Digital Vellum and Other Cures for Bit Rot

Marc Kosciejew, Ph.D.

In the 20 years since digital preservation expert Jeff Rothenberg sounded the alarm about the danger of losing digital information – and our historical legacy – several methods of ensuring long-term digital preservation have been explored. Learn about the challenges and potential solutions in this article.

Late last year an archivist accidentally discovered hidden within the pages of a Victorian-era scrapbook a rare 12th century edition of the Magna Carta issued by King Edward I, according to reports in the *Daily Mail* and *Smithsonian.com*. It is believed to have been placed there by a British Museum official sometime in the late 1800s and forgotten within the city archives of the English town of Sandwich.

Although this copy of one of the world's most famous historical and legal documents is in a battered condition – torn, water-stained, and missing sections – it remains relatively durable and legible; in fact, the document's handwriting, layout, and date (which survives at the bottom) permitted Magna Carta Project historians to verify its authenticity.

The discovery of this rare edition and the Victorian-era scrapbook serves as a timely reminder of the importance of the long-term preservation of information. These physical documents survived because of their paper and inscriptions, which, technically, can be handled, read, and used in much the same way today as in their respective pasts. An individual can still consult and regard each document much as Edward I did more than 700 years ago or the Victorian museum official did more than 100 years ago. Indeed, physical and print materials continue to prove their relative permanence despite technological progress and time's ravages.

The Digital Future

It is doubtful that individuals will be able to engage with today's digitized materials in 10 or 20 years, let alone 100 years, as they still will be able to interact with these centuries-old physical documents. Alarmingly, as more information is digitized, either by migrating physical and print artifacts and resources to online display or, increasingly, by creating it in digital formats rather than print, much of this information may become lost for future generations due to the corrosive phenomenon of bit rot.

> …bit rot's destructive ravages have already begun… erasing our recent history

Bit rot refers to the irrevocable degradation or loss of digital information when the infrastructure (the hardware and software) required to access, interpret, view, and use this information is no longer available or executable.

As Jeff Rothenberg warned in his seminal 1995 *Scientific American* article on digitization and preservation, "Ensuring the Longevity of Digital Information," "We are in imminent danger of losing [digital information and documents] even as we create them. We must invest careful thought and significant effort if we are to preserve these documents for the future. If we are unwilling to make this investment, we risk substantial practical loss, as well as the condemnation of our progeny for thoughtlessly consigning to oblivion a unique historical legacy."

Since he raised this warning, the danger has only grown in nature, size, and scope. And it continues to metastasize at an accelerating rate. If left unaddressed and unchecked, bit rot has the potential to wipe out contemporary (and future) history – and sooner than we may realize.

Brief History of Bit Rot

Bit rot is not a new phenomenon; it has been around for decades. In *The Discipline of Organizing*, Robert J. Glushko describes how "computer data began to be stored on magnetic tape and hard disk drives six decades ago, on floppy disks four decades ago, on CDs three decades ago, on DVDs two decades ago, on solid-state drives half-a-decade ago, and in 'cloud-based' or 'virtual' storage environments in the last decade." Indeed, bit rot's destructive ravages have already begun affecting our information and erasing our recent history.

VHS Tapes and Players

Video home system (VHS) tapes and players, for example, have basically become obsolete technologies. The recorded content and memories on VHS tapes are now becoming, or in many cases have become, lost due to equipment decay and disappearance. It was only about 15 years ago that VHS tapes and players were common technologies in many homes, offices, and schools, used for various purposes, including entertainment, work, and education.

Today, it is more likely to find VHS tapes and players in some kind of museum; even if one happens to have a VHS player and tapes to play on it, most televisions do not support it, lacking the appropriate scart socket feature to take the VHS video input.

CDs, DVDs, and Players

Another example of bit rot involves compact discs (CDs) and CD players. Although not quite "gone" like VHS technologies, they are quickly vanishing. Many people no longer purchase CDs, either for music or storage. Music is now commonly purchased or pirated online through digital music services. People are accessing and listening to their favorite songs in MP3 format on their mobile devices instead of on CDs played in CD players.

Information storage is now placed in so-called cloud computing services and transported via universal serial bus (USB) sticks – which, interestingly, are themselves slowly, perhaps

still imperceptibly, becoming "rotten"– instead of burnt onto, stored, and carried on CDs. Many companies and retailers no longer produce or sell CDs or CD players.

Much of the information contained within these technologies – and on the floppy disks, 8-track tapes, and cassette tapes and players that also used to be prevalent – is in many respects inaccessible and unusable. Unlike the recently discovered 700-year-old Magna Carta, this information is effectively lost.

Current ICT Issues

It is not only hindsight that reveals bit rot's ravages to information. A closer look at our current information communication technologies (ICTs) exposes bit rot's emergence. Digital video discs (DVDs) and Blu-Ray technologies have arguably begun a similar and as steep a descent as their predecessors; many computers and laptops no longer feature DVD or CD drives. In fact, the desktop computer itself is a gradually disappearing technology.

If the information stored on technologies that are approaching obsolescence is not digitized in the newest formats with the most up-to-date software, most of it will become lost. We must begin to develop ways to enable and ensure that digital information on current, widely used ICTs can migrate along with technological progress.

Bit rot can also emerge in the economic and managerial vagaries of business decisions and operations. For example, digital information is recorded in mainly proprietary digital documentary file types – such as .docs, .docx, .jpgs, and .pdfs – that are constantly updated. If the companies that own these file types and programs change them or if the companies are shuttered or choose not to continue to support the applications, people will not be able to access their information because the programs will not be available.

> "Preservation is often a key motive for digitization, but digitization alone is not preservation."

Challenges for Digital Preservation

In his keynote address at the 11th European Conference on Digital Libraries in 2007 entitled "Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries," Seamus Ross analyzes the challenges confronting digital preservation, including technological obsolescence. He observes:

> Digital objects break. Digital materials occur in a rich array of types and representations. They are bound to varying degrees to the specific application packages (or hardware) that were used to create or manage them. They are prone to corruption. They are easily misidentified. They are generally poorly described or annotated; they often have insufficient metadata attached to them to avoid their gradual susceptibility to syntactical and semantic glaucoma. Where they do have sufficient ancillary data, these data are frequently time constrained. Beyond maintaining the intactness of the bit stream (which is fairly straightforward), the long-term curation and preservation of digital materials is for the most part…a labor-intensive artisan or craft activity.

In his book, Glushko notes, "Preservation is often a key motive for digitization, but digitization alone is not preservation. Digitization creates preservation challenges because technological obsolescence of computer software and hardware require ongoing efforts to ensure the digitized resources can be accessed."

Glushko presents three major challenges of digitization and preservation:

1. Technological obsolescence
2. Expected useful lifetimes of physical storage media
3. The (un)availability of software and its associated computing environment

Technological Obsolescence

The first challenge of technological obsolescence "is a result of the relentless evolution of the physical media and environments used to store digital information in both institutional or business and personal organizing systems," Glushko says. He further explains that "as the capacity of storage technologies grows from kilobytes to megabytes to gigabytes to terabytes to petabytes, economic and efficiency considerations often make the case to adopt new technology to store newly acquired digital resources and raise questions about what to do with the existing ones."

Limited Useful Lifetimes of Media

The second challenge of expected useful lifetimes of storage media is, according to Glushko, seemingly paradoxical. He states that "even as the capacities of digital storage technologies increase at a staggering pace, the expected useful lifetimes of the physical storage media are measured in years or at best in decades."

In contrast, books that are created on high-quality and acid-free paper can last for centuries. Glushko notes that "the contrast between printed and digital resources is striking; books on library shelves don't disappear if

no one uses them, but digital data can be lost just because no one wants access to it within a year or two after its creation."

Unavailability of Software

The third challenge is the (un)availability of software and its associated computing environment required to use and "parse and interpret the resource at the time of preservation might no longer be available when the resource needs to be accessed." Glushko explains that "software and services that convert documents from old formats to new ones are widely available, but they are only useful if the old file can be read from its legacy storage medium."

For example, only a couple of decades ago, most digital documents were created using Word Perfect word processing software; presently, however, most are created using Microsoft Word and few with Word Perfect.

These challenges for digital preservation are of particular concern because of our growing reliance on ICTs for institutional and individual activities, documentation, and interactions and because many of our information assets and resources are now in digital formats.

A Forgotten Generation

In recent coverage by the BBC and in *The New York Times*, Google's vice-president, Vint Cerf, echoing Rothenberg's earlier warnings about the precarious longevity of digital information, warned that bit rot could result in a "forgotten generation, or even a forgotten century."

Cerf, who is also considered one of the fathers of the Internet, observed that "when you think about the quantity of documentation from our daily lives that is captured in digital form, like our interactions by e-mail, people's tweets, and all of the world wide web, it's clear that [because of bit rot] we stand to lose an awful lot of our history."

> …bit rot could result in a "forgotten generation, or even a forgotten century."

Yet, we migrate and/or create information online with the somewhat mistaken belief, or hope, that it will be better protected and preserved in the long term.

Cerf worries that we are "throwing all of our data into what could become an information black hole." This information black hole is the irony of digitization. It is often argued that digitization will ensure preservation of our information. We digitize information for preservation purposes without realizing that much of it may become useless or lost with technological progress. Backwards compatibility, for instance, is not always guaranteed or reliable.

Cerf notes, "We digitize things because we think we will preserve them, but what we don't understand is that unless we take other steps, those digital versions may not be any better, and may even be worse, than the artefacts that we digitized."

Bit rot, in other words, can render our supposedly preserved information unreadable or unintelligible, thereby erasing ourselves and our history – our digital artefacts, happenings, and memories – from future generations.

Digital Vellum: Panacea or Pipe Dream?

In a BBC interview with noted science journalist Pallab Ghosh, Cerf proposes that digital vellum is a possible solution to help prevent, or at least mitigate, bit rot's effects. Digital vellum could preserve old hardware and software to prevent their obsolescence and ensure the information dependent upon them will be recoverable and usable.

Digital Vellum Explained

He explains that *digital vellum* would involve a process of taking "an X-ray snapshot of the content and the application and the operating system together, with a description of the machine that it runs on, and preserve that for long periods of time. And that digital snapshot will recreate the past, in the future." This X-ray snapshot "should be transportable from one place to another. So I should be able to move it from the Google cloud to some other cloud, or move it into a machine I have."

Cerf continues, "No matter what the medium is in which digital bits are recorded, how long will we be able to read them, and how long will we make sense out of them? So the issue here is not just the physical bits, but what do they mean. If you use a program, for example, to create a spreadsheet, you have a complex file. You store the file away and you hold onto it for twenty or thirty years. And even pretending you can read the disc again, do you have the software that knows what the bits mean? So the digital vellum idea is not just physical medium, but an ecosystem which is able to remember what bits mean over long periods of time."

Standards Needed

A major component of digital vellum would be standardized descriptions to help ensure information remains accessible, understandable, and usable. Cerf states that "the key here is when you move those bits [of information] from one place to another, that you still know how to unpack them to correctly interpret the different parts. That is all achievable if we standardize the descriptions. And that's the key issue here: how

do I ensure in the distant future that the standards are still known, and I can still interpret this carefully constructed X-ray snapshot?"

It is not enough to have a picture. The structure of the information in the snapshot must be standardized and still known, or at least accessible and understandable for consultation and reference, to enable its use and ensure its own migration in the continual progress of technological change and development.

Results So Far

The practical applications of Cerf's digital vellum have begun to be explored. According to *The Guardian*'s Ian Sample, Carnegie Mellon University researchers are taking digital snapshots of old computer hard drives while they run different software programs, which are then uploaded to a new computer that can mimic the photographed one, resulting in its ability to read otherwise defunct files.

The researchers have been able to "resurrect" an early version of Word Perfect, Mystery House (Apple II's graphic adventure game from 1982), and Doom (the original first-person shooter game from 1993). The concept and practicality of digital vellum, however, remain works in progress. As Cerf acknowledges, "It's not without its rough edges but the major concept has been shown to work."

Economic, Legal Issues

Digital vellum, in addition to being a conceptual, practical, and technological work in progress, is also confronted by economic and legal issues. First, digital vellum is not necessarily a viable commercial project or undertaking for many businesses, at least not in the short or medium term. Its further development requires committed financial, knowledge, and other technological and computational investments unlikely to be made by most companies. There remain few business incentives to motivate many companies to work on this concept.

The concept and practicality of digital vellum, however, remain works in progress

There are some disincentives to developing digital vellum. Since most digital information depends upon proprietary products – in many cases with proprietary codes, copyright, patents, licensing, and other legalities technically included or incorporated into the information itself – it remains largely, if not wholly, dependent upon the owners' commitment to keep the required products updated or in existence. Buying these legal rights and protections from companies that close, sell the products and designs to other parties, discontinue product designs or entire lines, or stop making improvements or supporting updates to them could be prohibitive.

In other words, financial costs and legal constraints may help contribute to bit rot. Cerf therefore proposes that "the rights of preservation might need to be incorporated into our thinking about things like copyright and patents and licensing. We're talking about preserving them for hundreds to thousands of years."

Other Possible Solutions

Other possible solutions for preventing bit rot include open formats, platform diversification, checksums, microform, and greater research.

Open Formats

Complementing Cerf's call for digital vellum, open formats, for example, for digital files and storage could be employed to ensure that digital information is accessible, interpretable, and usable in the long term. Open formats would have open standards not tied to or owned by any company, and they would be freely available, allowing anyone to use and build software to use for their purposes. Like digital vellum, however, this possible solution is perhaps the most difficult to convince companies that own the proprietary programs to adopt.

Platform Diversification

Platform diversification could also help mitigate bit rot's effects by making many digital copies of information available in multiple locations and programs. Because bit rot affects different hardware and software at different times – everything does not degrade at the same rate – digital copies do not rot simultaneously. Creating digital copies for diverse platforms can help the digital preservation process by allowing more time to prepare and implement strategies to address bit rot before and as it emerges.

Checksums

Checksums are another possible solution to help combat bit rot. Checksums show the integrity of a digital file's constitution; they are used to check whether the same number of bits arrived and monitor it over time.

For instance, Rhian Sasseen, in a March 2014 *Al Jazeera* article on digital preservation challenges entitled "Bit rot: The Internet never forgets, or does it?" describes how the Library of Congress uses checksums to help maintain and monitor digital materials over time.

Digital information "received on more outdated and vulnerable formats, such as personal hard drives or CD-ROMS, are transferred to disc images, after which labels are created and photographed for documentation purposes. The labels are monitored for degradation alongside the data they describe," he wrote.

Microform

Microform, including both film and fiche, is arguably one of the most durable and useful technologies for long-term preservation of information. When microform is constructed from quality materials, exposed and processed to international standards, and placed in proper protective enclosures under appropriate conditions, it can survive upwards of 500 years. Its longevity can therefore be more secure and stable than most digital or digitization preservation efforts or processes.

It is also important to note that committed and ongoing research is crucial to battling bit rot. Ross in his previously cited article, for example, argues for greater research of digital preservation's challenges in order to "better understand what we might do to alleviate obstacles to the longevity of digital materials [and also] to define the uncertainties related to digital preservation and to convert these uncertainties into known, measureable, and mitigatable risks."

Ross presents nine research challenges for digital preservation that need more detailed focus and understanding: restoration, conservation, collection and repository management, risk management as preservation, functionality and interpretability, interoperability, automation in preservation, contextualization, and storage technologies and methods.

A Remembered Generation

Although bit rot presents serious threats to the successful long-term preservation of digital information, there are various conceptual and practical developments being explored to combat, mitigate, and prevent its destructive effects. Cerf's concept of digital vellum as one possible solution to bit rot's threats shows some practical promise.

Open formats, platform diversification, and checksums also offer some helpful possibilities in combatting bit rot. But, there are numerous research avenues that must be explored in more detail and depth to help provide other possible directions for anticipating and addressing bit rot.

Taking some of these steps today can help tomorrow remember. We do not have to become a forgotten generation. Planning and preparing for bit rot can help ensure that we are and remain a remembered generation. **END**

Marc Kosciejew, Ph.D., can be contacted at mkosciej@gmail.com. *See his bio on page 47.*