The Quest for Digital Preservation: Will a portion of Math History be lost forever?

Steve DiDomenico and Linda Newman

Abstract Libraries, archives, and museums have traditionally preserved and provided access to many different kinds of physical materials, including books, papers, theses, faculty research notes, correspondence, and more. These items have been critical for researchers to have a full understanding of their fields of study as well as the history and context that surround the work.

However, in recent years many of these equivalent materials only exist electronically on websites, laptops, private servers, and social media. These digital materials are currently very difficult to track, preserve, and make accessible. Future researchers may very well find a black hole of content: discovering early physical materials and late electronic records, but little information for the late 20th though early 21st Centuries. In other words, a portion of history–including the field of Mathematics—may be lost unless this electronic content is cared for properly.

This article will cover the issues surrounding Digital Preservation, including recommendations to make sure data is reasonably safe. Additionally a small number of discrete challenges and unsolved problems in the field of Digital Preservation with be posed, where Mathematicians may be able to help with analysis and new algorithms.

Steve DiDomenico
Northwestern University Library, 1970 Campus Drive Evanston, IL 60208, e-mail: steve@northwestern.edu

Linda Newman
University of Cincinnati, 2600 Clifton Ave., Cincinnati OH 45221 e-mail: newmanld@ucmail.uc.edu

1 Digital Preservation

As faculty, students, researchers, and professionals store more of their data in electronic rather than paper form, new knowledge and steps for proper care must be taken in order to sure this information is preserved. Without preservation, digital information will be lost to future researchers, creating a digital black hole in time where they are unable to find the information they need. Fortunately there is still time left to ensure existing digital content—including portions of Math research—will continue to be available for generations to come.

1.1 The Availability of Research Data

As media are able to store more and more digital data, it becomes very easy for even a single storage mishap to wipe out large amounts of information. For instance, a Cornell University engineering student accidentally left their research data on a USB flash drive plugged into a lab computer, which was subsequently lost or stolen and a backup copy didn't exist. [Steinhart 2012] While this example is only anecdotal evidence, there is an element of truth to the fragility of large amounts of digital information.

In the article *The Availability of Research Data Declines Rapidly with Article Age*, the authors dive deeply into the issue of whether the detailed data behind published research can be found. [Vines et al 2014] They asked a number of authors for the supporting data from their published papers. For authors that responded, they found that the data for older papers is increasingly hard to find, see Fig. 1.

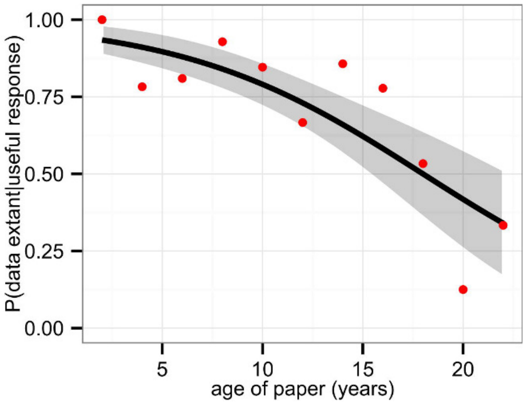

Fig. 1 Predicted probability that the research data were extant given a useful response was received [Vines et al 2014]

This graphically shows the probability that the data were extant over time, with a marked decrease over time. While the study only covered 20 years of research articles, it seems highly unlikely that the probability would increase, and more likely that the data's availability would be extremely low:

> 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. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year...Unfortunately, many of these missing data sets could be retrieved only with considerable effort by the authors, and others are completely lost to science. [Vines et al 2014]

The unavailability due to lost or inaccessible storage media is worth noting because records of scholarship are increasingly only digital. The question now is at what point will digital preservation and storage practices become so commonplace that these datasets will have a high likelihood of survival over time.

1.2 A Data Black Hole

Researchers in the future may find a data black hole as they look back at history, see Fig. 2. With a view far back enough in time—particularly before personal computers and mass digital storage media were invented—researchers will be able to find physical items such as books, research notes, letters, film, analog audio tape, and more. While this type of media include their own set of preservation issues care is usually reasonably straightforward, degradation under proper care happens slowly allowing more time to copy to new media, and experts in long-term longevity are not hard to find.

Likewise, a future researcher looking at content in the late 21st Century will hopefully be able to find digital equivalents likely in the form of ebooks, email, digital video, etc. when the issues of digital preservation are common knowledge and standards for proper care have matured.

However, when future researchers look back at a time in between the physical and primarily digital eras they may be unable to find a certain amount of content. During this black hole time period digital content will be lost and not cared for in a way that supports long-term preservation. Internet pioneer Vint Cerf has expressed concern about this as well calling the time period a "digital dark age" [Maffeo 2015] where technology has moved too quickly to be able to read or open old formats.

Data can become inaccessible for a variety of different reasons. A few examples include: missing content can be research data that was on obsolete storage media, the digital media may degrade too quickly before the data can be copied, or (as Vint Cerf suggests) the data reside in proprietary formats where the original application or hardware to read the data is unavailable. The information could have existed in the form of email correspondence where its historical significance seemed less important than handwritten letters, and was thus deleted instead of saved. Additionally, long-term storage and access may not be adequately addressed for content stored in

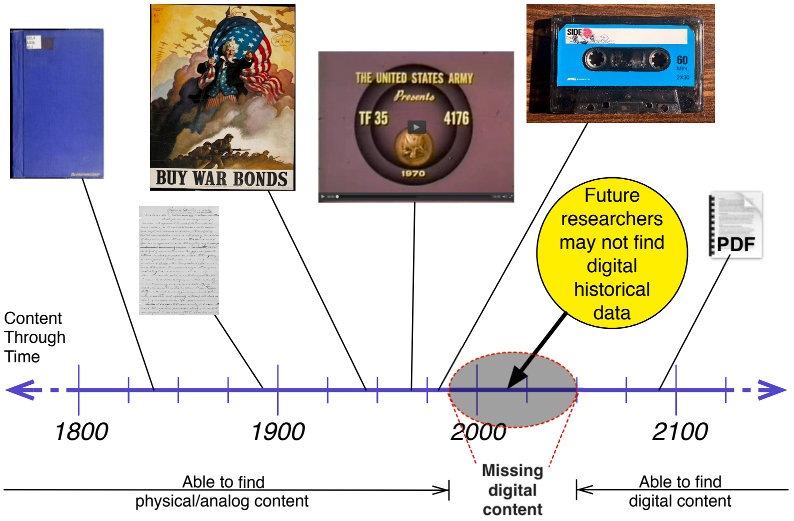
BUY WAR BONDS
THE UNITED STATES ARMY
Presents
TF 35 4176
1970
PDF
Future researchers may not find digital historical data
Content Through Time
1800
1900
2000
2100
Able to find physical/analog content
Missing digital content
Able to find digital content

Fig. 2 Content across time: Future researchers may be unable to find digital content within a certain portion of time.

blogs, wikis, personal websites, social media, cell phones, and other private or inaccessible locations.

1.3 Digital Preservation Standards

Most of these issues with inaccessible data can be solved through digital preservation standards, which include:

- **High-quality storage technologies** — Using high-quality storage that is refreshed regularly before it becomes obsolete works well to ensure data stays online (although it can be expensive).
- **Backups** — Creating regular backups of the data, at least one copy kept in a separate geographic location to cover disaster situations.
- **Metadata** — Creating good descriptions of the content so it can be found and understood later.
- **Maintenance** — Includes performing fixity checks to ensure the data hasn't unexpectedly changed, and performing test restores to ensure backups are working.
- **Provenance and Versioning** — Saving older copies rather than overwriting in case the older copy ends up being desirable, and tracking the files' histories.
- **Curation** — The difficult decision about what should be kept or deleted.
- **Normalization** — Converting data to more sustainable long-term formats.
- **Software** — Tools that help enable preservation, providing ingestion interfaces, automatic metadata extraction, fixity checking, and more.

Digital content can be broken down into two different categories. The first is digitized physical content: data where the content first originated from a physical object and then was digitally captured. Some examples include scanned books or a digital photo of a physical painting. Several issues arise when working with this content, including tradeoffs about what quality the item needs to be digitally captured at in order to meet preservation needs. This can include tough choices about the type and expense for equipment as well as staff time and expertise for working with the material. Additionally, digitization workflows must be carefully planned and documented—particularly if problems are discovered in the process.

Born-digital content is the second category: data that do not have physical equivalents and were created in the digital realm. Examples include photos from a digital camera, email, and web pages.

There are several issues for born-digital content. The first is the storage medium — many media degrade quickly (digital audio tape and burned CDs are notoriously fickle); though regardless of the speed of degradation, the data needs to be regularly refreshed to ensure stability and prevent digital obsolescence. The second issue are outdated formats that can be difficult or impossible to open, or may open with missing information. It can also be difficult to preserve content on websites, private servers, and social media. And lastly it may be hard to determine what to keep if data is in abundance.

1.4 Digital Repositories

These issues of digital preservation fall at the heart of the Library profession. A few of the American Library Association's Core Values of Librarianship [ALA 2004] include:

- **Collect and preserve the scholarly record** — "The Association supports the preservation of information published in all media and formats. The association affirms that the preservation of information resources is central to libraries and librarianship."
- **Provide access to content** — "All information resources that are provided directly or indirectly by the library, regardless of technology, format, or methods of delivery, should be readily, equally, and equitably accessible to all library users."
- **Support the creation of new knowledge, education, and lifelong learning** — "ALA promotes the creation, maintenance, and enhancement of a learning society..."

Note that these practices apply whether materials are in print or digital format.

In order to address the need for better digital preservation tools, libraries, museums, and other institutions have created open source, community-developed software solutions called Digital Repositories: preservation systems designed to help librarians, curators, and other specialists keep track of and maintain digital information over time. These systems can also provide users and patrons the ability to search, browse, and view content (in addition to other features such as restricted access controls, APIs for content ingestion, copyright support, and more).

Several open source applications have been developed out of these efforts, including:

- **DSpace** — A Java-based turnkey Digital Repository application allowing data capture, searchable indexes, content distribution, interface customization, and more.
- **Fedora** — A Java-based backend Digital Repository application supporting flexible metadata, high performance, fixity checking, and more — upon which other components can be added, such as the front-end access and ingestion interfaces:
  - **Hydra** — A Ruby-on-Rails-based framework that applies Rails concepts and architectures to allow rapid repository interface development for Fedora.
  - **Islandora** — A Drupal-based framework utilizing Drupal concepts and modules for repository interfaces to be easily developed for Fedora.

Maintaining these applications and communities is nearly as important as taking care of the content. DSpace, Fedora, and Hydra are stewarded by the not-for-profit, multi-institutionally supported company DuraSpace which helps with marketing and communications, sustainability planning, community development, fundraising, training, technical leadership, and more. The non-profit Islandora Foundation which plays a similar role in supporting Islandora. But it is important to note that much of the open source software development comes from a variety of different worldwide institutions, with technical and administrative decision-making performed by contributing members of these communities.

1.5 Consortial/Cooperative Preservation Services

In addition to repository software, a number of different services have been cooperatively developed, primarily driven by academic libraries, to achieve digital preservation requirements that would be very expensive to set up alone. A few of these services include:

- **Lots of Copies Keep Stuff Safe (LOCKSS)** — A peer-to-peer network system allowing institutions to keep multiple copies of data through use of specialized peer-to-peer software. The member-driven cooperative The MetaArchive is a notable example of a LOCKSS network.
- **Digital Preservation Network (DPN)** — A collaborative service where five ‘nodes’ —which may themselves be digital preservation systems with multiple institutional members — securely store content on datacenter-grade hardware and save copies with each other. Members of a node can upload their data where it is then copied to the data stores and remains in a dark unchanging archive for at least 20 years or more.

- **DuraCloud** — A not-for-profit service that provides an interface to upload and replicate content into cloud storage systems, such as Amazon S3, Amazon Glacier, Rackspace, San Diego Supercomputer Center, and the Digital Preservation Network. It also includes some other preservation and access features. Administered by DuraSpace.
- **Portico** — A not-for-profit company that provides data repository services, with a focus on online journals. Publishers enter into agreements with Portico that allow their content in Portico to move from 'dark' status to 'light' (for Portico members) in certain circumstances - such as a publisher bankruptcy or decision to stop journal publication.
- **HathiTrust** — Collaborative repositories where their members can place select digital data into a single digital preservation system, access the content, and share the data or leave it dark. HathiTrust has had an emphasis on monographs.
- **APTrust (Academic Preservation Trust)** — Similar to DuraCloud, but organized as a member cooperative, members place their digital assets into a system that replicates their content in a dark archive distributed in multiple places on the cloud, such as Amazon S3 and Amazon Glacier. The University of Virginia is the founding institution and provides the development and administrative staff. APTrust is one of the DPN nodes.

Note that many of these services include digital preservation features such as geographic dispersion, fixity checking, collaborative support, cost-sharing, and data security. These collaborative initiatives lower the barrier to entry and make it much easier for institutions to enable robust preservation services for their data.

These services have achieved different stages of maturity. No one service claims to solve all digital preservation problems. Many academic libraries participate in more than one endeavor. At the end of this article we present some of the technical challenges that remain in this domain, and which may benefit from a mathematical and engineering analysis.

1.6 What Faculty and Researchers Can Do Today

Faculty can be involved with these initiatives as a way to help ensure their content is available in the long-term. It's important to use library services and take advantage of professional resources. Faculty can talk to their Digital Librarian or Preservation specialists and discuss how best to preserve digital content, including placing it into a formal Digital Repository and/or for help with management of the personal collection. The book *I, Digital: Personal Collections in the Digital Era* describes this new area for professionals, "There is a growing community of practice related to the acquisition and management of personal digital collections, with many of the participants being archivists, special collections librarians, and manuscript creators." [Lee 2011] Additionally, publishing to Open Access journals helps ensure that faculty work will be stored in a trusted digital repository and that the content will be found later.

Depending on the project and grant agreements, it may even be necessary for faculty to store their grant-funded research data into a long-term repository. National Science Foundation grant proposals require a data management plan with "...plans for archiving data, samples, and other research products, and for preservation of access to them." [NSF 2015] The Institute for Museum and Library Services "..expects you to deposit data resulting from IMLS-funded research in a broadly accessible repository that allows the public to use the data without charge.." [IMLS 2015] And the Engineering and Physical Science Research Council says "Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party." [EPSRC 2011] Other organizations may have similar stipulations. Many of these requirements have been added in recent years so it is important to check to see if new grant proposals are affected.

For researchers, discovering and using digital resources (particularly in digital repositories) helps make the case that these relatively new tools are useful and the preservation and access services are worth the effort. Librarians, Archivists, Curators, and others can help researchers use these applications.

And for everyone (including faculty, researchers, and even others outside of academia), one of the most important first steps to ensure content remains accessible is to have one or more good backups and descriptions (metadata) of the content. Be sure to keep good notes for the people who may curate the data in the future. Also, cloud backup solutions are inexpensive for small-to-medium sized hard drives and an easy way to have off-site insurance against natural disasters or theft.

Performing these tasks is a start to ensure more content is available to future researchers—shrinking the time span of lost digital data—and hopefully at least a portion of Math History won't be lost for future generations.

2 Challenges and unsolved problems

These are unsolved problems in the field of Digital Preservation where Mathematicians may be able to help with analysis and new algorithms.

2.1 Checksums - Is there a better hash function for the digital preservation use case?

Checksums are hashes used to detect errors in data transfer and for authenticity validation. Checksums are routinely used by digital preservation applications to check that multiple, live copies of files are in fact copies of the same digital object. 'Collisions' can occur when the same checksum is generated for two files that should not in fact match. We think that there is a low likelihood of a collision caused by accidental degradation of bits, but a high likelihood of a collision caused by malicious alteration. The fear of malicious alteration has justified a cryptographic hash.

A brief and oversimplified history of checksums and cryptographic hashes could start with the Cyclic Redundancy Check (CRC) from 1961, which is not cryptographic, followed by the Message Digest Algorithm (commonly known as MD5) from 1991, which is cryptographic but has been found to be highly vulnerable. MD5 remains in widespread use. In 1995 the first 'Secure Hash Algorithm - SHA-1 was developed. This has also been found to be vulnerable. The National Institute of Standards and Technology (NIST) has directed US Agencies to stop using SHA-1. Mozilla plans to stop accepting SHA-1 SSL certificates by 2017. In 2002 SHA-2 (also called SHA-256) was deployed; this is widely assumed to be less vulnerable to malicious alteration than SHA-1. SHA-3 was deployed in 2012; it is also known as 'Keccak' and is the result of a multi-year NIST-sponsored contest.

CRCs are less complex than the SHA family. SHA-2 and SHA-1 are assumed to be slower than MD5 but empirical data may be unpublished and inconsistent. A 2014 study tried to address this evidence gap, and found that SHA-256 was 30% slower per gigabyte than MD5. [Duryee 2014] To be secure, a hash function may need to be slow, as the faster the hash can be calculated the more vulnerable it may be to brute force attacks.

This presents a problem for the digital preservation use case of preservation at mass scale, which requires constant file validation that allows for re-calculation, not just comparisons of stored hash values. The importance of confirming that there has been no accidental degradation is the primary objective of such file validation, but malicious alteration of files is not unimaginable in digital preservation networks. Is there perhaps a better algorithm for the digital preservation at mass scale use case, balancing these competing interests? Or if we find that we do not have the CPU power to generate SHA-256 hashes with the required frequency, should we simply return to CRC or MD5 hashes for day to day comparisons?

2.2 How many copies to we really need to keep stuff safe?

LOCKSS (Lots of Copies Keep Stuff Safe) is a digital preservation application in widespread use; most implementations store seven copies across seven geographically dispersed servers. Some digital preservation networks, such as DPN, APTrust and DuraCloud, are attempting to leverage the cloud in a similar fashion. APTrust stores 3 copies on Amazon S3 in Virginia and 3 copies on Amazon Glacier in Oregon, utilizing different 'availability zones' with separate power and internet. Amazon claims 99.999999999% (11 nines) reliability for S3 and Glacier in this configuration. (http://media.amazonwebservices.com/AWS%5FStorage%5FOptions.pdf) But these estimates may be based in part on hardware manufacturers' estimates of mean time to data loss (MTTDL), which may not tell us enough about the extent of data loss over a given period of time. [Rosenthal October 6, 2010] Greenan et al proposed a new measurement — Normalized Magnitude of Data Loss (NOMDL)

— and demonstrated that it is possible to compute this using Monte Carlo simulation, assigning failure and repair characteristics to hardware devices drawn from real-world data. [Greenan et al 2010] Whereas NOMDL is almost certainly a better measurement than MTTDL for digital preservation solutions, it still may not tell us exactly how many copies we need to keep, and to what extent we should achieve data independence among those copies.

David Rosenthal has written extensively about the problems with proving that we can keep a petabyte for a century and 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." [Rosenthal June 22, 2010] 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. [Rosenthal October 1, 2010]

Despite the lack of proof, those of us designing digital preservation solutions often have general assumptions that the likelihood of digital preservation increases with each of the following factors, although it is also not clear with the increase of any one factor, when the increased likelihood becomes a matter of diminishing returns.

- The number of copies
- Data independence of copies (Complete independence requires different storage technology and software, network independence, as well as organizational and geographic independence)
- The frequency of audits using hashes

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 all factors, high costs will force us to preserve much less.

In the operation of a digital preservation solution, if an audit reports the loss of integrity of a file, we should have at least two other copies that agree with each other, so we have a tie-breaker to tell us which file to keep. So keeping three copies is clearly the minimum to support such a tie-breaker, but if those three copies are not stored independently it is not unimaginable to conceive of losing all three copies at close to the same time. Is seven copies (the LOCKSS recommendation) enough or too much? Can we do a better job of modeling and making evidence-based decisions? In the absence of such data, should we accept what may be a higher risk in order to be able to afford to preserve more?

2.3 *Format Obsolescence - is there a better algorithm to trigger 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 existence 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.

Tools such as Droid, JHOVE and Apache Tika have been developed to identify formats and create metadata about them — leading some practitioners to emphasize preserving the formats we can identify and describe with such tools. But one study suggests that browsers still render much of the content these tools fail to identify, so perhaps our criteria are still inadequate. [Jackson 2012]

In some cases, emulation of a software environment could prove to be equally if not more feasible than format migration. 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.

3 Conclusions

Despite areas where better algorithms and more evidence-based decision-making could improve long-term digital preservation practices, digital preservation networks and institutional repositories are proving their worth today and are becoming essential components of disaster recovery plans for libraries and archives. [Mallery 2015] Faculty and researchers can increase the likelihood of preserving their digital output by talking to their Digital Librarian or Preservation specialists about placing the output of their research in their local institution's digital repository, asking for guidance on the data management guidelines of federal and other funding agencies, and asking for help with best practices for the management and backup of personal collections.

Performing these tasks is a start to ensure more content is available to future researchers—shrinking the time span of lost digital data—and hopefully at least a portion of Math History won't be lost for future generations.

4 Further Reading

For further reading, a list of additional resources (other than citations already listed) can be found in the original presentation bibliography stored at the University of Cincinnati's Digital Repository:
http://dx.doi.org/doi:10.7945/C23S3C

References

ALA - American Library Association (2004) Core Values of Librarianship. http://www.ala.org/advocacy/intfreedom/statementspols/corevalues Cited 21 Jan 2016

EPSRC - Engineering and Physical Sciences Research Council (2011) EPSRC policy framework on research data. https://www.epsrc.ac.uk/about/standards/researchdata/ Cited 21 Jan 2016

Duryee, Alex. (October 2014) What Is the Real Impact of SHA-256? A Comparison of Checksum Algorithms. AVPreserve.com. http://avpreserve.com/wp-content/uploads/2014/10/ChecksumComparisons%5F102014.pdf Cited 21 Jan 2016

Greenan, Kevin M., James S. Plank, and Jay J. Wylie. (2010) 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. https://www.usenix.org/legacy/events/hotstorage10/tech/full%5Fpapers/Greenan.pdf Cited 21 Jan 2016

IMLS - Institute of Museum and Library Services (2015) General Terms and Conditions for IMLS Discretionary Grant and Cooperative Agreement Awards For Awards Made After December 26, 2014. https://www.imls.gov/sites/default/files/gtc%5Fafterdec2014%5F11%5F2015.pdf Cited 21 Jan 2016

Jackson, Andrew N. (2012) Formats over Time: Exploring UK Web History. In arXiv:1210.1714 [cs]. Toronto, Canada. http://arxiv.org/abs/1210.1714 Cited 21 Jan 2016

Lee C (2011) Introduction to I, Digital. pp. 20. In: Lee, C (ed) I, Digital: Personal Collections in the Digital Era. Society of American Archivists, Chicago.

Maffeo L (2015) Google's Vint Cerf on how to prevent a digital dark age. Guardian News and Media Limited. http://www.theguardian.com/media-network/2015/may/29/googles-vint-cerf-prevent-digital-dark-age Cited 21 Jan 2016

Mallery, Mary, editor (2015) Technology Disaster Response and Recovery Planning: a LITA guide. American Library Association. https://books.google.com/books?id=1szLCQAAQBAJ&lpg=PT19&ots=0Tjx-LDPEp Cited 21 Jan 2016

NSF - National Science Foundation (2014) Proposal and Award Policies and Procedures Guide. NSF 15-1, OMB 3145-0058. http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg%5F2.jsp#dmp Cited 21 Jan 2016

Rosenthal, David S. H. (June 22, 2010) Bit Preservation: A Solved Problem? International Journal of Digital Curation 5, no. 1: 134?48. doi:10.2218/ijdc.v5i1.148. http://www.ijdc.net/index.php/ijdc/article/view/151 Cited 21 Jan 2016

Rosenthal, David S. H. (October 1, 2010) Keeping Bits Safe - ACM Queue. ACM Queue 8, no. 10. http://queue.acm.org/detail.cfm?id=1866298 Cited 21 Jan 2016

Rosenthal, David S. H. (October 6, 2010) DSHR's Blog: 'Petabyte for a Century' Goes Main-Stream. DSHR's Blog. http://blog.dshr.org/2010/09/petabyte-for-century-goes-main-stream.html Cited 21 Jan 2016

Steinhart G (2012) a picture is worth 1000 words... back it up! Twitter. https://twitter.com/gailst/status/237525591547580416 Cited 21 Jan 2016

Vines, T H et al. (2014) The Availability of Research Data Declines Rapidly with Article Age. Current Biology, Volume 24, Issue 1, 94 - 97. http://dx.doi.org/10.1016/j.cub.2013.11.014 Cited 21 Jan 2016