Follow up Information

Caring for Digital Materials:
Backups, copies, and what can go wrong

Jefferson Bailey, instructor

General Questions

Q. *If you make a copy of the original you lose some of the original do you not? Make one copy of original, work on the copy, keep original and leave it alone.*

Copying a digital file, especially with some of the special "forensic" tools, is bit-level, so it is an exact copy of the file, and the copy will have the same checksum as the original. However, opening a file in its original software and saving as a copy does introduce changes in the object (metadata mostly).

Q. *If you have a file stored on a floppy or other removable media and you save a copy to a hard drive, which would be the master or preservation copy?*

As long as you have an exact copy of the file (or the entire media device) then the preservation copy should be that one which is in a more secure & long-term preservation environment. Since physical media degrade, they would rarely be considered the “preservation copy” in the long term. The “preservation copy” is not tied to a specific media device.

Q. *What is near-line storage?*

Near-line storage is less accessible than immediate, online storage and more accessible than a “dark archive” which may take many days or weeks to fulfill a request to retrieve your data and deliver it to you. Most near-line storage would be able to fulfill a request within a day or two.

Q. *Are there resources describing the basic setup of a forensic workstation?*

Yes, there are a number of them listed in the resources:
http://www.connectingtocollections.org/courses/caring-for-digital-materials/

This blog post also provides a basic introduction:
http://mith.umd.edu/digital-curation-workstation/

Digital Preservation Tools

Q. *Can you discuss some of the digital preservation software options?*

Note: This list is not intended to be exclusive and represents questions posed during the webinar on April 10, 2013.

- BitCurator – The BitCurator project is a joint effort led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and the Maryland Institute for Technology in the Humanities (MITH) to develop a system for collecting professionals that incorporates the functionality of many digital forensics tools. You can find details and access the software at http://www.bitcurator.net/.

- Archivematica – Archivematica uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control the micro-services via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core and other best practice metadata standards. Archivematica implements format policies based on an analysis of the significant characteristics of file formats. You can find details and access the software at https://www.archivematica.org/wiki/Main_Page

- Fedora – Fedora (Flexible Extensible Digital Object Repository Architecture) provides a core repository service (exposed as web-based services with well-defined APIs). In addition, Fedora provides an array of supporting services and applications including search, OAI-PMH, messaging, administrative clients, and more. Fedora provides RDF support and the repository software is integrated with semantic triple store technology. Fedora helps ensure that digital content is durable by providing features that support digital preservation. You can find details and access the software at http://www.fedora-commons.org/

- Preservica – Preservica, from Tessella, is a vendor-supported digital preservation system. Preservica is a workflow-based application that accessions and curates content of all types. It provides services like information management, information processes, and information preservation functions. Examples include the ability to structure content, use content description (metadata), search and browse content and metadata, and provide quality-assured ingest processes. Preservica uses highly durable cloud storage with redundant copies and cloud services to underpin these capabilities. You can find details at http://www.digital-preservation.com/solution/preservica/

- LOCKSS – LOCKSS (Lot of Copies Keeps Stuff Safe) was developed as a distributed preservation system, primarily intended for electronic journals. It has developed into a robust solution that underpins other digital preservation efforts, such as MetaArchive. While the software is open source, to participate in the distributed preservation effort you must become a member of the LOCKSS Community. You can find details and access the software at http://www.lockss.org/

Storage Media

Q. *A friend at Disney told me they have a dept that rotates each of their digital animation objects to the newest media every seven years. Is there a list of expected longevity for each existing media type?*
See Agogified Blog**,** How Long Will Your Digital Storage Media Last? http://agogified.com/97

Q. *For an independent studio, how do you suggest tracking when it's time to migrate? Do you migrate in tandem with creating new materials/works? Do you plan to do massive migration at one time every 7 or so years?*

See above. Migration should be determined by physical media longevity.

Q. *How often do you check the functionality of backup storage media like DVDs?*
There's no right answer here. It should be an institutional decision. Do what is realistic for your staff, budget, and time.

Checksums

Q. *What are some recommended software tools for checksums? Do you have any recommendations for programs to generate checksums on Macs?*

Resource: http://www.mnhs.org/preserve/records/legislativerecords/carol/authentication.htm
Scroll down to the Tools section.

Wikipedia also has a good overview:
http://en.wikipedia.org/wiki/Comparison_of_file_verification_software

Q. *Will different checksum generating software always calculate identical checksums?*

Yes, specific algorithms produce the checksum so it will be the same regardless of what software is used to run the algorithm(s).

Q. *Any recommendations on software to create checksums on large amounts of data (up to 1Tb)?*

I would recommend something Unix-based, something along the lines of hashdeep.

Q. *What about longevity and/or access to the algorithm and checksum tools themselves? Will we be able to use them long-term?*

Yes, the algorithms will always be around and is not in danger of becoming obsolete. The different pieces of software are all using the same algorithms.

Fixity and Bitrot

Q. *Is it common to do the baseline check prior to bringing a digital collection (from a donor's collection) to the library/archive? Then run another check to see if anything was lost or corrupted upon ingest?*

Yes! Though this depends a bit on your accession workflow. You should generate fixity information (and make back-up copies as early as possible in your acquisitions workflow).

Q. *So if you do a fixity check you can see when something's gone wrong -- but what can you do about it? The thought that comes to my mind is keeping two copies on separate media and replacing the corrupted version with a copy of the (hopefully!) non-corrupted one, but is there any way to actually reverse the bitrot?*

Generally, you cannot reverse bitrot (and it isn't worth the effort (time/resource wise) to do so on an item-by-item basis. Yes, multiple copies, multiple place. Replace the corrupt file with a valid one and make a new back-up copy (at least one!).

Q. *Are there steps you can take to minimize the likelihood of bitrot?*

Generally, limited access to your preservation copy is the best step to take.

Cloud storage

Q. *We've been debating "cloud" versus getting a new backup hard drive. What is your basic suggestion regarding "cloud" storage?*

This is hard to answer. It depends on your institution's strategic planning and resources and expectations. They both have their advantages and disadvantages.

Q. *How reliable is cloud storage?*

Well Amazon guarantees a certain number of "nines" (meaning 99.9999 availability or something like that) but you are obviously dependent on a potentially-ephemeral corporate entity. Cloud storage should never be your **only** method of storage.

Q. *What is the safety level of "the cloud"? Can material backed up there be readily stolen?*

It depends on the service, I suppose, but the safety level is quite high for dedicated services like Amazon, etc. Security is more fraught at the access level, meaning users you give credentials to access the account.

RAID Storage

Q. *Can you explain RAID? What are the differences between software and hardware RAID?*

Wikipedia provides an acceptable answer: http://en.wikipedia.org/wiki/RAID

Q. *Can files become corrupt on a RAID storage system?*

Participant-supplied resource: Probability of RAID failures : http://www.raid-failure.com/raid10-50-60-failure.aspx

Q. *Why pdf /a as opposed to pdfx1a?*

From ftp://ftp.fu-berlin.de/tex/CTAN/macros/latex/contrib/pdfx/pdfx.pdf:

> PDF/X and PDF/A are umbrella terms used to denote several ISO standards that define different subsets of the PDF standard. The objective of PDF/X is to facilitate graphics exchange between document creator and printer and therefore, has all requirements related to printing. For instance, in PDF/X-1a, all fonts need to be embedded and all images need to be CMYK or spot colors. PDF/X-2 and PDF/X-3 accept calibrated RGB and CIELAB colors along with all other specifications of PDF/X-1a.
>
> PDF/A defines a profile for archiving PDF documents which ensures the documents can be reproduced the exact same way in years to come, a key element to achieve this is that the PDF/A documents shall be 100% self contained. All the information needed to display the document in the same manner every time is embedded in the file. A PDF/A document is not permitted to be reliant on information from external sources. Other restrictions include avoidance of audio/video content, JavaScript and encryption. Mandatory inclusion of fonts, color profile and standards based metadata are absolutely essential for PDF/A.

Other questions? Email info@heritagepreservation.org and include "Caring for Digital Materials" in the subject. Or join the community at http://www.connectingtocollections.org/ and ask your questions there!