


Connecting Collections ONLINE COMMUNITY

Practice Safe Archiving: Backups, Copies, and What Can Go Wrong



C2C: Caring for Digital Materials
Session 4: Practice Safe Archiving
2013-04-10, Jefferson Bailey
jbailey@metro.org

Copy, by Piotrek Chuchla from The Noun Project
(TNP): Disk Copy by Simon Child, from TNP

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Caring for Digital Materials Goals

1. Participants will have a better understanding of the inherent fragility of digital objects
2. Participants will acquire information to help them select preservation formats, metadata, and backup systems for digital objects
3. Participants will be able to identify one or more actions that can be taken to improve their institution's digital preservation efforts

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Caring for Digital Materials Sessions

Overview of digital preservation	Lauren Goodley	Tues., April 2, 2013 2:00 - 3:30 EDT
Convert it to preserve it: Digitization and file conversion	Jacob Nadal	Thurs., April 4, 2013 2:00 - 3:30 EDT
Describe it so you can find it: Metadata, finding aids, and asset management	Danielle Plumer	Tues., April 9, 2013 2:00 - 3:30 EDT
Practice safe archiving: Backups, copies, and what can go wrong	Jefferson Bailey	Wed., April 10, 2013 2:00 - 3:30 EDT
Partner to preserve: Digital preservation networks and collaboration	Liz Bishoff and Tom Clareson	Mon., April 15, 2013 2:00 - 3:30 EDT

We are here →


01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Session Outline:

Part 1:
Physical Media & Digital Information

Part 2:
Backup & Storage


Part 3:
The Levels of Digital Preservation




Floppy Disk by Mike Wirth, from The Noun Project (TNP), Computer by Claudine Rodriguez, from TNP, Hard Drive by Mike Wirth, from TNP, Information by Adrian Escudero, from TNP

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110


BITS, or, what's a digital object



01100100
01101001
01100111
01101001
01110100
01100001
01101100
00100000
01110000



<?xml version="1.0" encoding="ISO-8859-1"?>
<note>
<to>You/fo
<from>Me/From</to>
<heading>Hello</heading>
<body>Hi</body>
</note>



Floppies, <http://oldcomputers.net/floppydisks.html>
Wairua, <http://www.nytimes.com/2012/10/10/nyregion/orphaned-baby-wairua-to-arrive-at-new-york-aquarium.html>
Photoshop, Author screenshot

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

What's a bit?



4um

News: <http://www.nanofar.es/applications/gallery/details.php?cat=nanofar1m4-facilities&part=1>

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

An Ontology of Digital Objects

Physical Object - "an inscription of signs on some physical medium" [on some physical device]

Logical Object - "processable units...recognized by some application software" [in some format, with some metadata]

Conceptual Object - "recognized and understood by a person, or in some cases recognized and processed by a computer application" [information we understand]

Conditional Object (my addition): Digital materials are created and managed as well as acquired, preserved, and made available within certain social, financial, and institutional conditions. [dependent on you]

- At many different types of organizations
- Among many different types of colleagues
- In many different budgetary climates
- On many different projects
- With many different types of content

Thibodeau, Kenneth. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years." *The State of Digital Preservation: An International Perspective*. CLIR Report, 2002.

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01101110 01100001 01101000 01101111 01101110

Session Outline:

Part 1:
Physical Media & Digital Information



Part 2:
Backup & Storage

Part 3:
The Levels & Other Resources

Floppy Disk by Mike Wirth, from The Mouse Project (TMP).

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01101110 01100001 01101000 01101111 01101110

Physical Object: What can go wrong?

- **Obsolescence** - Hardware, Media, File System
- **Access** - Degradation, Longevity
- **Appraisal** - Backlog, Physical & Intellectual Control
- **Authenticity** - Alteration, Corruption

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01101110 01100001 01101000 01101111 01101110

Poll: Physical Formats

- What types of physical media do you have in your institution (can be collection material or media used locally). Check all that apply:
 - 5.25" floppy disks
 - 3.5" floppy disks
 - ZIP and Jaz disks
 - SD Cards
 - Tape (audio, video, data)
 - Optical (CD-R, DVD)
 - External Hard Drives (includes thumb drives)
 - Internal Hard Drives (and/or full computers)
 - Network-attached Storage
 - Other

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01101110 01100001 01101000 01101111 01101110

Physical Object: the actions

Media Types & Longevity (in years)

- Floppy: 3-5
- Optical: 2-10
- Spinning disk: 2-8
- Flash: 1-10
- Tape: 10-30

Dependent on Environment & Handling

- CLIR, "Care and Handling of CDs & DVDs" <http://www.clir.org/pub/reports/pub121/contents.html>
- UK Archives, "Care and Handling of Removable Media" <http://www.nationalarchives.gov.uk/documents/information-management/removable-media-care.pdf>

Longevity info: <http://agooified.com/97>

Image credit: <http://www.cdn-geo.com/cdn/cdn/DigitalMediaLifeExpectancyAndCare.html>

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01101110 01100001 01101000 01101111 01101110

Physical Object: the actions

Frontline:

- Vintage Drives & Machines
- Controller Cards
- Write Blocker
- Forensic Software
- Photo Station
- FRED



Backline:

- Storage, more storage, & even more storage
- Transfer tools
- Hardware & infrastructure



Apple II: <http://technical.ly/philip/2011/10/07/original-apple-ii-developed-by-steve-jobs-used-microsoft-made-microprocessor/>; Kryoflux, http://web.archive.org/web/20060601000000/http://info.php/products_id38_write_blocker, http://commons.wikimedia.org/wiki/File:Portable_Forensic_Cablebox.jpg; FRED, www.computerforensimarket.com

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01101110 01100001 01101000 01101111 01101110

Physical Object: the actions

Obsolescence: Get Bits off Media and into Systems

Using:

- On Disk Images
- Virus Checking
- BitCurator



Digital Forensics:

- CLIR, Digital Forensics and Born-Digital Content in Cultural Heritage Collections
<http://www.clir.org/pubs/abstract/reports/pub149>
- DPC, Digital Forensics and Preservation [Technology Watch Report]
http://www.dpconline.org/component/docman/doc_download/810-dpctw12-03pdf

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Physical Object: the actions

Access:

Metadata

- Inventory - at least know what you have
- UUID! (Universal Unique Identifier)
 - i.e. consistent filenames and identifiers
 - Needed to link multi-part objects
 - Helps track change through time
- Locate - at least know where it lives (cyberspace and meatspace)
- Describe - descriptive information, at least at collection level
- Photograph - people often take photographs of donated digital content for added contextual information

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Physical Object: the actions

Appraisal & Authenticity:

- Donor Agreements - see resources
- Describe what you have & how it changes
- Backlog/Backup & Storage - cover below
- Key point: align what you acquire and what you commit to preserve with your collection policy and institutional abilities!
- Nobody can save everything!

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Questions

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Logical Object... but first!

---FIXITY---

- Fixity = numeric string = "digital fingerprint" of a file
- More accurate than DNA
- Computed by algorithm: MD5, SHA1, SHA256
- AKA checksum, hash, message digest
- Any alteration to bits leads to a new checksum
- Fixity checks, audits

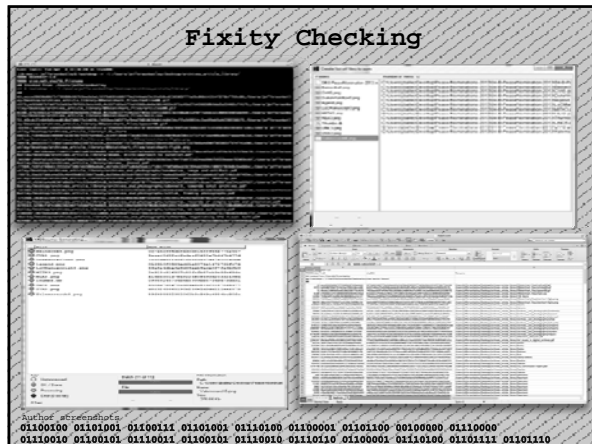
```
<fileobject>
  <filename>/Users/jeffersonbailey/Documents/fixity.txt</filename>
  <filesize>6</filesize>
  <ctime>2013-03-21T01:05:20Z</ctime>
  <mtime>2013-03-21T01:05:20Z</mtime>
  <atime>2013-03-21T01:12:13Z</atime>
  <hashdigest type='MD5'>b8887178222619ddad7b189dafdb8361</hashdigest>
  <hashdigest
    type='SHA256'>2290aaecba0445775d01a1ee0f5b55167cf457c4578f2208da0a2e412792c8
    c3</hashdigest>
  </fileobject>
```

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Poll: Fixity

- My institution generates and audits fixity information for digital content:
 - Yes
 - No
 - Don't know

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110



Physical/Logical: Fixity

Bitrot!

Some GUI tools:

- MD5 Summer & MD5 (see resources)
- Bagger (see resources)



Fixity Checking/Audit

- Scheduled check of integrity of digital content
- No visual way to tell if data has changed
- Data can corrupt (flipped bits) for intentional & unintentional reasons
- Checking fixity periodically is an essential activity to ensure data hasn't changed
- "Periodically" is an institutional decision

Image: Atlas of Digital Damages (Flickr group): <http://www.flickr.com/groups/2121762@N23/>

01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Logical Object: What Can Go Wrong?

- **Obsolescence** - Formats depend on software
- **Access** - Systems needs to know how to open a file
- **Appraisal** - Why preserve a file you can't open?
- **Authenticity** - Need to document changes to a digital object (including format)

01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Logical Object: the challenges

- "recognized and processed by software"
- Recognition and processing are independent
- Formats, wrappers, codecs
- A format is not a file extension
- Software obsolesces
- Unique appearance



Formats, <http://www.webdesignbot.com/free-icon/free-icon-set-adeable/>

01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Logical Object: the actions

- Identification: what is the object's format?
- Characterization: what are an object's characteristics?
- Validation: is the object what it says it is?
- Embedded metadata
- Working with donors to accept open formats
- Institutional policies around open formats
- Migration decisions are not so different than digitization decisions



Sustainability of Digital Formats
Planning for Library of Congress Collections

01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

The Conditional Object

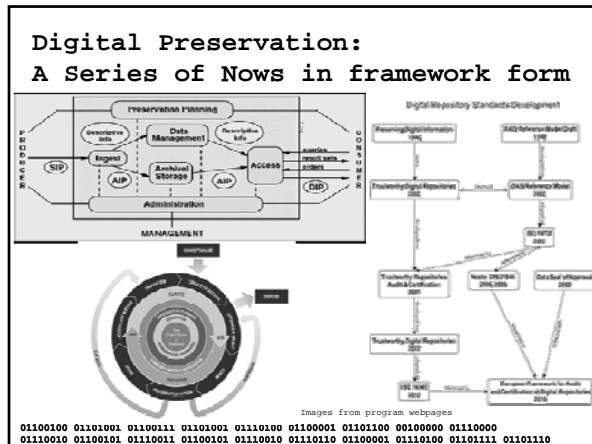
Planning Tools (see resources):

- Score Model
- DPC Decision Tree
- DRAMBORA
- OAIS & TRAC requirements

All are very wonky, but together help cover every planning and institutional question you should be asking.

Remember, every institution is different. You need to align your institutional capabilities with accepted best practices.

01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110



Session Outline:

Part 1:
Physical Media & Digital Information

Part 2:
Backup & Storage

Part 3:
The Levels & Other Resources

Icon: Hard Drive by Mike Wirth, from The Noun Project

01100100 01101001 01100111 01101001 01101100 01100001 01101100 00100000 01110000 01110010 01100101 01110011 01100101 01100110 01101110 01100001 01110100 01101111 01101110

Poll: Backups

- How many copies of digital content does your institution keep and in how many different geographic locations?
 - 1 copy, 1 place
 - 2 copies, 1 place
 - 2 copies, 2 places
 - 3 copies, 2 places
 - 4 or more copies, 3 or more places
 - Don't know

01100100 01101001 01100111 01101001 01101100 01100001 01101100 00100000 01110000 01110010 01100101 01110011 01100101 01100110 01101110 01100001 01110100 01101111 01101110

Backup

Number of Copies

- Minimum: Triple Deuces Rule
 - 2 copies, 2 places, 2 media types
- Better: 3 minimum copies, 2 places & media types
- Have inventory, location, fixity for all copies!
- Details dependent on:
 - Types of files
 - Institutional requirements
 - Institutional resources

Icon: Copy, by Piotr Chuchla from The Noun Project

01100100 01101001 01100111 01101001 01101100 01100001 01101100 00100000 01110000 01110010 01100101 01110011 01100101 01100110 01101110 01100001 01110100 01101111 01101110

Storage types

Types:

- Online
- Near-line
- Offline


Question to ask:

- How often do you access?
- Preservation copies separate from access copies?
- How are preservation & access copies created and/or managed?
- Do you systems/workflows "play nice" with other systems? With future systems?

Image: <http://venturebeat.com/2012/04/25/synform-raises-11m/>


01100100 01101001 01100111 01101001 01101100 01100001 01101100 00100000 01110000 01110010 01100101 01110011 01100101 01100110 01101110 01100001 01110100 01101111 01101110

Tech Stack



Some Terminology:

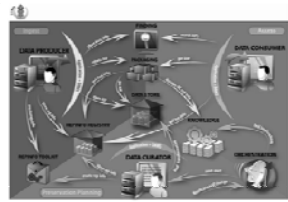
- LAMP
 - Linux, Apache, MySQL, Ps
- RAID
 - Redundant Array of Independent Disk
- NAS (Networked-Attached Storage)
 - Computer network
- SAN (Storage Area Network)
 - Storage network
- CMS (Content Management System)
 - WordPress, Drupal, Joomla
- DAMS (Digital Asset Management System)



LAMP: <http://www.montpelieropensource.com/services.htm>
RAID: <http://www.freshdy.com/2007/08/caldigit-hdpro-raid-storage-review.html>

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

System types:



Types:

- Vendor
- "Turnkey" solution
- Open-Source Software

Question to ask:

- Resources & expertise
- Requirements & needs
- Ties into existing systems
- Data in & data out

Realization:
No solution is permanent - Series of Nows!

Image: <http://www.digitalpreservationeurope.eu/video-training/prague-2008/?media=7>

01100100 01101001 01100111 01101001 01101000 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110


Digital Preservation: Series of Nows in Software Form



Images from program webpages

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Key Concepts: Backup & Storage



- Multiple Copies
- Multiple Place
- Multiple Media Types
- Unique Universal Identifies
- Inventory & Identify (what & where)
- Record & Monitor Fixity Information
- Work with IT
- Be Adaptative
- Systems Change: Data Shouldn't

Icons: Copy, by Piotrek Chuchla & Disk Copy by Simon Child from The Noun Project

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Questions


01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

Session Outline:

Part 1:
Physical Media & Digital Information

Part 2:
Backup & Storage

Part 3:
The Levels of Digital Preservation



Icon: Information by Adrian Escudero, from TNP

01100100 01101001 01100111 01101001 01101000 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01110010 01110110 01100001 01110100 01101111 01101110

NDSA Levels of Digital Preservation

What it covers:

- Concepts & process areas
- Accepted best practices & community input
- Baseline considerations, risk mitigation
- Progressive, accessible scalability
- Functional independence

What it doesn't cover:

- Institutional context
- Technology
- Activity
- Policy



Icon: Upstairs, by Jo Szczepanska from The Noun Project

01100100 01101001 01100111 01101001 01101010 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01110110 01100001 01110100 01101111 01101110

NDSA Levels of Digital Preservation

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Monitor your data)	Level 3 (Repair your data)
Storage and Geographic Location	<ul style="list-style-type: none"> - Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the contents off the medium and into your storage system 	<ul style="list-style-type: none"> - 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 	<ul style="list-style-type: none"> - At least one copy in a geographic location with a different disaster threat - Create/enclose monitoring process for your storage system(s) and media - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fidelity and Data Integrity	<ul style="list-style-type: none"> - Check file/folder integrity if it has been provided with the content - Create file/folder if it wasn't provided with the content 	<ul style="list-style-type: none"> - Check file/folder integrity on all ingests - Use write inhibitors when working with original media - Verify/Check high risk content 	<ul style="list-style-type: none"> - Check file/folder integrity at least annually - Maintain logs of file/folder integrity - Ability to reconstruct corrupted data - Verify no new errors are being introduced to all ingests
Information Security	<ul style="list-style-type: none"> - Identify who has read, write, move and delete authorization to individual files - Restrict who has these authorizations to individual files 	<ul style="list-style-type: none"> - Document access restrictions for content - Maintain logs of when performed what actions on files, including deletions and preservation actions 	
Metadata	<ul style="list-style-type: none"> - Inventory of content assets - Store administrative metadata - Store descriptive metadata and log events 	<ul style="list-style-type: none"> - Store administrative metadata - Store descriptive metadata and log events 	<ul style="list-style-type: none"> - Store standard preservation metadata - Store standard descriptive metadata
File Formats	<ul style="list-style-type: none"> - When you capture data input into the system of digital files encourage use of a limited set of known open formats and schemas 	<ul style="list-style-type: none"> - Inventory of file formats in use 	<ul style="list-style-type: none"> - Monitor file format obsolescence issues - Perform format migration, emulation and similar activities as needed

01100100 01101001 01100111 01101001 01101010 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01110110 01100001 01110100 01101111 01101110

Poll: Levels of Preservation

- What Level of Preservation level best describes your institution's currently digital preservation activity?
 - Not yet at Level 1
 - Mostly at Level 1
 - Mostly at Level 2
 - Mostly at Level 3
 - Mostly at Level 4
 - Don't know

01100100 01101001 01100111 01101001 01101010 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01110110 01100001 01110100 01101111 01101110

THANKS !

Thanks to Danielle, Kristen, & the other presenters

Contact me with any questions!

Jefferson Bailey
jbailey@metro.org
@jefferson_bail



Icon: Memory, by Andrew J. Young from The Noun Project

01100100 01101001 01100111 01101001 01101010 01100001 01101100 00100000 01110000
01110010 01100101 01110011 01100101 01100110 01110110 01100001 01110100 01101111 01101110