 Create Access Copy of Ingested Files ............... 8
3.2 Export Access Files from Disk Image(s) .......... 9
3.3 Normalize Access Files to Predetermined Preservation Formats ............... 10
3.4 Arrange and Rename Access Files as Needed ............... 10
3.5 Create Resource and Digital Object Records ............... 11

4.0 Appendix ............................................ 12
4.1 Appendix A: Sample Accession Record in Archivist's Toolkit ............... 12
4.2 Appendix B: Sample Accession Folder in Collection Files ............... 14
1. Locate digital storage media in hybrid collections

- Inventory and physically separate digital storage media
  - Assign unique identifier (e.g. rbrl-444-cd-001)
  - Label disk and set aside for additional processing
  - Create separation sheet (if necessary)
2. Implement read-only protection and scan for viruses

https://www.cru-inc.com/products/wiebetechnologies/usb_writeblocker
3. Copy files off storage media and extract metadata.

Data Accessioner  [http://dataaccessioner.org](http://dataaccessioner.org)
4. Create disk image of storage media (if necessary)

FTK Imager
https://accessdata.com/product-download/ftk-imager-version-4.2.0
5. Create file manifest (inventory)

<table>
<thead>
<tr>
<th>directory path</th>
<th>file name</th>
<th>last modified</th>
<th>size (bytes)</th>
<th>md5</th>
<th>file format</th>
<th>format version</th>
</tr>
</thead>
<tbody>
<tr>
<td>mss292_disk001/</td>
<td>CHART</td>
<td>1997-06-05T23:15:28.000</td>
<td>13777</td>
<td>31ac6ff0e9019efddac4db9f0826668b9</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>COLDWAT7</td>
<td>1997-09-18T09:01:20.000</td>
<td>130298</td>
<td>e37b6530fcd71864a3118962a8ebf5ff</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>ESSAYS1</td>
<td>1995-05-16T02:18:00.000</td>
<td>113955</td>
<td>e2a1fe0a85e57dec824f9f54894ef79</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>ESSAYS2</td>
<td>1996-10-12T21:03:36.000</td>
<td>130242</td>
<td>2a2040d91126a27e2229c38b215ca4</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>ESSAYS3</td>
<td>2001-04-18T08:10:44.000</td>
<td>121916</td>
<td>205ce18111b7c23272e4116bad102f46e</td>
<td>WordPerfect Document</td>
<td>6</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>ESSAYS4</td>
<td>1995-08-29T21:06:12.000</td>
<td>107690</td>
<td>574bd34fd1ad0149f4ff312e2ff4ab4</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>ESSAYS5</td>
<td>1997-06-28T20:13:18.000</td>
<td>183295</td>
<td>0a33ef7f7590a9f37cd14f5fed35b25</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>ESSAYS6</td>
<td>1997-09-06T07:41:42.000</td>
<td>157665</td>
<td>Sdfeai129b5c08ec37d41cc3bd64319d</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>LIBRARY</td>
<td>1997-06-02T22:48:56.000</td>
<td>54753</td>
<td>7123a475d51a002362fb44f20707c113</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>MED-ED</td>
<td>1997-04-04T19:29:54.000</td>
<td>61739</td>
<td>4ab8fd1699f14fa14d356d29dfc17bc</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>NEIGHBO</td>
<td>1995-10-24T01:25:08.000</td>
<td>13227</td>
<td>8a291020a602d76e799795741f7e7985</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>OBITCV</td>
<td>1997-08-09T07:34:10.000</td>
<td>18947</td>
<td>cd6122f15382148eb3ed18ddbe6f38</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>PERSLET1</td>
<td>1993-12-18T10:21:10.000</td>
<td>73873</td>
<td>7439d609a9420f1a096b5c3759c4697</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>PERSLET2</td>
<td>1992-12-30T06:56:28.000</td>
<td>81637</td>
<td>567c9baed327fbeb869a292f975a9</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>PERSLET3</td>
<td>1994-09-15T08:06:58.000</td>
<td>77101</td>
<td>cc278e6674873b1edf5e37478f3c8333</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>PERSLET4</td>
<td>1994-01-06T09:31:14.000</td>
<td>47693</td>
<td>30019999e71f0207a7e548094f9e845e</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>PERSLET5</td>
<td>1996-01-31T18:35:34.000</td>
<td>465600</td>
<td>e328a0f3691e788df71945eb87611c77</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
<tr>
<td>mss292_disk001/</td>
<td>PERSLET6</td>
<td>1995-02-18T18:48:28.000</td>
<td>68391</td>
<td>12239b7bff1c7bea6e033f76a0e4ac0c</td>
<td>WordPerfect Document</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Example of a file manifest (text file with comma separated values)
5. Create file manifest (inventory)

Karen’s Directory Printer
https://www.karenware.com/powertools/karens-directory-printer
6. Arrange and rename the files (if necessary)

ReNamer [https://www.den4b.com/products/renamer](https://www.den4b.com/products/renamer)
7. Incorporate descriptions into the finding aid

### ER 4: Abstracts Published Versions [digital files], 2013-2014

**Scope and Content:** View an inventory of this folder online

<table>
<thead>
<tr>
<th>Roll</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EB 2013 Poster</td>
</tr>
<tr>
<td>2</td>
<td>EB 728.19</td>
</tr>
<tr>
<td>3</td>
<td>EB 728.20</td>
</tr>
<tr>
<td>4</td>
<td>EB Poster</td>
</tr>
</tbody>
</table>

### ER 5: Posters [digital files], 2013-2018

**Scope and Content:** View an inventory of this folder online

<table>
<thead>
<tr>
<th>Roll</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8. Establish level of access for the digital records

Some options:

▪ Provide online descriptions of digital files using manifests
▪ Provide on-site only access in the archives
▪ Make digital files downloadable using file sharing application (e.g. Dropbox, Google Drive, FTP)
▪ Publish select digital records online in content management system
What is web archiving?

Web archiving essentially is collecting, preserving, and enabling access to materials available on the worldwide web.
How is web archiving done and why are institutions doing it?
Web Crawlers

Web crawler does the work to gather code from a live site into an archival format. Then a rendering tool must be used in order to see the code as it existed online, archival replay tools.

The standard format of these large data files is WARC

Heritrix--developed by the Internet Archive and freely available to use

HTTrack--developed by Xavier Roche and also freely available

Wget--formerly Geturl, a part of the GNU project
More Tools

- Archive-It
- Hanzo Archives
- OCLC WebHarvester
- Webrecorder
- Web Curator Tool
- Documenting the Now

API tools for Social Media capture

- Social Feed Manager
- Twarc
- ArchiveSocial
- Twitter Archiving Google Sheet
Why are we “saving” websites?

Content is moving online

Institutional mandates

Documenting spontaneous events, North Bay Fires, 2017 collected by Sonoma County Library

Take a look at the NDSA survey from 2017 on web archiving.
So why is Athens-Clarke County Library Heritage & Special Collections archiving the web?
Here’s a little bit to think about

Collection development policy--Does it fit into your current one? Can you rewrite a section to include web archiving.

What are you going to collect? Social Media?

Are you going to ask for permission or forgiveness?

Metadata, what schema are you going to use? Dublin Core, which fields, how do we make it standardized when working in the wild west?

Think about your audience
How do we communicate to the public what we are actually doing and why we are doing it?
Bibliography

Digital Preservation Basics

Mary Willoughby
Digital Conversion and Curation Librarian
Digital Library of Georgia
# NDSA Levels of Preservation

## Table 1: Version 1 of the Levels of Digital Preservation

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protect your data</td>
<td>Know your data</td>
<td>Monitor your data</td>
<td>Repair your data</td>
</tr>
<tr>
<td><strong>Storage and Geographic Location</strong></td>
<td><strong>File Integrity</strong></td>
<td><strong>Information Security</strong></td>
<td><strong>Metadata</strong></td>
</tr>
<tr>
<td>- Two complete copies that are not collocated</td>
<td>- Check file integrity on ingest if it has been provided with the content</td>
<td>- Identify who has read, write, move and delete authorization to individual files</td>
<td>- Inventory of content and its storage location</td>
</tr>
<tr>
<td>- For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system</td>
<td>- Check file integrity on ingest</td>
<td>- Restrict who has those authorizations to individual files</td>
<td>- Ensure backup and non-collocation of inventory</td>
</tr>
<tr>
<td>- At least three complete copies</td>
<td>- Use write-blockers when working with original media</td>
<td>- Maintain logs of who performed what actions on files, including deletions and preservation actions</td>
<td>- Store administrative metadata</td>
</tr>
<tr>
<td>- At least one copy in a different geographic location</td>
<td>- Monitor file format obsolescence issues</td>
<td>- Store standard technical and descriptive metadata</td>
<td>- Store transformative metadata and log events</td>
</tr>
<tr>
<td>- Document your storage system(s) and storage media and what you need to use them</td>
<td>- Check file integrity of content at fixed intervals</td>
<td>- Perform audit of logs</td>
<td>- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs</td>
</tr>
<tr>
<td>- At least three copies in geographic locations with different disaster threats</td>
<td>- Maintain logs of file integrity info; supply audit on demand</td>
<td>- Ability to detect corrupt data</td>
<td>- Inventory of file formats in use</td>
</tr>
<tr>
<td>- Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems</td>
<td>- Ability to replace/repair corrupted data</td>
<td>- - Ability to delete files</td>
<td>- Monitor file format obsolescence issues</td>
</tr>
<tr>
<td>- Check file integrity of all content in response to specific events or activities</td>
<td>- Ensure no one person has write access to all copies</td>
<td>- Perform format migrations, emulation and similar activities as needed</td>
<td>- - Check file integrity of all content in response to specific events or activities</td>
</tr>
</tbody>
</table>

NDSA Levels of Preservation

● Tool to help evaluate status and structure decision making in 5 key areas:
  ○ Storage and Geographic Location
  ○ File Fixity and Data Integrity
  ○ Information Security
  ○ Metadata
  ○ File Formats

● Stay tuned for revision coming in 2019
● https://ndsa.org//activities/levels-of-digital-preservation/
Storage and Geographic Location

NDSA Level 1:

- 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.

Recommendations:

- Get files off of obsolete or aging media (lookin’ at you, CDs and DVDs!).
- Don’t rely exclusively on one type of storage-- all have strengths and weaknesses. Evaluate costs and risks over time, not just what is least expensive up front.
- Cloud Storage can be a great way to get geographic dispersal, BUT... Know your provider’s policies and terms of service! How do you get your data back? What is the cost for retrieval and how are charges calculated? How is their data backed up? What happens if they go out of business?
File Fixity and Data Integrity

NDSA Level 1:

- Check file fixity on ingest if it has been provided with the content
- Create fixity info if it wasn’t provided with the content

Recommendations:

- BagIt! Bags are portable and easy to validate.
  - Multiple tools for creating and validating LC style Bags
  - GUI https://github.com/LibraryOfCongress/bagger
  - Python/Command Line https://github.com/LibraryOfCongress/bagit-python

- Many other options to calculate checksums to monitor fixity:
  - Fixity (utility) https://www.weareavp.com/products/fixity/
  - HashMyFiles (Windows) https://www.nirsoft.net/utils/hash_my_files.html
  - FITS (File Information Tool Set) https://projects.iq.harvard.edu/fits/home
Bagger
Information Security

NDSA Level 1:

- Identify who has read, write, move, and delete authorization to individual files.
- Restrict who has those authorizations to individual files.

Recommendations:

- Limit access to digital objects to necessary personnel.
- Control physical access to servers and hardware.
- Use the lowest level of access sufficient to perform a task.
Metadata

NDSA Level 1:

- Inventory of content and its storage location
- Ensure backup and non-collocation of inventory

Recommendations:

- Leverage existing descriptive and administrative metadata whenever possible.
- When possible, package digital objects and essential metadata together for long term storage (ahem, Bag them cough cough).
- Extract technical metadata from files to help guide preservation actions.
File Formats

NDSA Level 1:

- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs

Recommendations:

- Choose files based on open, published standards.
- Use file formats suggested by LOC’s Recommended Formats Statement for digitization projects to create master files. [http://www.loc.gov/preservation/resources/rfs/TOC.html](http://www.loc.gov/preservation/resources/rfs/TOC.html)
- LOC Sustainability of Digital Formats site [https://www.loc.gov/preservation/digital/formats/](https://www.loc.gov/preservation/digital/formats/)
- Identify unknown formats.

- Examine technical metadata about your files to guide preservation actions.
  - FITS
  - PRONOM format registry
More information

- *An Introduction to Digital Preservation*, Elizabeth La Beaud (Mississippi Digital Library). (DLG/HomePLACE sponsored webinar recorded 2019-03-28) https://drive.google.com/open?id=1ikXGTa6R_U8jdd6TV_RwBD8hhGXl5Z5T

- Community Owned digital Preservation Tool Registry. https://coptr.digipres.org/Main_Page

References


Thank you!

Any questions?