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)

Collection	Accession	Date	Media Identifier	Action	Staff
RBRL.462	2019.07.ER	2019-02-22	RBRL.462.F.001	Files received as email attachments from donor.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-02-22	RBRL.462.F.001	Copied files to accession folder on G:\Russell ERecords\Preservation Copies.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-06	RBRL.462.CD.001	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-06	RBRL.462.CD.001	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.002	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.002	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.003	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.003	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.004	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.004	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.005	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.005	Blank disc, no content. Removed CD from collection and destroyed.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.006	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.006	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.007	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.007	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.008	Virus check with Trend Micro OfficeScan. No security risks were detected.	Brandon Pieczko
RBRL.462	2019.07.ER	2019-03-07	RBRL.462.CD.008	Copied files to accession folder on G:\Russell ERecords\Preservation Copies with Data Accessioner v1.1. No errors were detected.	

Contents


```
<collection xmlns="http://dataaccessioner.org/schema/dda-1-1" name="Senator Andre Lauren Benjamin Papers">
▼<accession number="2017.001">
 ▼<ingest_note>
    Senator Andre Lauren Benjamin Papers transferred by Brandon Pieczko on Tue Apr 04 13:54:15 EDT 2017
  </ingest_note>
  <ingest_time>00:01:07.67270</ingest_time>
  <source_note>3.5" Sony HD floppy disk</source_note>
 ▼<folder name="mss292_disk001" last_modified="1600-12-31T19:00:00.000">
  ▼<file name="CHART" last_modified="1997-06-05T23:15:28.000" size="13777" MD5="31ac6ff0e9019efddac4db9f
   ▼<premis:object xmlns:premis="info:lc/xmlns/premis-v2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-i
    xsi:type="premis:file">
    ▼<premis:objectIdentifier>
       <premis:objectIdentifierType>uuid</premis:objectIdentifierType>
       <premis:objectIdentifierValue>bd3d6cff-2353-4738-9986-cb0374d892b6</premis:objectIdentifierValue
     </premis:objectIdentifier>
    ▼<premis:objectCharacteristics>
       <premis:compositionLevel>0</premis:compositionLevel>
      ▼<premis:fixity>
         <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
         <premis:messageDigest>31ac6ff0e9019efddac4db9f082668b9</premis:messageDigest>
         <premis:messageDigestOriginator>OIS File Information</premis:messageDigestOriginator>
       </premis:fixity>
       <premis:size>13777</premis:size>
      ▼<premis:format>
       ▼<premis:formatDesignation>
          <premis:formatName>WordPerfect Document</premis:formatName>
          <premis:formatVersion>5.1</premis:formatVersion>
        </premis:formatDesignation>
```

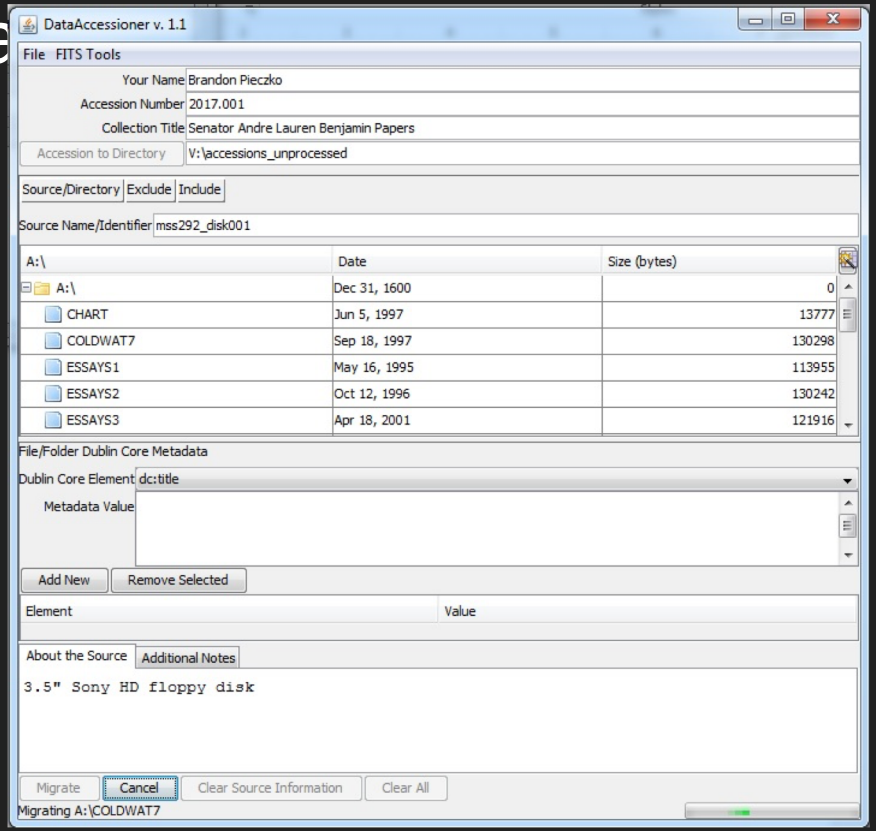
1. Locate digital storage media in hybrid collections

- 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

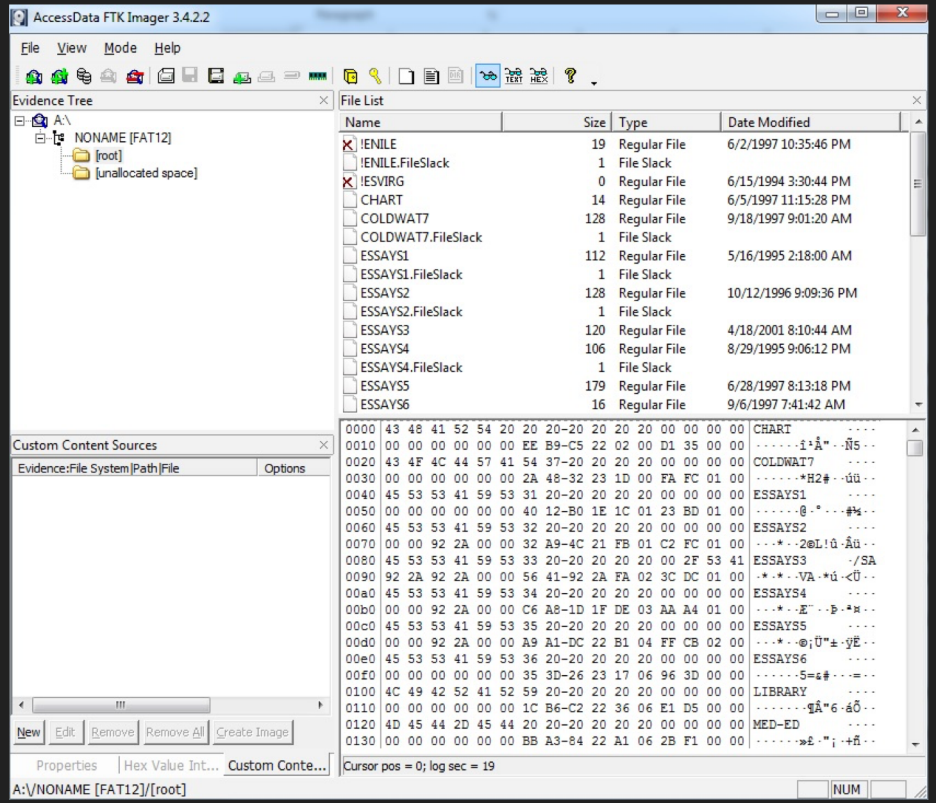
https://www.cru-inc.com/products/wiebetech/usb_writeblocker/

3. Copy files off storage media and extract metadata



Data Accessioner http://dataaccessioner.org

4. Create disk image of storage media (if necessary)



FTK Imager

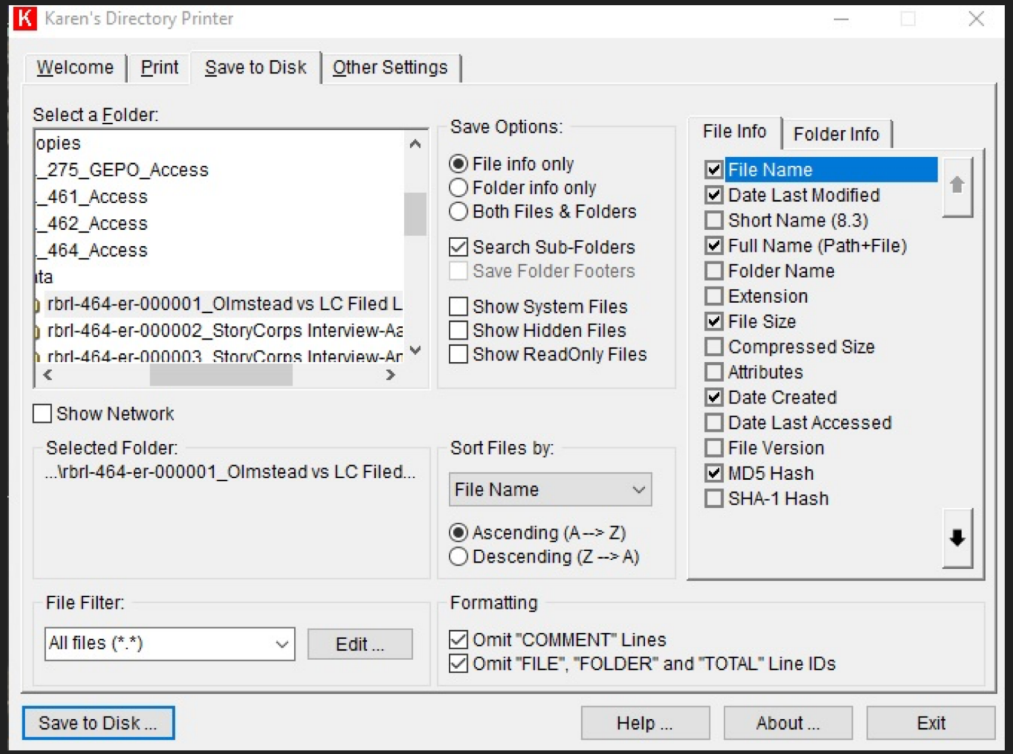
https://accessdata.com/product-download/ftk-imager-version-

5. Create file manifest (inventory)

directory path	file name	last modified	size (bytes)	md5	file format	format version
mss292_disk001/	CHART	1997-06-05T23:15:28.000	13777	31ac6ff0e9019efddac4db9f082668b9	WordPerfect Document	5.1
mss292_disk001/	COLDWAT7	1997-09-18T09:01:20.000	130298	e37b6530fdc71864a3118962a8ebf5ff	WordPerfect Document	5.1
mss292_disk001/	ESSAYS1	1995-05-16T02:18:00.000	113955	e2a1fe0a85e57dec824f99f54894ef79	WordPerfect Document	5.1
mss292_disk001/	ESSAYS2	1996-10-12T21:09:36.000	130242	2a2040d911286a27e2229c38bab15ca4	WordPerfect Document	5.1
mss292_disk001/	ESSAYS3	2001-04-18T08:10:44.000	121916	205ce18111b7c2327e411b6ad012f46e	WordPerfect Document	6
mss292_disk001/	ESSAYS4	1995-08-29T21:06:12.000	107690	574bd34f1dac0149f4ff312e2fff4ab4	WordPerfect Document	5.1
mss292_disk001/	ESSAYS5	1997-06-28T20:13:18.000	183295	0a33ef7f7590af937cd14f5fed35b25b	WordPerfect Document	5.1
mss292_disk001/	ESSAYS6	1997-09-06T07:41:42.000	15766	5dfea129b5c0e8cc37d413ccbd64319d	WordPerfect Document	5.1
mss292_disk001/	LIBRARY	1997-06-02T22:48:56.000	54753	7123a475d51a002362fb442ff07cc113	WordPerfect Document	5.1
mss292_disk001/	MED-ED	1997-04-04T19:29:54.000	61739	4ab8fd1699f14fa14d356d29df1c17bc	WordPerfect Document	5.1
mss292_disk001/	NEIGHBO	1995-10-24T01:25:08.000	13227	8a291020a602d76e799795741fce7985	WordPerfect Document	5.1
mss292_disk001/	OBITCV	1997-08-09T07:34:10.000	18947	cd6122f15382148eb3eadd18ddbe0f38	WordPerfect Document	5.1
mss292_disk001/	PERSLET1	1993-12-18T10:21:10.000	73873	7439d6f09a9420f1a096b5c375c94697	WordPerfect Document	5.1
mss292_disk001/	PERSLET2	1992-12-30T06:56:28.000	81637	567c9baed327ffbea869a292fae775a9	WordPerfect Document	5.1
mss292_disk001/	PERSLET3	1994-09-15T08:06:58.000	77101	cc278e66748731be7d5e37478f3c8333	WordPerfect Document	5.1
mss292_disk001/	PERSLET4	1994-01-06T09:31:14.000	47693	30019999e71f0207a7e548094fea845e	WordPerfect Document	5.1
mss292_disk001/	PERSLET5	1996-01-31T18:35:34.000	46560	e328a0f3691e788d7f1945eb87611c77	WordPerfect Document	5.1
mss292_disk001/	PERSLET6	1995-02-18T18:48:28.000	68391	12239b7bff1c7bea6e033f76a0e4ac0c	WordPerfect Document	5.1

Example of a file manifest (text file with comma separated values)

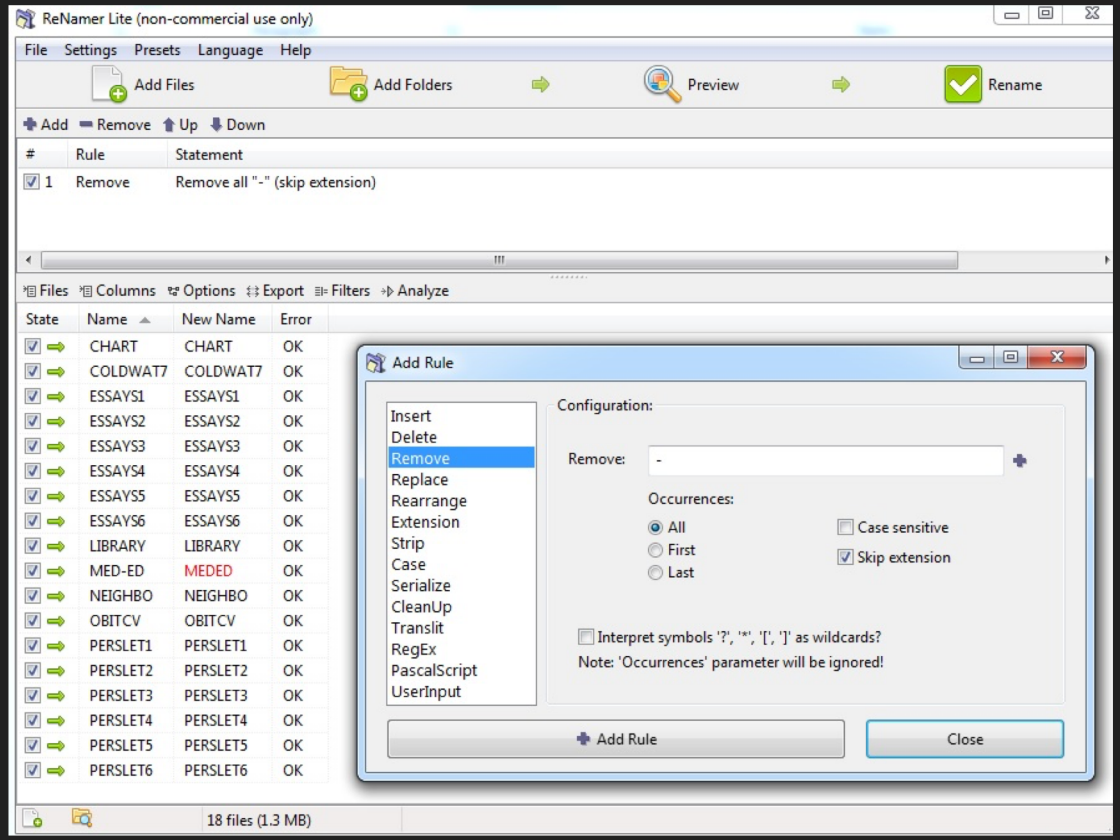
5. Create file manifest (inventory)



Karen's Directory Printer

https://www.karenware.com/powertools/karens-directory-

6. Arrange and rename the files (if necessary)



ReNamer https://www.den4b.com/products/renamer

7. Incorporate descriptions into the finding aid

Series 3: Presentation Abstracts and Posters, 2013-2018

Scope and Content: Contains abstracts for case studies published in *The FASEB [Federation of American Societies for Experimental Biology] Journal* between 2013 and 2014, and posters created by Nancy Canolty and Marcus Jennings for presentations they gave at conferences including the American Association of Anatomists and the Salk/Fondation Ipsen/Science Symposium on Biological Complexity between 2013 and 2018.

Roll

1 EB 2013 Poster

Roll

2 EB 728.19

Roll

3 EB 728.20

Roll

4 EB Post

Digital

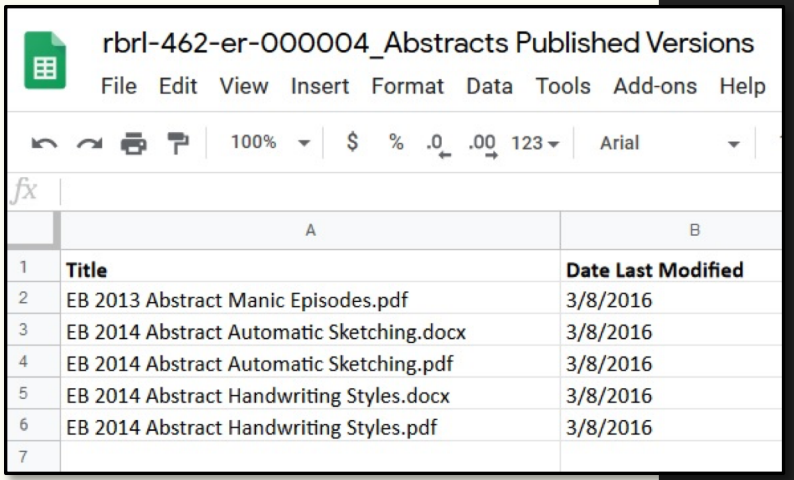
ER 4 Abstracts Published Versions [digital files], 2013-2014

Scope and Content: View an inventory of this folder online

Digital

ER 5 Posters [digital files], 2013-2018

Scope and Content: View an inventory of this folder online



rbrl-462-er-000004_Abstracts Published Versions

File Edit View Insert Format Data Tools Add-ons Help

	A	B
1	**Title**	**Date Last Modified**
2	EB 2013 Abstract Manic Episodes.pdf	3/8/2016
3	EB 2014 Abstract Automatic Sketching.docx	3/8/2016
4	EB 2014 Abstract Automatic Sketching.pdf	3/8/2016
5	EB 2014 Abstract Handwriting Styles.docx	3/8/2016
6	EB 2014 Abstract Handwriting Styles.pdf	3/8/2016
7		

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

Web Archiving

Ashley Shull
Archives and Special Collections Coordinator
Athens-Clarke County Library

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

Getting the word out

How do we communicate to the public what we are actually doing and why we are doing it?

Bibliography

- Lyman, P. (2002). "Archiving the World Wide Web". *Building a National Strategy for Preservation: Issues in Digital Media Archiving*.
- NDSA (2018). Contributor Katherine Kim. Web Archiving in the United States_A 2017 Survey: *Open Science Framework*.

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

	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
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

https://ndsa.org//activities/levels-of-digital-preservation/

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:

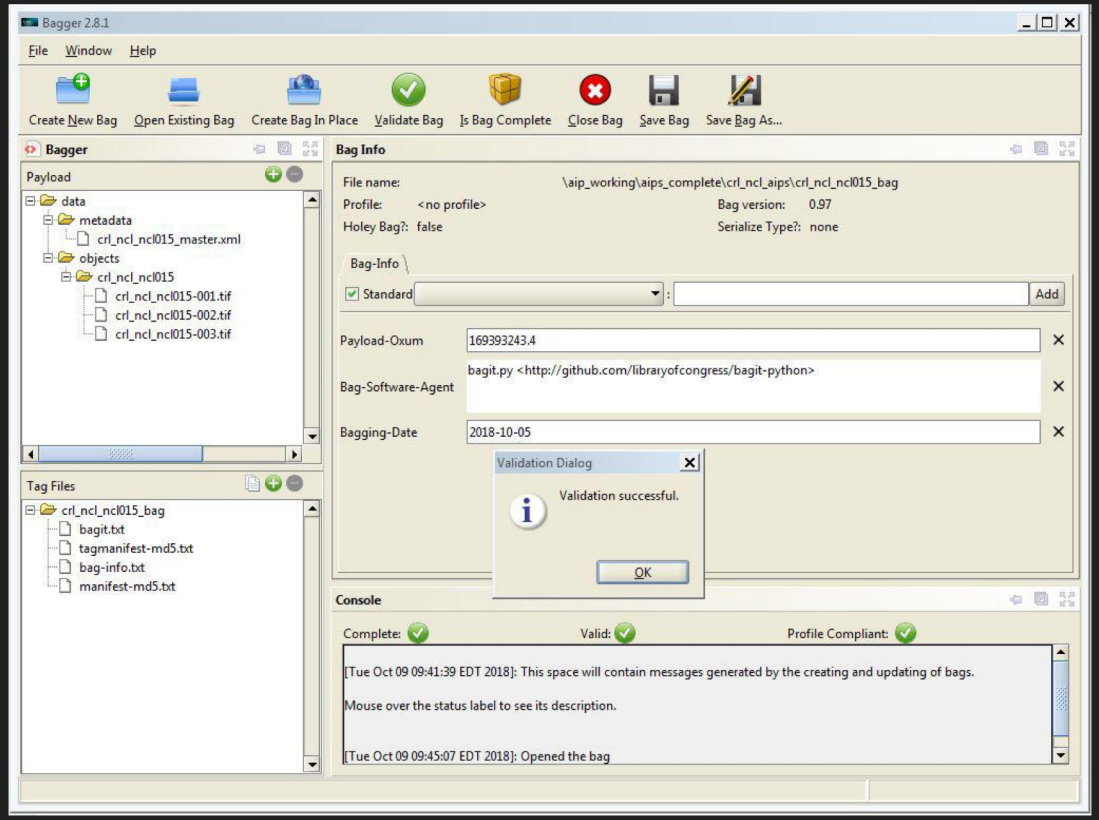
- Check file fixity on ingest if it has been provided with the content
- Create fixity info if it wasn't provided with the content

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


Bagger 2.8.1
File Window Help
Create New Bag
Open Existing Bag
Create Bag In Place
Validate Bag
Is Bag Complete
Close Bag
Save Bag
Save Bag As...
Bagger
Payload
data
metadata
crl_ncl_ncl015_master.xml
objects
crl_ncl_ncl015
crl_ncl_ncl015-001.tif
crl_ncl_ncl015-002.tif
crl_ncl_ncl015-003.tif
Tag Files
crl_ncl_ncl015_bag
bagit.txt
tagmanifest-md5.txt
bag-info.txt
manifest-md5.txt
Bag Info
File name: \aip_working\aips_complete\crl_ncl_aips\crl_ncl_ncl015_bag
Profile: <no profile>
Bag version: 0.97
Holey Bag?: false
Serialize Type?: none
Bag-Info
Standard
Add
Payload-Oxum 169393243.4
Bag-Software-Agent bagit.py <http://github.com/libraryofcongress/bagit-python>
Bagging-Date 2018-10-05
Validation Dialog
Validation successful.
OK
Console
Complete:
Valid:
Profile Compliant:
[Tue Oct 09 09:41:39 EDT 2018]: This space will contain messages generated by the creating and updating of bags.
Mouse over the status label to see its description.
[Tue Oct 09 09:45:07 EDT 2018]: Opened the bag

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
- LOC Sustainability of Digital Formats site 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

- Digital Preservation at the Library of Congress**.** https://www.loc.gov/preservation/digital/

References

BagIt Python. (2019, April 04). Retrieved July 10, 2019 from https://github.com/LibraryOfCongress/bagit-python

Bagger. Retrieved July 10, 2019 from https://github.com/LibraryOfCongress/bagger

File Information Tool Set (FITS). Retrieved July 10, 2019 from https://projects.iq.harvard.edu/fits/home

Kunze, J., Littman, J., Madden, E., Scancella, J., & Adams, C. (2018, September 17). *The BagIt file packaging format (V1.0)* [IETF Internet-Draft]. Retrieved from https://tools.ietf.org/html/draft-kunze-bagit-17

National Digital Stewardship Alliance. *Levels of digital preservation*. (n.d.). Retrieved July 10, 2019, from https://ndsa.org/activities/levels-of-digital-preservation/

PRONOM the technical registry. Retrieved July 10, 2019 from http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Thank you!

Any questions?