The Quest for Digital Preservation

Will a portion of Math History be lost forever?

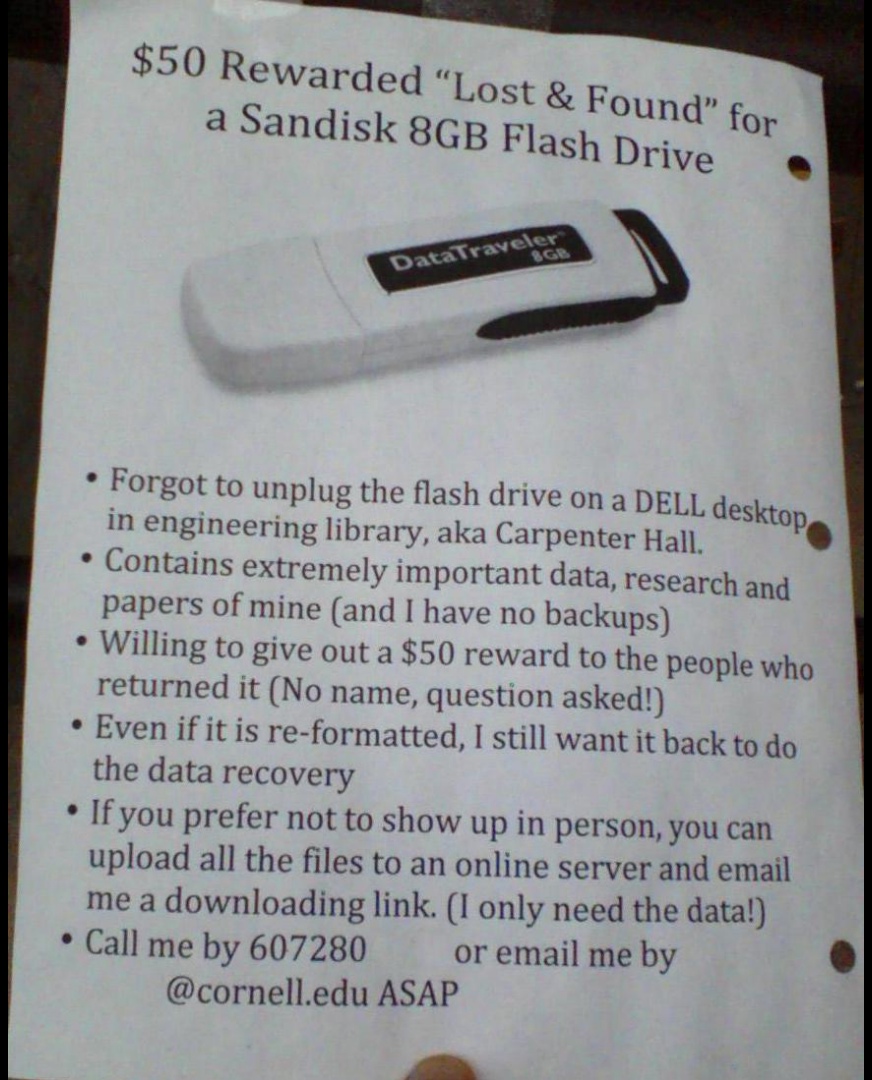
Steve DiDomenico
Head, Enterprise Systems
Northwestern University Library
steve@northwestern.edu

Linda Newman
Head of Digital Collections and Repositories
University of Cincinnati Libraries
newmanld@ucmail.uc.edu

Today we'll be talking about:

- What is Digital Preservation
- Why is it important
- Libraries, museums, and other institutions' recent work in this area
- What you (Mathematicians) can do and how you can help
- Time for questions

An Informal Survey

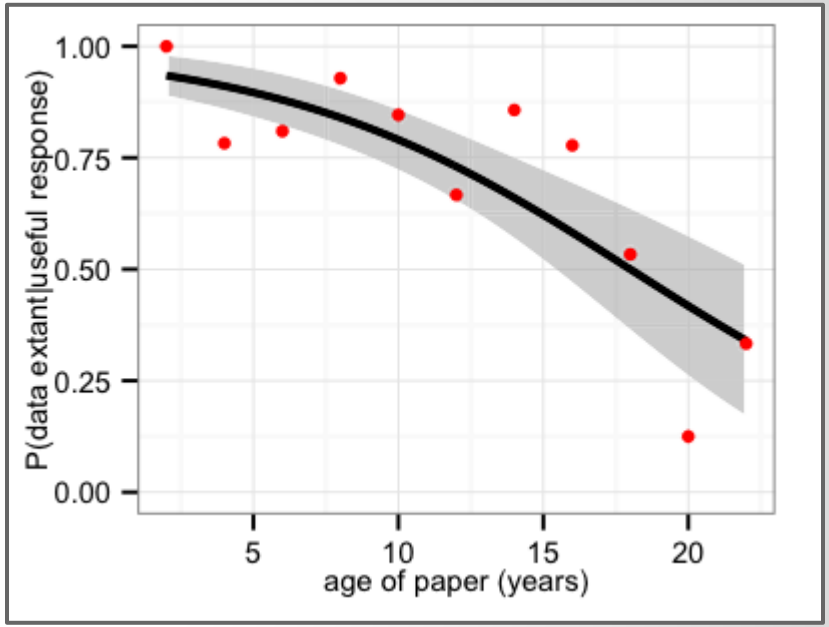


Tweeted in 2012 by Gail Steinhart, Head of Research Services, Mann Library, Cornell University

https://twitter.com/gailst/status/237525591547580416

The Availability of Research Data Declines Rapidly with Article Age

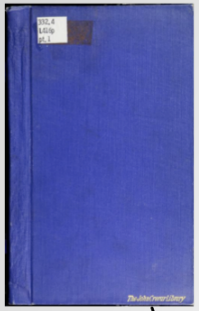
"The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either **lost or on inaccessible storage media**."

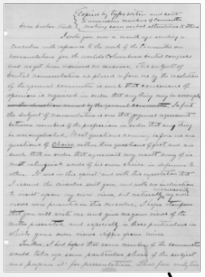


Predicted probability that the research data were extant given a useful response was received

1. Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2013). *The Availability of Research Data Declines Rapidly with Article Age*. Current Biology, 24(1), 94–97. doi:10.1016/j.cub.2013.11.014

Content





Future researchers may not find digital historical data

PDF

Time

1800 1900 2000 2100

Able to find physical/analog content

Missing digital content

Able to find digital content

A Library’s mission is to:

- Collect and preserve the scholarly record
- Provide access to this content
- Support the creation of new knowledge

Applies whether materials are in print or digital

Digital Preservation

Some issues with storing digital information include:

- Backups — Keeping separate geographic locations for disaster situations
- Maintenance (performing fixity checks, test restores)
- Versioning
- Costs of high-quality storage
- Curation: What can we not afford to lose?

Digital Preservation

Digitized physical content: Data where the content originated from a physical object

Examples: scanned books, digitally captured photo of a painting

Issues include:

What quality do we capture at in order to make sure the item is preserved?

What equipment do we use to capture?

How do we note our digitization processes for the future (particularly if there is a problem)?

Digital Preservation

Born-digital content: Data that doesn't have a physical equivalent

Examples: photos from a digital camera, email, web pages

Issues include:

- Storage medium (floppy disks, burned data discs that only last a few years, etc.)
- Format — Obsolete formats difficult or impossible to open, may have missing data
- Difficult to preserve content on websites, private servers, social media.
- There can be a lot of content, hard to determine what to keep

Digital Preservation

For both born-digital and digitized materials:

Can we preserve what we need before it’s too late?

Does metadata exist (or can it be created) so we know the content that we have?

Digital Repository: A preservation system designed to help librarians, curators, and other specialists keep track of and maintain digital information over time. It can also provide users and patrons access to the content.

(In addition to some other other things, such as access controls, APIs for content ingestion, copyright support, and more)

Libraries, museums, and other institutions have created community-developed, open source software such as:

avalon
MEDIA SYSTEM

Avalon Media System

media.northwestern.edu

Search

Sign in | Help

Repository | audio + video

Browse

Search

Browse by

Format

Moving Image (82)

Sound Recording (70)

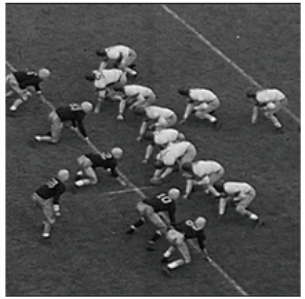
Date

Genres

Collection

Unit

Robert Marcellus Master Class Audio Archives



Northwestern University Football Films

Using the System

The Audio+Video Repository houses audio and video collections.

- Use the search box, or browse using the terms on the left side to discover content.

Consortial/Cooperative Preservation Services

- LOCKSS
- Digital Preservation Network (DPN)
- Portico
- DuraCloud
- APTrust
- HathiTrust

What can you do

- **Use Library Services!**
  - Faculty members: contact your Digital Librarians/Preservation specialists to discuss how to preserve your content
  - Publish to Open Access journals
  - Researchers: Discover digital resources that have become available
- For personal items, make sure you have good backups and metadata (descriptions) of your data

Puzzle #1 - Checksums

Challenge/puzzle #1:

- Is there a better hash function – a better checksum algorithm for the digital preservation use case?

- Checksums are hashes used to detect errors in data transfer and for authenticity validation.
- ‘Collisions’ occur when the same checksum is generated for two files that should not match.
- Low likelihood of a collision caused by accidental degradation (bit rot)?
- High likelihood of a collision caused by malicious alteration? (justifies a cryptographic hash)

Puzzle #1 - Checksums

Checksums & Cryptographic Hashes – an oversimplified history:

- 1961: CRC (Cyclic Redundancy Check) – not cryptographic

The following checksums are created with cryptographic hash functions (here cryptographic means that it is practically impossible to invert and recreate input data from the hash value). These checksums are less vulnerable to collisions:

- 1991: MD5 (Message Digest algorithm) – cryptographic but found to be highly vulnerable
- 1995: SHA-1 – (Secure Hash Algorithm)
- SHA-1 still in wide use but NIST directive to US. Agencies to stop using SHA-1 by 2010. Mozilla plans to stop accepting SHA-1 SSL certificates by 2017.
- 2002: SHA-2 – Less vulnerable than SHA-1.
- 2012: SHA-3 – also known as ‘Keccak’ developed after 5 year NIST-sponsored contest – not yet in widespread use.

Puzzle #1 - Checksums

Checksums - performance:

- CRCs are less complex than the SHA family of cryptographic hash functions.
- SHA-2 and SHA-1 are commonly assumed to be slower than MD5 but empirical data is unpublished and inconsistent. A 2014 study tried to address this evidence gap, and found SHA-256 30% slower per gigabyte than MD5.[1]
- To be secure, hash functions may need to be slow – the faster the hash the more vulnerable it may be to brute force attacks.
- For preservation at mass scale, has CPU power kept up with the amount of constant file validation we need, allowing re-calculation, not just comparison of stored values?

2. Duryee, Alex. “What Is the Real Impact of SHA-256? A Comparison of Checksum Algorithms.” AVPreserve.com, October 2014. http://www.avpreserve.com/wp-content/uploads/2014/10/ChecksumComparisons_102014.pdf.

Puzzle #1 - Checksums

- With digital preservation at mass scale, are cryptographic checksums, in addition to robust system security, necessary? (The answer has been considered to be yes, to prevent malicious alteration by bad actors.)

Challenge/puzzle #1:

❖ Is there a better hash function – a better checksum algorithm for the digital preservation use case?

Puzzle #2 - How many copies do we need?

Challenge/puzzle #2:

- How many copies do we need to preserve a given amount of data over a specified period of time?

- LOCKSS stores 7 copies across 7 geographically dispersed servers.
- Academic Preservation Trust uses Amazon S3 (3 copies) and Amazon Glacier (3 copies), with geographic distance between S3 (Virginia) and Glacier (Oregon). Copies are in different 'availability zones' with separate power and internet. Amazon claims 99.999999999% (11 nines) reliability for S3 and Glacier in this configuration.[3]
- But are these estimates based on projections and models or empirical studies?
- Hardware manufacturers' estimates of mean time to data loss (MTTDL) may not tell us much about the extent of data loss over a given period of time.[4]

3. http://media.amazonwebservices.com/AWS_Storage_Options.pdf

4. Rosenthal, David S. H. "DSHR's Blog: 'Petabyte for a Century' Goes Main-Stream." *DSHR's Blog*, October 6, 2010. http://blog.dshr.org/2010/09/petabyte-for-century-goes-main-stream.html and Greenan, Kevin M., James S. Plank, and Jay J. Wylie. "Mean Time to Meaningless: MTTDL, Markov Models, and Storage System Reliability." In Proceedings of the 2nd USENIX Conference on Hot Topics in Storage and File Systems. Boston, Mass.: USENIX Association, 2010. https://www.usenix.org/legacy/events/hotstorage10/tech/full_papers/Greenan.pdf.

Puzzle #2 - How many copies do we need?

- Greenan et al propose a new measurement – Normalized Magnitude of Data Loss (NOMDL) – and demonstrate that it is possible to compute this using Monte Carlo simulation based assigning failure and repair characteristics to hardware devices drawn from real-world data.[3]
- David Rosenthal has written extensively about the problems with proving that we can keep a petabyte for a century. He has concluded that "the requirements being placed on bit preservation systems are so onerous that the experiments required to prove that a solution exists are not feasible."[4] He proposed a bit half-life measurement (the time after which there is a 50% probability that a bit will have flipped), but argues that it is not possible to construct an experiment that proves that a given number of copies of files stored in a particular configuration will keep a petabyte safe for a century.[5]

5. Greenan, Kevin M., James S. Plank, and Jay J. Wylie. "Mean Time to Meaningless: MTTDL, Markov Models, and Storage System Reliability." In Proceedings of the 2nd USENIX Conference on Hot Topics in Storage and File Systems. Boston, Mass.: USENIX Association, 2010. https://www.usenix.org/legacy/events/hotstorage10/tech/full_papers/Greenan.pdf.

6. Rosenthal, David S. H. "Bit Preservation: A Solved Problem?" *International Journal of Digital Curation* 5, no. 1 (June 22, 2010): 134–48. doi:10.2218/ijdc.v5i1.148. http://www.ijdc.net/index.php/ijdc/article/view/151.

7. Rosenthal, David S. H. "Keeping Bits Safe - ACM Queue." *ACM Queue* 8, no. 10 (October 2010) (October 1, 2010). http://queue.acm.org/detail.cfm?id=1866298

Puzzle #2 - How many copies do we need?

Where does this us when designing digital preservation solutions?

We have general comfort with assumptions that

- The more copies of our files the better.
- The more independent the copies are from each other the better. (different storage technology, network independence, as well as organizational and geographic independence)
- The more frequently the copies are audited the better.

But as the size of the data increases, the time and cost for audits increases, the per-copy cost increases, and the number of storage options that are feasible decrease. If we maximize the number of copies, the frequency of audits and network independence, high costs will force us to preserve much less.

Puzzle #2 - How many copies do we need?

- Should librarians and archivists accept a higher level of risk in order to preserve more?
- Is it better to preserve more quantity with some data loss than little quantity with unproven projections of no data loss?

Challenge/puzzle #2:

❖ How many copies do we need to preserve a given amount of data over a specified period of time?

A more definitive answer to this question would help us make better informed choices (and get funding).

Puzzle #3 - Format Obsolescence

Challenge/puzzle #3:

- Is there a better algorithm that could be used to initiate format preservation actions?

- Format obsolescence predictions have not proven to be as dire as previously thought – the web continues to be able to render much earlier content, even for formats no longer in active use.
- The importance of open source documentation may greatly increase a format's chances of being rendered in the future.
- Popularity of a format may be a simple test, but popularity can wax and wane quickly and unexpectedly as happened with the Flash video format.

Puzzle #3 - Format Obsolescence

- We've used tools such as Droid, JHOVE and Apache Tika to identify formats and create metadata about them – leading some practitioners to emphasize preserving those formats we can identify. But one study suggests that browsers still render much of the content these tools fail to identify, so perhaps our format criteria are inadequate.[6]
- In some cases, emulation of a software environment could prove to be equally if not more feasible than format migration.

8. Jackson, Andrew N. "Formats over Time: Exploring UK Web History." In *arXiv:1210.1714 [cs]*. Toronto, Canada, 2012. http://arxiv.org/abs/1210.1714.

Puzzle #3

Librarians and archivists would like to know when to take preservation ‘actions’ – to intervene and establish a method of format migration or emulation.

At present we are balancing opinions about format renderability and brittleness without knowing whether factors such as wide-spread use (popularity), self-documentation (open source) and rendering complexity (layers) might be better and possibly more quantifiable predictors.

Challenge/Puzzle #3:

- Is there a better algorithm that could be used to initiate format preservation actions?

Conclusion

- Data is at risk if it isn't preserved properly
- Talk with libraries at your institution and make sure your research will be found for later generations

Questions?

Steve DiDomenico
Head, Enterprise Systems
Northwestern University Library
steve@northwestern.edu

Linda Newman
Head of Digital Collections and Repositories
University of Cincinnati Libraries
newmanld@ucmail.uc.edu

These slides are available on Slideshare at: http://www.slideshare.net/newmanld/the-quest-for-digital-preservation-will-part-of-math-history-be-gone-forever

A bibliography is available at: http://www.slideshare.net/newmanld/the-quest-for-digital-preservation-mathfest-2015-bibliography