A reprint from

American Scientist
the magazine of Sigma Xi, The Scientific Research Society

Avoiding a Digital Dark Age

Kurt D. Bollacker

When I was a boy, I discovered a magnetic reel-to-reel audio tape recorder that my father had used to create "audio letters" to my mother while he was serving in the Vietnam War. To my delight (and his horror), I could listen to many of the old tapes he had made a decade before. Even better, I could make recordings myself and listen to them. However, all of my father's tapes were decaying to some degree—flaking, stretching and breaking when played. It was clear that these tapes would not last forever, so I copied a few of them to new cassette tapes. While playing back the cassettes, I noticed that some of the sound quality was lost in the copying process. I wondered how many times I could make a copy before there was nothing left but a murky hiss.

A decade later in the 1980s I was in high school making backups of the hard drive of my PC onto 5-¼-inch floppy disks. I thought that because digital copies were "perfect," and I could make perfect copies of perfect copies, I couldn't lose my data, except by accident. I continued to believe that until years later in college, when I tried to restore my backup of 70 floppy disks onto a new PC. To my dismay, I discovered that I had lost the floppy disk containing the backup program itself, and thus could not restore my data. Some investigation revealed that the company that made the software had long since gone out of business. Requests on electronic bulletin board systems and searches on Usenet turned up nothing useful. Although all of the data on them may have survived, my disks were useless because of the proprietary encoding scheme used by my backup program.

Data longevity depends on both the storage medium and the ability to decipher the information

The Dead Sea scrolls, made out of still-readable parchment and papyrus, are believed to have been created more than 2,000 years ago. Yet my barely 10-year-old digital floppy disks were essentially lost. I was furious! How had the shiny new world of digital data, which I had been taught was so superior to the old "analog" world, failed me? I wondered: Had I had simply misplaced my faith, or was I missing something?

Over the course of the 20th century and into the 21st, an increasing proportion of the information we create and use has been in the form of digital data. Many (most?) of us have given up writing messages on paper, instead adopting electronic formats, and have exchanged film-based photographic cameras for digital ones. Will those precious family photographs and letters—that is, email messages—created today survive for future generations, or will they suffer a sad fate like my backup floppy disks? It seems unavoidable that most of the data in our future will be digital, so it behooves us to understand how to manage and preserve digital data so we can avoid what some have called the "digital dark age." This is the idea—or fear!—that if we cannot learn to explicitly save our digital data, we will lose that data and, with it, the record that future generations might use to remember and understand us.

Save Our Bits!

The general problem of data preservation is twofold. The first matter is preservation of the data itself: The physical media on which data are written must be preserved, and this media must continue to accurately hold the data that are entrusted to it. This problem is the same for analog and digital media, but unless we are careful, digital media can be more fragile.

The second part of the equation is the comprehensibility of the data. Even if the storage medium survives perfectly, it will be of no use unless we can read and understand the data on it. With most analog technologies such as photographic prints and paper text documents, one can look directly at the medium to access the information. With all digital media, a machine and software are required to read and translate the data into a human-observable and comprehensible form. If the machine or software is lost, the data are likely to be unavailable or, effectively, lost as well.

Preservation

Unlike the many venerable institutions that have for centuries refined their techniques for preserving analog data on clay, stone, ceramic or paper, we have no corresponding reservoir of historical wisdom to teach us how to save our digital data. That does not mean there is nothing to learn from the past, only that we must work a little harder to find it. We can start by briefly looking at the historical trends and advances in data representation in human history. We can also turn to nature for a few important lessons.

The earliest known human records are millennia-old physical scrapings on whatever hard materials were available. This medium was often stone, dried clay, bone, bamboo strips or even tortoise shells. These substances were very durable—indeed, some specimens have

Over the past two decades, Kurt D. Bollacker has romped through the fields of artificial intelligence, digital libraries, linguistics, databases and electrocardiology. He currently is the digital research director of the Long Now Foundation and gets his hands dirty as a freelance data miner and builder of collaborative knowledge-creation tools. He also works on the Rosetta Project. He received his Ph.D. in computer engineering from the University of Texas at Austin in 1998. Email: kurt@longnow.org

survived for more than 5,000 years. However, stone tablets were heavy and bulky, and thus not very practical.

Possibly the first big advance in data representation was the invention of papyrus in Egypt about 5,500 years ago. Paper was lighter and easier to make, and it took up considerably less space. It worked so well that paper and its variants, such as parchment and vellum, served as the primary repositories for most of the world's information until the advent of the technological revolution of the 20th century.

Technology brought us photographic film, analog phonographic records, magnetic tapes and disks, optical recording, and a myriad of exotic, experimental and often short-lived data media. These technologies were able to represent data for which paper cannot easily be used (video, for example). The successful ones were also usually smaller, faster, cheaper and easier to use for their intended applications. In the last half of the 20th century, a large part of this advancement included a transition from analog to digital representations of data.

Even a brief investigation into a small sampling of information-storage media technologies throughout history quickly uncovers much dispute regarding how long a single piece of each type of media might survive. Such uncertainty cannot be settled without a time machine, but we can make reasonable guesses based on several sources of varying reliability. If we look at the time of invention, the estimated lifespan of a single piece of each type of media and the encoding method (analog or digital) for each type of data storage *(see the table, above right)*, we can see that new media types tend to have shorter lifespans than older ones, and digital types have shorter lifespans than analog ones. Why are these new media types less durable? Shouldn't technology be getting better rather than worse? This mystery clamors for a little investigation.

To better understand the nature of and differences between analog and digital data encoding, let us use the example of magnetic tape, because it is one of the oldest media that has been used in both analog and digital domains. First, let's look at the relationship between information density and data-loss risk. A standard 90-minute analog compact cassette is 0.00381 meters wide by about 129 meters long, and a typical digital audio tape (DAT) is 0.004 meters wide by 60 meters long. For audio encodings of similar quality (such as 16 bit, 44.1 kilohertz for digital, or 47.6 millimeters per second for analog), the DAT can record 500 minutes of stereo audio data per square meter of recordable surface, whereas the analog cassette can record 184 minutes per square meter. This means the DAT holds data about 2.7 times more densely than the cassette. The second table *(below)* gives this comparison for several common consumer audio-recording media types. Furthermore, disk technologies tend to hold data more densely than tapes, so it is no surprise that magnetic tape has all but disappeared from the consumer marketplace.

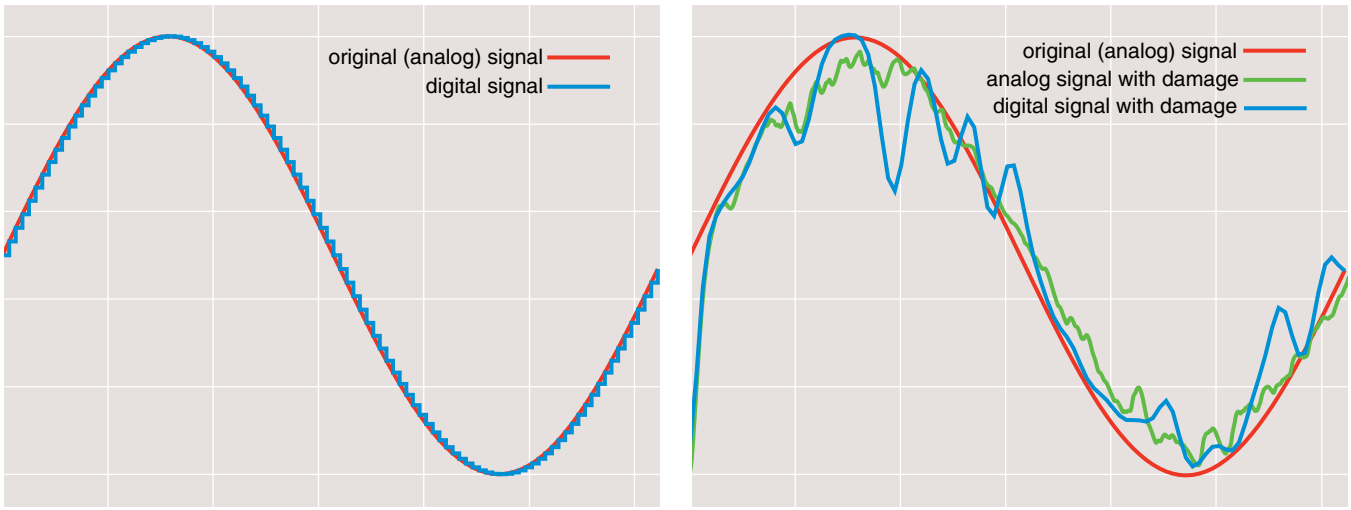
type of medium	data medium	approximate year of invention	ideal expected lifetime of medium
analog	clay/stone tablet	8000 BC	>4,000 years
analog	pigment on paper	3500 BC	>2,000 years
analog	oil painting	600	centuries
analog	silver halide black and white photographic film	1820	>100 years
analog	modern color photographic film	1860	decades
analog	phonograph record	1877	>120 years
analog/digital	magnetic tape	1928	decades
analog/digital	magnetic disk	1950	3–20 years
analog/digital	polycarbonate optical WORM disk	1990	5–20 years

When we compare the different data-storage media that have appeared over the course of human history, a trend emerges: Digital data types are expected to have shorter lifetimes than analog ones.

However, enhanced recording density is a double-edged sword. Assume that for each medium a square millimeter of surface is completely corrupted. Common sense tells us that media that hold more data in this square millimeter would experience more actual data loss; thus for a given amount of lost physical medium, more data will be lost from digital formats. There is a way to design digital encoding with a lower data density so as to avoid this problem, but it is not often used. Why? Cost and efficiency: It is usually cheaper to store data on digital media because of the increased density.

type of medium	audio data medium	recording capacity (minutes per square meter)
analog	6.35 millimeter wide 190.5 millimeters per second reel-to-reel magnetic tape	13.8
analog	33-1/3 RPM vinyl album	411
analog	90-minute audio cassette	184
digital	compact disk (CD)	8,060
digital	60-meter digital audio tape (DAT)	500
digital	2 terabyte 89-millimeter hard drive	4,680,000

As technology has advanced, the density of data storage on analog and, subsequently, digital recording media has tended to increase. The downside of packing in data, however, is that more of the information will be lost if a portion of the recording medium becomes damaged.



A simple audio tone is represented as a sine wave in an analog signal, and as a similar wave but with an approximated stepped shape in a digital signal *(left)*. If the data receive simulated damage, the analog signal output is more resistant to damage than the digital one, which has wilder swings and higher error peaks *(right)*. This result is largely because in a digital recording, all bits do not have the same worth, so damage causes random output error.

A possibly more important difference between digital and analog media comes from the intrinsic techniques that comprise their data representations. Analog is simply that—a physical analog of the data recorded. In the case of analog audio recordings on tape, the amplitude of the audio signal is represented as an amplitude in the magnetization of a point on the tape. If the tape is damaged, we hear a distortion, or "noise," in the signal as it is played back. In general, the worse the damage, the worse the noise, but it is a smooth transition known as *graceful degradation.* This is a common property of a system that exhibits *fault tolerance,* so that partial failure of a system does not mean total failure.

Unlike in the analog world, digital data representations do not inherently degrade gracefully, because digital encoding methods represent data as a string of binary digits ("bits"). In all digital symbol number systems, some digits are worth more than others. A common digital encoding mechanism, pulse code modulation (PCM), represents the total amplitude value of an audio signal as a binary number, so damage to a random bit causes an unpredictable amount of actual damage to the signal.

Let's use software to concoct a simulated experiment that demonstrates this difference. We will compare analog and PCM encoding responses to random damage to a theoretically perfect audiotape and playback system. The first graph in the third figure *(above)* shows analog and PCM representations of a single audio tone, represented as a simple sine wave. In our perfect system, the original audio source signal is identical to the analog encoding. The PCM encoding has a stepped shape showing what is known as *quantization error,* which results from turning a continuous analog signal into a discrete digital signal. This class of error is usually imperceptible in a well-designed system, so we will ignore it for now.

For our comparison, we then randomly damage one-eighth of the simulated perfect tape so that the damaged parts have a random amplitude re-

ZIP code digit value	POSTNET code	POSTNET code with missing middle digit
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		

The U.S. Postal Service uses an encoding scheme for ZIP code numbers called POSTNET that uses an error-correcting code. Each decimal digit is represented as five bars. If, say, the middle bar disappears, each number is still distinguishable from all the others.

The Phaistos Disk, housed at the Heraklion Archaeological Museum in Crete, is well preserved and all its data are visible, but the information is essentially lost because the language in which it is written has been forgotten. (Photograph courtesy of Wikimedia Commons.)

sponse. The second graph in the third figure *(facing page, top)* shows the effect of the damage on the analog and digital encoding schemes. We use a common device called a *low-pass filter* to help minimize the effect of the damage on our simulated output. Comparing the original undamaged audio signal to the reconstructions of the damaged analog and digital signals shows that, although both the analog and digital recordings are distorted, the digital recording has wilder swings and higher error peaks than the analog one.

But digital media are supposed to be better, so what's wrong here? The answer is that analog data-encoding techniques are intrinsically more robust in cases of media damage than are naive digital-encoding schemes because of their inherent redundancy—there's more to them, because they're continuous signals. That does not mean digital encodings are worse; rather, it's just that we have to do more work to build a better system. Luckily, that is not too hard. A very common way to do this is to use a binary-number representation that does not mind if a few bits are missing or broken.

One important example where this technique is used is known as an error correcting code (ECC). A commonly used ECC is the U.S. Postal Service's POSTNET (Postal Numeric Encoding Technique), which represents ZIP codes on the front of posted envelopes. In this scheme, each decimal digit is represented as five binary digits, shown as long or short printed bars *(facing page, bottom)*. If any single bar for any decimal digit were missing or incorrect, the representation would still not be confused with that of any other digit. For example, in the rightmost column of the table, the middle bar for each number has been erased, yet none of the numbers is mistakable for any of the others.

Although there are limits to any specific ECC, in general, any digital-encoding scheme can be made as robust as desired against random errors by choosing an appropriate ECC. This is a basic result from the field of information theory, pioneered by Claude Shannon in the middle of the 20th century. However, whichever ECC we choose, there is an economic tradeoff: More redundancy usually means less efficiency.

Nature can also serve as a guide to the preservation of digital data. The digital data represented in the DNA of living creatures is copied into descendents, with only very rare errors when they reproduce. Bad copies (with destructive mutations) do not tend to survive. Similarly, we can copy digital data from medium to medium with very little or no error over a large number of generations. We can use easy and effective techniques to see whether a copy has errors, and if so, we can make another copy. For instance, a common error-catching program is called a *checksum function*: The algorithm breaks the data into binary numbers of arbitrary length and then adds them in some fashion to create a total, which can be compared to the total in the copied data. If the totals don't match, there was likely an accidental error in copying. Error-free copying is not possible with analog data: Each generation of copies is worse than the one before, as I learned from my father's reel-to-reel audiotapes.

Because any single piece of digital media tends to have a relatively short lifetime, we will have to make copies far more often than has been historically required of analog media. Like species in nature, a copy of data that is more easily "reproduced" before it dies makes the data more likely to survive. This notion of *data promiscuousness* is helpful in thinking about preserving our own data. As an example, compare storage on a typical PC hard drive to that of a magnetic tape. Typically, hard drives are installed in a PC and used frequently until they die or are replaced. Tapes are usually written to only a few times (often as a backup, ironically) and then placed on a shelf. If a hard drive starts to fail, the user is likely to notice and can quickly make a copy. If a tape on a shelf starts to die, there is no easy way for the user to know, so very often the data on the tape perishes silently, likely to the future disappointment of the user.

Comprehensibility

In the 1960s, NASA launched *Lunar Orbiter 1*, which took breathtaking, famous photographs of the Earth juxtaposed with the Moon. In their rush to get astronauts to the Moon, NASA engineers created a mountain of magnetic tapes containing these important digital images and other space-mission-related data. However, only a specific, rare model of tape drive made for the U.S. military could read these tapes, and at the time (the 1970s to 1980s), NASA had no interest in keeping even one compatible drive in good repair. A heroic NASA archivist kept several donated broken tape drives in her garage for two decades until she was able to gain enough public interest to find experts to repair the drives and help her recover these images.

Contrast this with the opposite problem of the analog Phaistos Disk *(above left)*, which was created some 3,500 years ago and is still in excellent physical condition. All of the data it stores (about 1,300 bits) have been preserved and are easily visible to the human eye. However, this disk shares one unfortunate characteristic with my set of 20-year-old floppy disks: No one can decipher the data on either one. The language in which the Phaistos disk was written has long since been forgotten, just like the software to read my floppies is equally irretrievable.

These two examples demonstrate digital data preservation's other challenge—comprehensibility. In order to survive, digital data must be understandable by both the machine reading them and the software interpreting them. Luckily, the short lifetime of digital media has forced us to gain some experience in solving this problem—the silver lining of the dark clouds of a looming potential digital dark age. There are at least two effective approaches: choosing data representation technologies wisely and creating mechanisms to reach backward in time from the future.

Make Good Choices …

In order to make sure digital data can be understood in the future, ideally we should choose representations for our data for which compatible hardware and software are likely to survive as well. Like species in nature, digital formats that are able to adapt to new environments and threats will tend to

survive. Nature cannot predict the future, but the mechanism of mutation creates different species with different traits, and the fittest prevail.

Because we also can't predict the future to know the best data-representation choices, we try to do as nature does. We can copy our digital data into as many different media, formats and encodings as possible and hope that some survive.

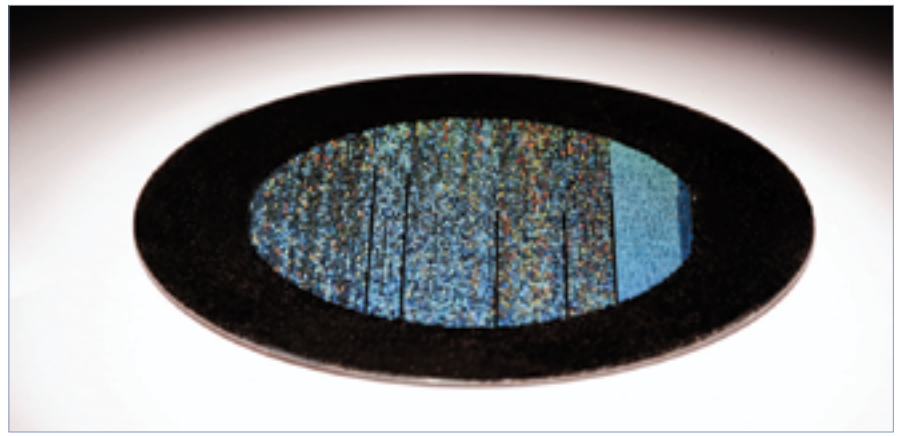
Another way to make good choices is to simply follow the pack. A famous example comes from the 1970s, when two competing standards for home video recording existed: Betamax and VHS. Although Betamax, by many technical measures, was a superior standard and was introduced first, the companies supporting VHS had better business and marketing strategies and eventually won the standards war. Betamax mostly fell into disuse by the late 1980s; VHS survived until the mid-2000s. Thus if a format or media standard is in more common use, it may be a better choice than one that is rare.

… Or Fake It!

Once we've thrown the dice on our data-representation choices, is there anything else we can do? We can hope we will not be stuck for decades, like our NASA archivist, or left with a perfectly readable but incomprehensible Phaistos disk. But what if our scattershot strategy of data representation fails, and we can't read or understand our data with modern hardware and software? A very common approach is to fake it!

If we have old digital media for which no compatible hardware still exists, modern devices sometimes can be substituted. For example, cheap and ubiquitous optical scanners have been commonly used to read old 80-column IBM punchcards. This output solves half of the problem, leaving us with the task of finding hardware to run the software and interpret the data that we are again able to read.

In the late 1950s IBM introduced the IBM 709 computer as a replacement for the older model IBM 704. The many technical improvements in the 709 made it unable to directly run software written for the 704. Because customers did not want either to lose their investment in the old software or to forgo new technological advances, IBM sold what they called an *emulator* module for the 709, which allowed it to pretend to be a 704 for the purposes of running the old software. Emulation is now a common technique used to run old software on new hardware. It does, however, have a problem of recursion—what happens when there is no longer compatible hardware to run the emulator itself? Emulators can by layered like Matryoshka dolls, one running inside another running inside another.



The Rosetta Project aims to preserve all of the world's written languages with a metal disk that could last up to 2,000 years. The disk records miniaturized versions of more than 13,000 pages of text and images, etched onto the surface using techniques similar to computer-chip lithography. (Photograph by Spencer Lowell, courtesy of the Long Now Foundation, http://www.longnow.org.)

Being Practical

Given all of this varied advice, what can we do to save our personal digital data? First and foremost, make regular backup copies onto easily copied media (such as hard drives) and place these copies in different locations. Try reading documents, photos and other media whenever upgrading software or hardware, and convert them to new formats as needed. Lastly, if possible, print out highly important items and store them safely—there seems to be no getting away from occasionally reverting to this "outdated" media type. None of these steps will guarantee the data's survival, but not taking them almost guarantees that the data will be lost, sooner or later. This process does seem to involve a lot more effort than my grandparents went to when shoving photos into a shoebox in the attic decades ago, but perhaps this is one of the costs for the miracles of our digital age.

If all this seems like too much work, there is one last possibility. We could revert our digital data back to an analog form and use traditional media-preservation techniques. An extreme example of this is demonstrated by the Rosetta Project, a scholarly endeavor to preserve parallel texts of all of the world's written languages. The project has created a metal disk *(above)* on which miniaturized versions of more than 13,000 pages of text and images have been etched using techniques similar to computer-chip lithography. It is expected that this disk could last up to 2,000 years because, physically, the disk has more in common with a stone tablet than a modern hard drive. Although this approach should work for some important data, it is much more expensive to use in the short term than almost any practical digital solution and is less capable in some cases (for example, it's not good for audio or video). Perhaps it is better thought of as a cautionary example of what our future might look like if we are not able to make the digital world in which we find ourselves remain successful over time.

Bibliography

Balistier, Thomas. 2000. *The Phaistos Disc: An Account of Its Unsolved Mystery.* New York: Springer-Verlag.

Besen, Stanley M., and Joseph Farrell. 1994. Choosing how to compete: Strategies and tactics in standardization. *Journal of Economic Perspectives* 8:117–131.

Camras, Marvin. 1988. *Magnetic Recording Handbook.* New York: Van Nostrand Reinhold Co.

The IBM 709 Data-Processing System. http://www-03.ibm.com/ibm/history/exhibits/mainframe/mainframe_PP709.html

Koops, Matthias. 1800. *Historical Account of the Substances Which Have Been Used to Describe Events, and to Convey Ideas, from the Earliest Date, to the Invention of Paper.* London: T. Burton.

Pohlmann, Ken C. 1985. *Principles of Digital Audio,* 2nd ed. Carmel, Indiana: Sams/Prentice-Hall Computer Publishing.

The Rosetta Project. http://www.rosettaproject.org

United States Postal Service, Domestic Mail Manual 708.4—Special Standards, Technical Specifications, Barcoding Standards for Letters and Flats.