A Whole Lotta Preservin’ Goin’ On

By Sonoe Nakasone and Ben Pennell

https://goo.gl/WORvz5
Rough outline

1. Presentation ~ 1 hr
   - Definitions: Digital Curation, Digital Preservation, Digital Provenance
   - Digital Curation Lifecycle
   - Metadata
   - Access, Use, Discoverability
   - Carolina Digital Repository

2. Bagger Workshop ~ 1 hr

https://goo.gl/WORvz5
Sonoe Nakasone

Lead Librarian for Metadata Technologies, North Carolina State University Libraries

Former metadata librarian for the Carolina Digital Repository, University of North Carolina at Chapel Hill Libraries

https://goo.gl/WORvz5
Ben Pennell

Lead developer, Carolina Digital Repository
University of North Carolina at Chapel Hill Libraries
Zoning out is OK

Slides will be provided
Digital Curation
Digital Preservation
Digital Provenance
Digital Curation is the custodianship and management of digital objects throughout their life. The digital curation lifecycle is ongoing and iterative.
“Digital preservation combines policies, strategies and actions that ensure access to digital content over time.”

(ALCTS Preservation and Reformatting Section)
“Digital Provenance is **chronology** related to management of a digital object. Digital provenance typically describes **agents responsible** for digital objects and **key events** that occur over the course of the digital object’s life cycle” (Li and Sugimoto 2014 (Abridged))
Digital Curation Centre Curation lifecycle (modified). Original: http://www.dcc.ac.uk/resources/curation-lifecycle-model
Closer look: Digital Curation Lifecycle
conceptualization and creation

conceive and plan the creation of
digital objects, including data capture
methods and storage options

produce digital objects and assign
administrative, descriptive, structural
and technical archival metadata

Definitions: http://www.dcc.ac.uk/digital-curation/what-digital-curation
Images:
http://www.clipartkid.com/painting-board-clipart-2x6ch2-clipart/
appraise and dispose

evaluate and select for long-term preservation
dispose of objects not selected for preservation


Images:
unintentionally modifying digital objects
tools to help:
write blockers
disc images
Prof. Lopez
SSN: xxx-xx-xxxx
DOB: yyyy-mm-dd
Credit card: xxxx xxxx xxxx xxxx xxxx
Email: xxxxxx@gmail.com
Employee id: xxxxxxxxxxx
Preserve, share, and promote scholarly and creative work.
Open Archival Information System (OAIS) Reference Model

Conceptual Framework
Content Standard
Interoperability

OAIS: http://dx.doi.org/10.7207/twr14-02
SIPs, AIPs (and DIPs)

- Submission Information Packages (SIP)
- Archival Information Packages (AIP)
- Dissemination Information Packages (DIP)

Images:
http://www.clipartkid.com/dips-royalty-free-stock-image-image-12438266-ON0X2c-clipart/
http://www.freeimages.com/photo/big-ape-1386762
Submission Information Package (SIP)

myfirstbag/
|-- data
  |-- 27613-h
    |-- images
    |   |-- q172.png
    |   |-- q172.txt
    |-- manifest-md5.txt
    |   | 49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png
    |   | 408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt
|-- bagit.txt

BagIt-Version: 0.97
Tag-File-Character-Encoding: UTF-8

Archival Information Package (AIP)

Object

Structural and semantic information

Preservation Information
Descriptive information

Packaging information

Metadata

<PREMIS>
.MODS>
preserve and transform

- virus scanning
- fixity checks
- replication
- distribution
- normalization
- migration
Checksums and fixity checks

```
Microsoft Word Doc

2550 4446 2d31 2e37 0a34 3020 6f62 6a0a 2849 6e74 6974 7929 0a65 6e64 6f62 6a0a 35 0a 6f62 6a
2849 6465 6e74 6974 7929 0a65 6e64 6f62 6a0a 2841 646f 6265 29 0a 65 6e 64 6f 62 6a 0a 38 0a 30
6f62 6a 0a 3c 0a 2f 69 6c 74 65 72 20 2f 66 6c 61 74 65 44 65 63 6f 64 65 0a 2f 4c 65 6e 67 0a
39b7c937b152203bdc597dc3a10e4b2
```
Why do files change?

- Intentional change
- Accidental change
- Loss during moves and migrations
- Data degradation
Break!

5 min

Stand, stretch, pee
Metadata

Important throughout the lifecycle.

Technical

Structural

Preservation

Provenance

(Descriptive)

(Administrative)
Technical Metadata

- file properties
- creation and origin
- migration
- FITS (File Information Tool Set)
  - ex. EXIF tool to extract file headers
Structural Metadata

hierarchy, order, and other relationships between objects and object parts
Preservation Metadata Maintenance Activity (PREMIS)

- virus scanning
- fixity checks
- replication
- distribution
- normalization
- migration

<eventType>
   http://id.loc.gov/vocabulary/preservationEvents/virusCheck
</eventType>
<eventDateTime>2015-02-13T15:55:32</eventDateTime>
<eventDetail>File passed pre-ingest scan for viruses.</eventDetail>
<linkingAgentIdentifier>
   <linkingAgentIdentifierType>Name</linkingAgentIdentifierType>
   <linkingAgentIdentifierValue>
      ClamAV (ClamAV 0.98.5/20062/Thu Feb 12 17:12:36 2015)
   </linkingAgentIdentifierValue>
   <linkingAgentRole>Software</linkingAgentRole>
</linkingAgentIdentifier>
</event>
Provenance metadata

- chain of custody - Agents / responsible parties for actions
- original media location
- manifests of accessions
- deposit agreements
- PROV-DM (data Model) and PROV-O (ontology)
Putting it together...
Example workflow for a hard drive

1. Repository staff receive a hard drive with 14 TB of data.

2. A curator uses a write blocker to access the drive for a preliminary assessment.

3. Once determined in scope for the collection, the curator decides to make a disc image of the hard drive to do further appraisal and selection.

4. Once the curator has deselected out of scope materials, the objects are packaged into a SIP using Bagger and ingested into the repository.

5. The repository runs virus scans, fixity checks, and many other microservices are run on the object. All actions are recorded in the PREMIS log.

6. Periodically, the system performs preservation services. The curator reevaluates files for retention and migration needs on a 5 year schedule or as needed.
Access, Use, Discovery

access controls, encryption, redaction, versioning

rights, licensing

descriptive metadata, linking, web discoverability

sharing, harvesting, open data
access controls

redaction

versioning

encryption

access levels and roles
rights and licenses

<premis>
  <rights>
    <rightsStatement>
      <rightsBasis>copyright</rightsBasis>
      <copyrightInformation>
        <copyrightStatus>under copyright</copyrightStatus>
        <copyrightNote>copyright 1996</copyrightNote>
      </copyrightInformation>
    </rightsBasis>
  </rightsStatement>
</rights>
</premis>
descriptive metadata and discovery
sharing, harvesting, open data

Open Access

OAI-PMH

Web APIs

iiiF

Images:
An Example of Theory in Practice

A brief look at UNC Chapel Hill's digital repository

CAROLINADIGITALREPOSITORY
Preserves and provides access to born digital and at risk materials, including:

- Special collections materials
- Institutional repository
- Open Access Materials
- Content requiring long-term preservation
The repository provides:

- Preservation activities
- Arrangement and description tools
- Ingest of SIPS
- Flexible access controls
Implementation

Locally developed software

Using open source and community written tools

Built on top of a Fedora Repository
A couple of trends...

**Open Access**

Open Access Mandates  
Open Educational Resources  
Identifiers (ex. ORCID, UNF, GRID)

Tips:  
[UNC Open Access Policy](#)  
[OER Commons](#)  
[NCSU Alt-textbooks](#)  
[GRID](#)

**Researcher Engagement**

Research data  
Data Management Plans (DMPs)  
Data repositories

Tips:  
[Data Snafu](#)  
[DMP Tool](#)  
[NCSU Libraries Data Management](#)  
[Figshare](#)  
[SHARE](#)
Mahalo for listening
Resources

How to use Bulk Extractor for Personally Identifying Information

BitCurator digital forensics software and information

PREMIS

Open Archives Information Systems (OAIS) Reference Model

Using PROV with OAIS and PREMIS

Pinecone: open source software for long term, short-term storage

DuraSpace technologies

AP Trust

Digital Preservation Network (DPN)

Archivematica

Conferences:

National Digital Stewardship Alliance’s Digital Libraries Federation Conference

Open Repositories


Digital Preservation Definition:


Digital Curation Definition:

Bagger Workshop

with Ben Pennell
What's a bag?

A specific directory structure, used for transmission or preservation. Containing:

- The things you wanted shipped
- Information about the things that were shipped
Mixing up some metaphors

Amazon Boxes and Bagit Bags
What's a bag have to offer?

- It's standard - other people or systems can understand it
- Can tell if there are missing or unexpected files
- Detect changed files with checksums
Downloading Bagger

Go Here: https://github.com/LibraryOfCongress/bagger

Click on “Releases.” Download version 2.7.3

Or just go here: http://bit.ly/2ikjgXC

If you run into problems…

1) Make sure Java is installed. If not, install Java.
2) Make sure home path variable is set. Find your advanced system settings, navigate to environment. Add path and set home variable.
Running Bagger

1. Go to the location where you downloaded Bagger to

2. Go into the "bin" folder inside of it

3. Double click bagger.bat
Example bag

- Go to [https://goo.gl/i87s3L](https://goo.gl/i87s3L)
- Click the download icon in the top right corner
- Unzip treasures.zip where you downloaded it
- In Bagger, click the Open Existing Bag button
- Select the unzipped "treasures" folder and click Open.
<table>
<thead>
<tr>
<th><strong>Bag Info</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>File name:</strong> /Users/bbpennel/Desktop/treasures</td>
</tr>
<tr>
<td><strong>Profile:</strong> &lt;no profile&gt;</td>
</tr>
<tr>
<td><strong>Bag version:</strong> 0.97</td>
</tr>
<tr>
<td><strong>Holey Bag:</strong> false</td>
</tr>
<tr>
<td><strong>Serialize Type:</strong> none</td>
</tr>
</tbody>
</table>

**Bag Info**

- **Payload-Oxum**: 966539.7
- **Bagging-Date**: 2017-01-03
- **Organization**: University of Hawai‘i
- **Bag-Size**: 944.1 KB
- **Title**: Treasures from the University of Hawai‘i Library
Bag structure
Bag inspection (TSA free!)

- In your file browser, go into the "treasures" folder
- Double click on manifest-md5.txt
- Double click on tagmanifest-md5.txt
- Double click on bag-info.txt
- Look inside the "data" folder
Bag validation

There are two types of validation in Bagger

- **Valid Bag** - Checks that all payload and tag files are present, have the right file sizes, there are no extra files, and performs a fixity check.
- **Complete** - Checks that all payload and tag files are present, there are no extra files, and they all have the correct file size.
Validation Experimentation

● Click on "Validate Bag" and "Is Bag Complete", observe results
● Open treasures.txt in Notepad, make a minor change to it and save
  ○ Click "Validate Bag" and "Is Bag Complete", what are the results?
● Delete treasures.txt
  ○ Click "Validate Bag" and "Is Bag Complete", what are the results?
Make a bag in place

- Download https://goo.gl/uyWyi2
- Unzip small-treasures.zip
- View the contents of the unzipped "small-treasures" folder
- In Bagger, click Create Bag In Place
- Select the "small-treasures" unzipped folder
- View the contents of the "small-treasures" folder again
Metadataing

In the "Bag Info" panel, try adding yourself as the author (hint, there is a checkbox and an add button)

Click on the "Save Bag" button to update the bag

You can view your new metadata by looking in bag-info.txt
Creating a totally new bag

1. Click Create New Bag
2. Click okay with no profile selected
3. Add some new files or folders to the bag
4. Save the bag
5. "Save in" - the name you provide will be the name of the bag
6. Serialize Type - Save the bag as a folder or zip file
7. Manifests - Whether to include or not, what kind of checksum
Other Bagger features

- Profiles - Can set up standard metadata templates to start from
- Console
- Holey Bag
Other resources
METS Manifests

Sections

Header
Descriptive metadata
Administrative metadata
File section
Structural maps
Structural links
Behavior

Advantages over bags

Associate metadata with objects
Include information about object structures and relationships
Embed identifiers
Pinecone

https://github.com/UNC-Libraries/pinecone

ben.pennell@gmail.com

stnakaso@ncsu.edu