Bit Rot and Silent Data Corruption in Digital Audiovisual Preservation
Jeffrey Lauber

Introduction

As atoms decay, so too do bits. Though relatively mild compared to other threats to data integrity, the potential for and prevalence of bit rot—or silent data corruption in general—in data storage systems poses a set of risks and consequences that should not be overlooked. Though not decaying in a literal, physical sense, the unreliability of contemporary digital storage, namely hard-disk drives, makes long-term preservation of large, complex sets of digital bits—audiovisual data a case in point—increasingly difficult and susceptible to random and spontaneous corruption. Moreover, these small, silent corruptions run the risk of going unnoticed by keepers of digital information, with potentially catastrophic results. Understanding small data corruptions of the sort, and understanding the most reliable measures for mitigating their risks, has the potential to curb the possibility of irreplaceable loss. Much has been written about data corruption, data storage reliability, and data integrity; this paper attempts to distill a wide array of information from various fields of study to provide keepers of digital heritage with an overview of bit rot and silent corruption, specifically as it pertains to the preservation of digital audiovisual data. The paper will begin with an overview of silent data corruption, followed by a discussion of the role data storage plays in both perpetuating and alleviating its presence (with a specific focus on hard-disk storage). It will then outline those concerns which are specific to digital audiovisual content. Finally, it will overview some of the strategies used by archivists and data technicians in the prevention of data corruption to point towards practical ideals for long-term digital preservation.

Overview of Bit Rot and Other Silent Errors

The term *bit rot* is a rather colloquial and figurative one used to describe a type of physical error or corruption in digital data, figurative because digital bits are incapable of physically rotting. Though used with slight semantic difference depending on the person, the field of study, or the subject matter, aggregating some definitions of bit rot proves fruitful for understanding its basic premise. The *Network Dictionary* defines bit rot as: "a colloquial computing term used either to describe gradual decay of storage media or to facetiously describe the spontaneous degradation of a software program over time."[1] Though useful as a foundation, this definition fails to make clear its ambiguous use of the term *decay*. Supplementing with a definition from the *Hacker Dictionary* alleviates some of this uncertainty: "Hypothetical disease the existence of which has been deduced from the observation that unused programs or features will often stop working after sufficient time has passed, even if 'nothing has changed.' The theory explains that bits decay as if they were radioactive. As time passes, the contents of a file or the code in a program will become increasingly garbled."[2] Considering bit rot as a phenomenon akin to that of radioactive decay is an apt comparison. Still, these definitions fail to indicate the ways this decay manifests itself in the digital realm, and expanding a bit further seems necessary. Working with the radioactive decay analogy, Rosenthal describes bit rot as: "a process that randomly flips the bits the system stores with a constant small probability per unit time. In this model, we can treat bits as radioactive atoms, so that the time after which there is a 50% probability that a bit will have flipped is the *bit half-life*."[3] The concept of bit half-life moves one closer to understanding the ways in which bit rot can be understood and measured,

---

[1] Jielin Dong, ed. *Network Dictionary*. Javvin Technologies, 2007. 68.
[2] *Hacker Dictionary*. http://www.hacker-dictionary.com/terms/bit-rot.
[3] David S.H. Rosenthal. "Keeping Bits Safe: How Hard Can It Be?" *Communications of the ACM* 53, no. 11 (November 2010): 50.

and can be explored further in Rosenthal's article; important in this definition for now is the granularity of the phenomenon (i.e. that it occurs at the bit-level) and its temporality (i.e. that its probability of occurrence increases with time). A final consideration in defining bit rot is that it is a form of silent or latent data corruption, i.e. bit-level errors that occur with "no indication from the drive that an error has occurred"[4] and that will often go undetected unless explicitly looked for.[5] The primary risk associated with corruptions of the sort is irrevocable data loss.[6]

Combination and distillation of these various definitions helps generate a clearer and more comprehensive foundation from which to discuss its various causes, risks, and remedies in the remainder of this paper. In that respect, bit rot can be defined as a category of data corruption by which a single bit (or small subset of bits) is flipped without indication as a result of aging and decaying physical data storage media, rendering the data altered or entirely unusable.

An understanding of bit rot would be incomplete without a discussion of its common causes and its prevalence in data repositories. The DRAMBORA project cites bit rot as a manifestation of hardware failure, i.e. the storage system is responsible for corrupting the data (this should not suggest that human error is irrelevant; human interactions with storage systems have the undoubted potential to cause the system to corrupt data).[7] Baker et al cite numerous causes by which latent failures may occur in digital storage systems, including accidental overwrite by humans, a failed component within the system, hardware and software

---

[4] Lakshmi N. Bairavasundaram et al. "An Analysis of Data Corruption in the Storage Stack." *ACM Transactions on Storage* 4, no. 3 (2008): 8:1.
[5] Matthew Addis. "Long Term Data Integrity for Large Audiovisual Archives." Proceedings of 2010 Joint Technical Symposium, 66th Congress of the International Federation of Film Archives (FIAF), Norway, Oslo. PrestoPRIME, 2010: 19.
[6] Matthew Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. Publication no. 3.2.1. PrestoPRIME. 2010: 5.
[7] Richard Wright et al. "The Significance of Storage in the 'Cost of Risk' of Digital Preservation." *The International Journal of Digital Curation* 4, no. 3 (2009): 110.

obsolescence, and attack.[8] But perhaps most pertinent to bit rot is age and wear of use of physical storage media over time: Elerath, writing specifically about hard-disk drives, notes the consequence of this decay as: "degradation of the magnetic properties of the media […] a process in which the magnetic media is not capable of holding the proper magnetic field to be correctly interpreted as a 0 or a 1."[9] Especially important to consider in terms of archives is the fact that the likelihood of this phenomenon occurring increases with age of the storage medium, posing a challenge to long-term digital preservation.[10] Added to this, a consideration of the storage medium's age must also include consideration of its use: every instance of use of the storage medium increases the probability of flipped bits, such that the mere processes of detecting latent faults (i.e. reading and analyzing stored data) are potentially taxing to the storage medium and can increase the likelihood of corruption.[11]

It is worth noting that bit rot and similar latent data errors are not entirely common occurrences in data repositories. Yet the vast increase in data on both the file level (e.g. high fidelity audiovisual files, which require large file sizes) and the repository level (i.e. the proliferation of digital data that exists in the world and must be preserved in archives) will continue to make errors of the sort essential to understand in coming years. Measuring the prevalence of bit rot in digital data storage systems is no easy feat: an accurate, quantized bit error rate for contemporary storage media would potentially require exabytes of data observed and analyzed over many years. Still, notable studies have managed to generate average bit error

---

[8] Mary Baker et al. "A Fresh Look at the Reliability of Long-Term Digital Storage." Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems, Belgium, Leuven. 2006: 3-4.
[9] Jon Elerath. "Hard-Disk Drives: The Good, the Bad, and the Ugly." *Communications of the ACM* 52, no. 6 (June 2009): 43.
[10] Lawrence L. You et al. "PRESIDIO: A Framework for Efficient Archival Data Storage." *ACM Transactions on Storage* 7, no. 2 (July 2011): 6:5.
[11] Baker et al. "A Fresh Look at the Reliability of Long-Term Digital Storage." 8.

rates for digital storage media, and these rates prove useful in pointing to the prevalence of bit rot and other latent errors in data repositories.

Toigo, for instance, notes that silent corruptions account for one non-recoverable error in every 67TB of hard-disk storage. Though this ratio appears rather small, Toigo importantly notes that with the amount of high-capacity disk drives employed in storage arrays, this error rate is significant.[12] Rosenthal, citing a study on silent corruption conducted by CERN, notes that of $9.7x10^{16}$ bytes written to RAID arrays, $1.92x10^{8}$ were found to be subject to silent corruption within the first six months of observation (meaning $1.2x10^{-9}$ of the data was permanently corrupted in this timeframe).[13] Addis notes that a bit error rate of one bit in 10^{14} is typical for hard-disk drives; considering that a one-terabyte HDD contains approximately 10^{13} bits of data, failure to improve reliability of hard-disk drives will continue to increase the likelihood of encountering silent corruption in data repositories.[14]

The errors resulting from bit rot and other latent errors manifest themselves in digital files in various ways depending on a number of factors, including file wrapper, encoding, and compression. At the most basic level, latent errors result in data that is either inaccessible or corrupted.[15] In the case of inaccessible data, bit errors result in files which cannot be executed and whose contents cannot be revealed; for corrupted data, files may be executed but their contents appear altered. In image and audiovisual files, the visual characteristics of data corruption are seemingly infinite and will depend on the type of file, the specific bit(s) affected, and the file's compression and encoding schemes. A more comprehensive breakdown of how certain data errors manifest in image and audio files, including visual examples, can be found in

---

[12] Jon William Toigo. "Bit Rot Eroding RAID." *Storage Magazine*, July 2012.
[13] Rosenthal. "Keeping Bits Safe: How Hard Can It Be?" 51.
[14] Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. 47.
[15] Baker et al. "A Fresh Look at the Reliability of Long-Term Digital Storage." 5.

PrestoPRIME's *Threats to Data Integrity from Use of Large-Scale Data Management Environments* report.[16]

Despite its relatively low rate of occurrence, it is clear that when considering large quantities of digital data—as well as the fact that these forms of corruption can potentially go unnoticed for long periods of time—bit rot and silent corruption pose a definite risk for digital data repositories.

Data Storage and Corruption

Bit rot and other forms of data corruption are tied inextricably to the medium on which the bits are stored. The most susceptible to risks of the sort are hard-disk drives (HDDs), whose life expectancy is rather short and whose mechanics become less reliable over time. Aside from the dire consequences of this for the data itself, the short lifespan of a HDD has also been the cause of financial and operational concern for digital repositories when storage needs to be replaced every few years. In the succinct words of Jeff Rothenberg: "digital information lasts forever—or five years, whichever comes first."[17]

HDDs are significantly prone to failure and, despite methods of increasing their reliability (e.g. RAID configurations), it is generally agreed that they cannot be sufficiently trusted on their own for digital preservation. Failure rates of HDDs have been quantified by storage manufacturers using a Mean Time Between Failure (MTBF) model, which indicates the average time that will pass between drive failures. It is important to note at the outset that, in most cases, manufacturer claims differ wildly from statistics gathered in external studies on disk

[16] Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. 57-79.
[17] Jeff Rothenberg. "Ensuring the Longevity of Digital Information." *CLIR Resources*, February 22, 1999: 2.

reliability, manufacturer stats representing a rather inflated sense of trustworthiness.[18] Part of the reason for this is simply misleading figures. The PrestoPRIME report *Threats to Data Integrity* helps to demystify MTBF in that respect: it is noted that the MTBF for HDDs is considered to be 1,000,000 hours, but this figure can easily lead one astray if not interpreted properly. For instance, one might calculate that a MTBF of 1,000,000 hours results in a HDD that will last 100 years, an assumption which, despite being mathematically correct, has not proven true in real-world study and practice. What needs to be considered, then, is the fact that MTBF measures an average: "MTBF is a statistic on the *average* time between failures for a population of things, e.g. drives or tapes. So […] if you have 100 hard drives, each with a MTBF of 1 million hours, then on average a drive will fail in this set every 10,000 hrs, i.e. one drive a year."[19] Added to this, the ways in which manufacturers arrive at their statistics is an important consideration: in one of Addis' examples, for instance, a manufacturer might test ten drives—if nine fail after one year and one lasts 100 years, the manufacturer could still claim a MTBF of ten years for that set of drives, even with potentially 90% data loss after the first year.[20] Baker et al. note that despite manufacturer claims of up to 100 years reliability for HDDs, the life expectancy in real-world practice is really only two-to-five years.[21]

Though time to or between failures of storage media is indeed an essential facet of digital preservation, the quantity of data being preserved plays an essential role as well. In general, the capacity of HDDs has doubled every 18 months throughout the history of their existence.[22] This means an exponentially greater quantity of bits is being stored on HDDs now than in previous

---

[18] Elerath. "Hard-Disk Drives: The Good, the Bad, and the Ugly." 41.
[19] Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. 40.
[20] *Ibid.*
[21] Baker et al. "A Fresh Look at the Reliability of Long-Term Digital Storage." 2.
[22] Addis. "Long Term Data Integrity for Large Audiovisual Archives." 2.

years. However, error rates have not improved in tandem, resulting in an ever-increasing probability of bit errors in digital repositories. As Addis notes:

> The read error rates reported by Elerath from NetApp in 2007, e.g. 1 error in $8x10^{14}$ bytes read being considered 'medium,' are little improved over error rates reported by Chen et al in 1994 of 1 error in 10^{14} bits read, which was over a decade earlier. This lack of improvement in error rates has serious implications on the design and implementation of systems that ensure data safety.[23]

Yet in many digital repositories HDDs are not employed in isolation. One measure that has been widely taken is to configure HDDs into RAID arrays for long-term digital preservation. The redundancy and parity of these arrays offer a much greater level of data protection than individual HDDs. RAID arrays that dedicate disk space to parity have proven to be an adequate method of alleviating the risks of bit errors. If, in a RAID with parity, a bit error—or even an entire disk failure—occurs, parity information is capable of reconstructing the corrupted data.[24] Still, RAID arrays are not without fault and error. Toigo notes that the rates of multiple disk failures are about 1,500 times higher than manufacturer claims, in part due to bit rot and other silent corruptions which cause approximately one error for every 67TB of disk, a significant ratio for large-scale RAID arrays.[25] Aside from its mere susceptibility to silent corruptions, a significant problem with RAID lies in the difficulty of detecting the corruptions in the first place. An IBM report shows that parity scrubs—a typical method of RAID error detection by which stored data is checked against its parity data—are in fact somewhat problematic and cannot be relied on as a primary means of silent corruption detection.[26] Wright et al. sum it up nicely:

---

[23] Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. 47.
[24] Greg Schulz. "Protecting Data." In *Resilient Storage Networks: Designing Flexible Scalable Data Infrastructures*, 320. 1st ed. Digital Press Storage Technology. Burlington, MA: Elsevier Science & Technology, 2004.
[25] Toigo. "Bit Rot Eroding RAID."
[26] J.L. Hafner, V. Deenadhayalan, W. Belluomini, and K. Rao. "Undetected Disk Errors in RAID Arrays." *IBM Journal of Research and Development* 52, no. 4-5 (2008): 417.

"Whilst RAID can mitigate against detected failures, it does not solve the problem of 'silent' errors. Indeed, if there are additional errors in the software or firmware used to implement RAID, then this can actually make the problem worse."[27]

In these ways and more, bit rot and other silent corruptions caused by physical storage media not only pose serious risks to the preservation of digital data, but also have the potential to elude both human and machine attempts at detection. Consequentially, the longer the error remains latent and undetected, the greater the likelihood of total loss of content. Whether or not silent corruptions occur far less frequently than other causes of data loss in digital archives, the dangers of their presence—especially for large-volume repositories and for large file-size content, namely audiovisual—should not be overlooked.

Corruption in Audiovisual Data

Digital audiovisual data is at significantly greater risk than text and still image data when considering silent corruption. As noted above, the relatively large file sizes and quantities of data inherent to audiovisual content make it more probable that a bit will be corrupted and more catastrophic when it does. As Wright et al. aptly note: "The frustration for audiovisual archivists is that digital technology has taken us one step forward, and is now taking us two steps back. […] file based digital technology has *no* ability to cope with loss (corruption; uncorrectable errors), beyond the 'external redundancy' option of multiple copies," especially troubling given the error compensation and concealment options present on both analog and digital tape recorders past.[28]

[27] Wright et al. "The Significance of Storage in the 'Cost of Risk' of Digital Preservation." 113.
[28] *Ibid.,* 118.

One of the great risks of silent corruption with regards to audiovisual data results from the aforementioned increase in storage media capacity over time. The average 2TB HDD, for instance, can record approximately 4,680,000 minutes of audio per square meter of physical storage space, a significantly higher number than previous audio storage media such as compact disks (8,060 minutes per square meter) or reel-to-reel magnetic tape (13.8 minutes per square meter).[29] The much greater capacity of these disks makes it far more likely that silent corruption will affect bits stored on them. But image data—which accounts for a large percentage of digital data when present in storage repositories—has the highest probability of being affected by silent corruptions. A 2013 SMPTE presentation notes that while one second of high quality uncompressed audio (with six 24 bit channels at 96KHz) requires 1728 Kilobytes, one second of 2K cinemascope video at 24fps requires 123,264 Kilobytes, even with a lossless compression factor. This significantly greater data stream makes it far more likely that video data will be subject to bit rot.[30]

Bit rot, and data corruption in general, manifest in varying ways depending on a number of factors, among them file wrapper type and encoding scheme. For instance, when considering uncompressed audio, loss of one byte of data in a WAV file with a 44.1kHz sample rate results in one corrupted sample but does not spread to the remaining samples; a WAV file with one bad sample would be entirely useable and nearly indistinguishable from the original. In contrast, an MP3 file with similar errors would result in a file that either cannot be opened or whose aural qualities are severely altered.[31] Compression in the case of MP3 makes far more catastrophic the effects of data corruption in relation to its uncompressed counterpart.

---

[29] Addis. "Long Term Data Integrity for Large Audiovisual Archives." 3.
[30] François Helt, Benoît Février, Frantz Delbecque, Xavier Brachet, Hans-Nikolas Locher, and Marc Bourhis. "French Cinema Goes IMF." SMPTE 2013 Annual Technical Conference & Exhibition, White Plains, NY. 2013: 6.
[31] Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. 51-52.

A similar risk from compression results from files with JPEG 2000 encoding, especially pertinent given its wide use in the audiovisual archiving community. It has been found that corruption of a mere 0.01% of bytes in a compressed JPEG 2000 file—even with lossless encoding—can affect 50% of the file data, and the cost-to-risk ratio between uncompressed, lossless, and lossy compression schemes are not often balanced.[32] The PrestoPRIME report illustrates the severe visual defects resulting from corruption of just a single byte in TIFF, BMP, PNG, and JPEG files, among others, with various encoding schemes—as video files are essentially a series of still images, studies like these offer a solid indication of the risks of data corruption on that content.[33]

Despite the increased cost and spatial demand, uncompressed audiovisual files should be considered ideal for long-term digital preservation: "It's [sic] simplicity and resilience through inherent redundancy makes it a relatively reliable and long-lived way to store content. It is less likely to become obsolete and less likely to be affected seriously by data corruption before it does become obsolete."[34] Yet one might wonder whether the larger size of an uncompressed audiovisual file (i.e. a larger bit stream) puts the file at a greater risk of having a bit or byte corrupted. Wright acknowledges that this implication is indeed present: "For example, if the bit corruption rates of 10^{-9} reported in the CERN study occurred in the audiovisual domain, then for data files that are 10^{13} bits in size (approx. 1 TB, which is an hour of uncompressed HD), it would seem inevitable that these files will become corrupted quite rapidly when stored on disk." So why do all large audiovisual files *not* become corrupted? Wright notes that this is due to the fact that corruption is more likely to occur at the block level than the bit level, and is usually

[32] *Ibid.*, 52.
[33] *Ibid.,* 73-80.
[34] Addis. "Long Term Data Integrity for Large Audiovisual Archives." 33.

spatially correlated (i.e. affecting a sequence of blocks rather than a random assortment). Thus, corruption of large audiovisual files is a constant possibility but not an endemic concern.[35]

Attempts both technical and theoretical have been made to curb the risks of silent corruption in audiovisual data. One example, from a technical standpoint, is the JPEG 2000 standard's inclusion of various tools for detection of and protection against data corruption. The standard's core coding system includes optional mechanisms for detecting an eliminating corrupted data, and allows data to be separated to avoid propagation; the JPEG 2000 Wireless option offers solutions through insertion of error-correcting code and redundancy.[36] From a more theoretical position, some have suggested the possible benefit of splitting large files into multiple smaller data chunks, which could then be both easily replicated for redundancy and stored in different locations to avoid total loss of content.[37] More specifically, it has been suggested that audiovisual files might be split in a number of different ways (e.g. by shot, scene, frame, audio stream, video stream, etc.), allowing for data integrity measures to be carried out on a more granular level and allowing discrete parts of a large file to be stored in separate locations, avoiding total loss in the case of bit or disk failure.[38]

Yet it cannot be overlooked that measures such as these are beyond feasibility for many audiovisual archives due to limits in time and resources. The risks inherent to the preservation of such large and complex files on relatively unreliable storage media will not likely subside in the near future. Still, archivists have and continue to take practical measures to mitigate these risks, and though many of these methods have been at the core of digital preservation practice for a significant amount of time, they are worth revisiting and will be discussed below.

---

[35] Wright et al. "The Significance of Storage in the 'Cost of Risk' of Digital Preservation." 117.
[36] Helt et al. "French Cinema Goes IMF." 6.
[37] Addis. "Long Term Data Integrity for Large Audiovisual Archives." 36.
[38] Wright et al. "The Significance of Storage in the 'Cost of Risk' of Digital Preservation." 118.

Mitigating Risk of Data Corruption

PrestoPRIME notes six essential measures that can and should be taken to ensure long-term preservation of digital data: use longer lived storage technology; use more reliable storage technology; make more copies; encode so content is more resilient; use concealment; and check often and fix quickly.[39] Each of these will be elaborated below, not necessarily as recommendations so much as indicators of the various suggestions that have been made in the digital preservation world.

Though off to a rather sluggish start, efforts to introduce longer-lived alternate digital data storage formats to the market have sprouted in recent years due to acknowledgement of the unimproved reliability of hard-disk drives. Projects like Piql,[40] which preserves digital data in human-readable form on motion picture film; and DOTS (Digital Optical Technology System),[41] which preserves human-readable data on a non-magnetic metal alloy tape, boast lifespans of a century or more. Whether or not these products come into wide enough use so as to be affordable, practical, and open for audiovisual archives is yet to be determined, but their introduction to the archival world marks an important example of the ways in which keepers of digital data are attempting to mitigate the risks of unreliable storage technology.

Use of more reliable storage media than HDDs has become common practice in the archival world. It is widely acknowledged that data storage tape—namely LTO—is far more cost effective and reliable than hard-disks, with reports of data loss minimal or entirely nonexistent in comparison.[42] Though not without its own risks (especially given the fact that errors in tape

[39] Addis. "Long Term Data Integrity for Large Audiovisual Archives." 24.
[40] https://www.piql.com/
[41] http://group47.com/
[42] Addis. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. 49-51.

drives are prevalent), storing digital information, with redundancy, on tape has proven to be a far more effective means of digital preservation.

It is no revelatory statement to note that making more copies of digital information keeps it safer. Redundancy in digital preservation is perhaps the most widely-acknowledged measure to ensure long-term survival of digital information, and the most widely practiced in digital repositories. With that in mind, one might consider instead the ways in which digital redundancy might approach an ideal. For one, it has been suggested that replicas be stored on different storage systems or media.[43] This protects against failures that affect a particular storage system or are associated with a specific storage medium. Another suggestion is to keep three backups of each dataset—one study shows that for audiovisual data, in all cases surveyed, those who retained three copies of their data (one online or near-online for use, two for backup) had never experienced irreplaceable loss for that data.[44]

The potential consequences of storing data in compressed form have already been elaborated above. Encoding to ensure data resiliency over time—and to minimize artifacts in files which have experienced bit corruption—means encoding with a scheme that keeps the bits as close to their original, uncompressed stream as possible. In those many practical environments in which uncompressed digital preservation is not feasible, one might instead consider using file wrappers and encoding schemes that are open and have proven to be resilient in managing large, complex datasets.

[43] Rosenthal. "Keeping Bits Safe: How Hard Can It Be?" 51.
[44] J-H Chenot and C. Bauer. *Data Damage and Its Consequences on Usability*. Deliverable no. D2.1. Digital AV Media Damage Prevention and Repair. DAVID Consortium, 2013: 21.

Employing error concealment techniques—such as interpolation of digital video frames to "correct" a corrupted frame or block of data—are contentious and need further research. As such, they will not be elaborated on here.

Finally, regular data auditing is suggested to promote more rapid identification of data corruption and, in turn, greater likelihood of correcting their effects. Fixity checking through checksum generation and verification is a widely-employed method of data auditing. Simple audits of the sort to identify errors and replace corrupted data, if carried out frequently and regularly, can go a long way in increasing the life expectancy of a bit.[45] In larger data storage infrastructures, auditor units can be programmed to repeatedly crawl servers to test data integrity, isolating and replacing corrupted data when detected, and ensuring that all copies are consistent and up to date.[46]

Other methods for ensuring long-term data integrity and mitigating the risk of bit corruption are myriad, and all might benefit from further study. These range from the simple, such as IBM's suggestion to store redundant metadata as parity information in RAID systems;[47] to the more complex, such as the idea of digital vellum, by which entire digital data ecosystems would be preserved for the long-term.[48] Whether or not any of these measures become widely adopted in digital audiovisual preservation—or digital preservation in general—understanding the full range of options at the disposal of archivists for mitigating the risk of data corruption is essential to the practice.

---

[45] Rosenthal. "Keeping Bits Safe: How Hard Can It Be?" 51.
[46] Z. Zou and Q. Kong. "Secure Provable Data Possession for Big Data Storage." In *Big Data and Smart Service Systems*, 37. Elsevier Science & Technology, 2016.
[47] Hafner et al. "Undetected Disk Errors in RAID Arrays." 421-23.
[48] Marc Kosciejew, Ph.D. "Digital Vellum and Other Cures for Bit Rot." *Information Management Journal* 49, no. 3 (May/June 2015): 20-25.

Conclusion

The significance of bit rot and other silent corruptions in digital repositories is admittedly low compared to other data loss threats. Yet the dangers of even a single corrupt bit—and of that corruption going unnoticed for long periods of time—cannot be overlooked. In Heydegger's words: "A single bit can be extremely significant for robustness: the corruption of just one single bit out of the entirety of a million bits proves to be destructive for information consistency."[49] Small, silent corruptions' manifestations in digital audiovisual data prove to be potentially catastrophic to content integrity, and the risks inherent to hard-disk drive storage have not improved in tandem with increases in their data capacity. Audiovisual archivists have to field these concerns more than most, their large file sizes and complex file formats imbued with an inherently greater risk. Understanding when, where, how, and why small, silent corruptions occur—and in turn understanding the measures that can be taken to mitigate their risk—are essential considerations in the long-term preservation of digital data.

[49] V. Heydegger. "Just One Bit in a Million: On the Effects of Data Corruption in Files." *Lecture Notes in Computer Science*, 2009: 321.

Works Cited

Addis, Matthew. "Long Term Data Integrity for Large Audiovisual Archives." Proceedings of 2010 Joint Technical Symposium, 66th Congress of the International Federation of Film Archives (FIAF), Norway, Oslo. PrestoPRIME, 2010.

Addis, Matthew. *Threats to Data Integrity from Use of Large-Scale Data Management Environments*. Publication no. 3.2.1. PrestoPRIME. 2010.

Bairavasundaram, Lakshmi N., Garth R. Goodson, Bianca Schroeder, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. "An Analysis of Data Corruption in the Storage Stack." *ACM Transactions on Storage* 4, no. 3 (2008): 8:1-:28.

Baker, Mary, Mehul Shah, David S.H. Rosenthal, Mema Roussopoulos, Petros Maniatis, TJ Giuli, and Prashanth Bungale. "A Fresh Look at the Reliability of Long-Term Digital Storage." Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems, Belgium, Leuven. 2006.

Chenot, J-H, and C. Bauer. *Data Damage and Its Consequences on Usability*. Deliverable no. D2.1. Digital AV Media Damage Prevention and Repair. DAVID Consortium, 2013.

Dong, Jielin, ed. *Network Dictionary*. Javvin Technologies, 2007. 68.

Elerath, Jon. "Hard-Disk Drives: The Good, the Bad, and the Ugly." *Communications of the ACM* 52, no. 6 (June 2009): 38-45.

Hacker Dictionary. http://www.hacker-dictionary.com/terms/bit-rot.

Hafner, J.L., V. Deenadhayalan, W. Belluomini, and K. Rao. "Undetected Disk Errors in RAID Arrays." *IBM Journal of Research and Development* 52, no. 4-5 (2008): 413-25.

Helt, François, Benoît Février, Frantz Delbecque, Xavier Brachet, Hans-Nikolas Locher, and Marc Bourhis. "French Cinema Goes IMF." SMPTE 2013 Annual Technical Conference & Exhibition, White Plains, NY. 2013.

Heydegger, V. "Just One Bit in a Million: On the Effects of Data Corruption in Files." *Lecture Notes in Computer Science*, 2009: 315-26.

Kosciejew, Marc, Ph.D. "Digital Vellum and Other Cures for Bit Rot." *Information Management Journal* 49, no. 3 (May/June 2015): 20-25.

Rosenthal, David S.H. "Keeping Bits Safe: How Hard Can It Be?" *Communications of the ACM* 53, no. 11 (November 2010): 47-55.

Rothenberg, Jeff. "Ensuring the Longevity of Digital Information." *CLIR Resources*, February 22, 1999.

Schulz, Greg. "Protecting Data." In *Resilient Storage Networks: Designing Flexible Scalable Data Infrastructures*, 313-26. 1st ed. Digital Press Storage Technology. Burlington, MA: Elsevier Science & Technology, 2004.

Toigo, Jon William. "Bit Rot Eroding RAID." *Storage Magazine*, July 2012.

Wright, Richard, Ant Miller, and Matthew Addis. "The Significance of Storage in the 'Cost of Risk' of Digital Preservation." *The International Journal of Digital Curation* 4, no. 3 (2009): 104-22.

You, Lawrence L., Kristal T. Pollack, and Darrell D.E. Long. "PRESIDIO: A Framework for Efficient Archival Data Storage." *ACM Transactions on Storage* 7, no. 2 (July 2011): 6:1-:60.

Zou, Z., and Q. Kong. "Secure Provable Data Possession for Big Data Storage." In *Big Data and Smart Service Systems*, 27-41. Elsevier Science & Technology, 2016.