

Digital longevity

by [Howard Besser](#), 1999

Chapter in Maxine Sitts (ed.) **Handbook for Digital Projects: A Management Tool for Preservation and Access**, Andover MA: Northeast Document Conservation Center, 2000, pages 155-166

With a vast number of resources being committed to reformatting into digital form, we need to begin considering how we can assure that that digital information will continue to be accessible over a prolonged period of time. In this chapter we will first outline the general problem of information in digital form disappearing. We will then look closely at 5 key factors that pose problems for digital longevity. Finally, we will turn our attention to a series of suggestions that are likely to improve the longevity of digital information, focusing primarily on metadata. Though this chapter was written for the digital imaging community, the observations here will be useful for all communities wishing to assure the longevity of any type of digital information.

The Short Life of Digital Information

Though the advent of electronic storage is fairly new, a substantial amount of information stored in electronic form has deteriorated and disappeared. Archives of videotape and audiotape such as fairly recent interviews designed to capture the last cultural remnants of Navajo tribal elders may not be salvageable (Sanders 1997).

Though most people tend to think that (unlike analog information) digital information will last forever, we fail to realize the fragility of digital works. Many large bodies of digital information (such as significant parts of the Viking Mars mission) have been lost due to deterioration of the magnetic tapes that they reside on. But the problem of storage media deterioration pales in comparison with the problems of rapidly changing storage devices and changing file formats. It is almost impossible today to read files off of the 8-inch floppy disks that were popular just 20 years ago, and trying to decode Wordstar files from just a dozen years ago can be a nightmare. Vast amounts of digital information from just 20 years ago is, for all practical purposes, lost.

To prevent further loss, we need to come to grips with the problems of longevity in the digital world. We need to see how preservation in the digital world differs from what we have become accustomed to in the analog world. In the analog world, all of our efforts to preserve a work focused on that work as an artifact. As we begin to engage in preservation of information in digital form, we need to make a conceptual leap from preserving a physical object to preserving informational content that may be completely disembodied from any physical artifact.

In the following sections we will address five key factors that pose digital longevity problems: the Viewing Problem, the Scrambling Problem, the Inter-relation Problem, the Custodial Problem, and the Translation Problem.

The Viewing Problem

Digital information created in the past requires the maintenance of an infrastructure and knowledge base in

order to view it. For example, to view an older word processing file, one needs software that understands the encoding schemes of the original software and can display that properly on the screen; without this, all one will be able to see is gibberish. But to keep these files alive over time, we will need to also keep software needed to run it, or keep knowledge of the encoding scheme and be able to produce software that uses the encoding scheme to properly display our digital files on the screen.

In the analog world, previous formats persisted over time. Cuneiform tablets, papyrus, and books all exist until someone or something (fires, earthquakes, etc.) takes action to destroy them. But the default for digital information is not to survive unless someone takes conscious action to make them persist. Oftentimes in the past, we have found old manuscripts or books squirreled away in basements or attics. But the word processing files of today found in the attics or basements of the future won't be readable unless their authors take some concrete action to make them persist. Even if we can read the floppy disks that we find and discover that there are files on them, we won't likely be able to decipher those files and display them properly.

When we discover older analog works, at least we can view them and their structure even if we had lost the ability to decode their language. And the subsequent discovery of works like the rosetta stone allows us to decode their structure and meaning. And when we discover old film (either still or moving images), even if we don't have the right projector for that format, we can still hold it up to the light and see what's on it.

But digital information requires an elaborate set of knowledge and/or computing environment in order to decipher it. It is usually encoded and to view it requires applications software which runs on a particular operating system, and which needs a particular hardware platform. It is usually stored on a physical device (like a hard disk drive, floppy disk, or CD-ROM), which requires a particular type of driver connected to a particular type of computer.

And each piece of that infrastructure is changing at an incredibly rapid rate -- in a way that allows the computer industry to repeatedly sell the same type of product to the same person (because the individual "needs" a faster or newer version). The rapid changes in hardware and software versions creates a headache for those interested in digital longevity. This includes problems with file formats, storage devices, operating systems, and hardware.

Most of today's word processors cannot read files created with older word processors. Most organizations have trouble even opening files created with the most popular word processor of only a dozen years ago (Wordstar). In fact, today's popular word processors (such as Microsoft Word) cannot even read files created with earlier versions of the same word processor (and often can only read files created in the most recent 2 versions). How can we ever hope that the files we create today will be readable in our information environments 100 years from now?

When today's word processors are able to open files from the more recent versions, often these files lose their formatting, and boldface, underlining, centering, and indentation change or disappear. But at least most of our older word processing files are primarily ascii text interspersed with formatting commands. So attempts to resurrect such a file at least have some hope of finding words and phrases contained within it. But for file formats not based upon ascii text (such as multimedia file formats), there is little hope that archeologists a century from now will be able to decipher anything at all within these files. Formats such as TIFF, AVI, the various versions of MPEG, etc. will pose even more longevity problems than word processing files.

Changing storage devices also pose problems for the future. In less than 20 years we have gone through removable storage devices including: 8" floppies, 5.25" floppies, 3.5" floppies, CD ROMs, and DVDs (and, with increases in storage density, there is little hope that the movement to new storage devices will subside anytime soon). Today, when we discover an 8" floppy, we have to first find an appropriate 8" disk drive, attach that to a computer and operating system that has an appropriate driver and can read it, and after doing all of this, we still have the problems outlined above in deciphering the file format. With our changes in operating systems (CP/M, MS DOS, Windows, Windows 95, Window NT, Windows 2000, ...) and hardware platforms (8088, 8086, 286, 386, 486, Pentium, Pentium II, Pentium III,...), we're creating a literal Tower of Babel in the proliferation of combinations needed to view a file.

Though digital longevity would seem to require it, how can we ever hope to deal with all these permutations and combinations? Think of all the formats we'd have to save, or all the emulations we'd need to allow us to decipher just the currently existing files.

The Scrambling Problem

In order to solve short-term problems resulting from the use of digital technology, we've engaged in practices that may result in long-term peril. Two noteworthy examples of this are how we have dealt with storage constraints and with digital commerce.

In the past, because large-scale storage was costly and bandwidth was fairly narrow, many repositories responded to these constraints by compressing their master images or multimedia. According to the reasoning that dominated until recently, compressed master files take up less storage, are easier to deliver to users with slow network connections, and are more convenient to handle internally. In recent years a number of institutions have come to question this tenet, as storage costs have plummeted and network speeds have dramatically increased. Yet the notion that one should compress even the master files still persists in many institutions.

There are a number of problems created by compression. First of all, we don't yet really understand the long-term effects of compression. Common lossy image compression formats such as JPEG essentially try to throw away information that is not too distinguishable to the human eye (colors that are close to one another get combined; spectral ranges beyond human perception are eliminated). But we don't yet understand whether some of this eliminated data will prove useful to future applications that will employ machine (rather than human) vision -- applications that may perform functions such as color analysis, comparing and overlaying images, etc. Use of lossy compression today may preclude certain uses of these images in the future.

Another very important issue is that both lossy and lossless compression add still another level of complexity to the encoding of a file, making it even more difficult for future archeologists trying to decipher its contents.

In a similar way, a number of efforts to enhance digital commerce may pose threats to longevity. Encryption schemes to inhibit unauthorized use add a level of complexity to a file's encoding, again increasing the problem for future archeologists trying to decipher a file's contents. And it's difficult to believe that all the pieces of complex digital commerce schemes like container architecture (which rely both on encryption and on the continued existence of an authority that can approve a payment transaction and release the appropriate key to decrypt the file) will survive long enough to ensure access to a digital file for more than a decade.

Most of these scrambling schemes are proprietary, and most don't adhere to widely-accepted standards. The level of complexity that scrambling adds makes it difficult to believe that anyone will be able to decode today's scrambled files even 50 years from now.

The Inter-relation Problem

In the digital world, information is increasingly inter-related to other information. The WorldWide Web is a primary example of how any given work may incorporate or point to a number of other works. And frequently a given work may actually consist of more than one distinct files that may or may not be displayed as if they are a single file (such as when a user views what looks like a single display, but is actually composed of a digital image residing in one file, and its title and other descriptive metadata residing within a separate file).

Today web designers are encouraged to engage in "good practice," taking advantage of the hypertext aspects of the WorldWide Web by breaking up documents into small pieces each stored in a separate file. These pieces can then be reassembled at viewing time so that they resemble the original full document, or the various pieces can be re-contextualized in different forms for different purposes. This means that even "simple" works may consist of several files, and that any given file may be part of more than one work.

On today's Web it is difficult to strive to make our own works persist when they point to and integrate with works owned by others. Because our current scheme for referencing Web files (the URL) is based upon a file's location, any time the location of a file we reference changes, our links break and users experience the most common error message on the Web ("404 Not Found"). Usually this problem is caused by some simple reorganization at the pointed-to Website (the renaming of a file, or of a folder/directory somewhere above it in the storage hierarchy, or the renaming of a server, etc.). But this common act of file/site management wreaks havoc on any works that point to or incorporate files from that site.

Another critical subset of the Inter-relation problem is the issue of determining the boundary of a set of information (or even of a digital object). Today the boundaries of a digital work are no longer confined to a single file. Frequently a Webpage will incorporate images, graphics, and buttons that are stored in separate files (and sometimes even on separate servers managed by separate organizations). Even traditional works like a journal article, report, or essay are frequently broken up into several separate files which are either assembled together at viewing time by a user's browser, or (for the stylistic purpose of not presenting the user with displays exceeding 2 screenfuls in length) remain separate linked files that a user must click between.

If we want to take action to preserve one of these complex works, we need to develop guidelines on where the boundaries of the work lie. If a work incorporates pieces owned or managed by another organization (icons, logos, images, text, etc.), does saving a copy of those pieces raise intellectual property questions? If we want to be able to show future researchers what kind of information was organized and distributed by an organization today, should we try to save that organizations homepage and every page that the homepage links to? What about the pages linked to by those other pages? Where are the boundaries? This is not unlike the problem faced today by lecturers who want to demonstrate their website in a lecture hall not equipped with an Internet connection; they must decide how many layers of inter-related files to download onto a demonstration machine.

Sidebar: Definitions of Digital Longevity Terms

The key technical approaches for keeping digital information alive over time were first outlined in a 1996 report to the Commission on Preservation and Access (Task Force 1996).

- *Refreshing* involves periodically moving a file from one physical storage medium to another to avoid the physical decay or the obsolescence of that medium. Because physical storage devices (even CD ROMs) decay, and because technological changes make older storage devices (such as 8 inch floppy drives) inaccessible to new computers, some ongoing form of refreshing is likely to be necessary for many years to come.
- *Migration* is an approach that involves periodically moving files from one file encoding format to another that is useable in a more modern computing environment. (An example would be moving a Wordstar file to WordPerfect, then to Word 3.0, then to Word 5.0, then to Word 97.) Migration seeks to limit the problem of files encoded in a wide variety of file formats that have existed over time by gradually bringing all former formats into a limited number of contemporary formats.
- *Emulation* seeks to solve a similar problem that migration addresses, but its approach is to focus on the applications software rather than on the files containing information. Emulation backers want to build software that mimics every type of application that has ever been written for every type of file format, and make them run on whatever the current computing environment is. (So, with the proper emulators, applications like Wordstar and Word 3.0 could effectively run on today's machines.)

Both a migration and an emulation approach require refreshing.

The Custodial Problem

Though a number of traditions have developed concerning which organizations should take responsibility for preserving and maintaining various types of analog material (correspondence, manuscripts, printed matter), no such traditions exist yet for digital material. As a result of this, much current material originating in digital form falls through the cracks, and is unlikely to be accessible to future generations.

For example, special collections librarians who aggressively pursue print-based collection development in their particular specialty areas claim that it should be the responsibility of their organizations' computing staff to pursue collection development of material originating in digital form. Yet those computing staff claim that it should be the subject-matter specialists' responsibility to pursue collection development of digital materials. Meanwhile, much of this fragile material is not collected at all.

Another example is correspondence, which in an analog world left a paper trail. Most organizations follow guidelines for saving significant amounts of paper-based correspondence. But few organizations have developed similar guidelines for saving electronic correspondence, and few individuals have even the slightest idea of how they might save their own personal correspondence even if they were eager to do so. This problem is becoming even more acute as more and more important correspondence originates in digital form.

One final example is from the domain of literary creation. In the analog world, authors used to leave important traces of their creative process in the form of numerous drafts, marked-up manuscripts, and

correspondence. Today they use word-processors and email for both drafts and correspondence. And frequently they only save a very few of their drafts and none of their correspondence.

A major question we face in the coming years is: Who should be responsible for saving material in electronic form? Should individuals carry this responsibility themselves? Or should social entities (such as businesses, libraries, archives, and professional societies) aggressively intervene to save material? And how will they decide what to save?

Another critical question is: How should they go about saving it? Our field still needs to develop guidelines and best practices so that organizations and individuals who want to make the effort to try to make digital information persist will know how to do so.

A key function of archives is in ensuring the authenticity of a work. They do this by amassing "evidence" and by maintaining a "chain of custody". But when works are subject to repeated acts of "refreshing" as most approaches to digital longevity propose (see sidebar), these traditional ways of ensuring authenticity break down. Files repeatedly copied to new strata face the likelihood that changes will be introduced into these files, and we know little about how to control mutability across repeated "refreshments".

The Translation Problem

When content is translated into new delivery devices (such as digital forms), the change of form often serves to change part of the meaning. Conversions from analog to analog face this problem, as do conversions from analog to digital (a photograph of a painting is not the same as that painting, and a digital representation of an object is not the same as that object). (Besser 1987)

Because we can make identical copies of digital files, this has led some people to the mistaken belief that digital-to-digital conversion will not face the same translation problems that analog-to-digital conversions face. This is not true because, though the bits in the file's contents may be identical, the applications environment used to view the file most certainly will be different. In fact, the very reason for converting the file is because we are unable to successfully sustain that application's environment over time.

Many people have experienced this as their word processor "successfully" imports a document created with an earlier version of the same word processor, while losing formatting (such as centering, underlining, and font changes) or punctuation (losing apostrophes or double-quotes). This can also be true in emulation environments because the creators of these environments choose which aspects of the environment to emulate, and they cannot possibly emulate every single aspect. (For example, a recent emulation of one of the earliest computer games [Moon Dust] was shown to its original designer [Jaron Lanier] who contended that it was a completely different game than the one he designed because the pacing was different.)

In saving a work, it is critical that we save parts of that work's environment which might not be immediately obvious to save. For example, anyone is likely to recognize that we must save the image of every page in a digitized book. But for that book to be useable, we must also save important behaviors of that book, such as the metadata and accompanying behaviors which will allow future users to turn pages, skip from the table of contents to a particular chapter, or go back and forth between the main body of text and citations or footnotes (Making of American II 1998). Saving just the page images of a book without its behaviors would be like saving a video game with the interactions ñ some kind of representation, but missing one of the most critically important functions.

With works that start out in digital form, we need to better understand the aspects of that work's original environment that are critical to viewing that work, and we need to figure out ways to sustain all the important behaviors of a work as we move the work's contents through generations of migration or emulation (Besser and Gilliland-Swetland 1999). We also need to understand how each new viewing environment affects the nature of a work. (For example, many filmmakers would contend that their film is radically changed when shown on a video screen. How will today's multimedia creators feel about their works being shown in future environments where cathode ray tubes are no longer available for display?)

Paths to Improving Digital Longevity

Given these formidable problems, how can we hope to assure the longevity of today's works that we want to preserve?

Broad Approaches

First of all, we need to recognize that a great deal of knowledge exists over how to preserve bits over time. For more than a quarter of a century the data processing community has had experience in moving large centralized bodies of bits from one physical storage medium to another. Our community needs to study the experience of corporate and university data processing departments to learn about their experiences and to obtain cost figures, then we need to examine how these might be applied to the less highly centralized bodies of digital information that our community has.

While studying this experience, we also need to keep in mind that preserving bits is only a small part of the problem. This problem is dwarfed by the much larger problems of assuring that file formats will be accessible, and of problems involving organization, policy, and roles and responsibilities.

In the thousands of years since the Library at Alexandria was destroyed, redundancy has been a key to the preservation of information. The existence of multiple copies of a work geographically dispersed among a number of sites has helped preserve works from both natural and human-created disasters (ranging from fires and earthquakes to accidental obliteration of a set of works). Any long-term preservation strategy for digital information must incorporate cooperative relationships among physically dispersed locations and organizations. We need to develop international cooperative projects where organizations are willing to store and refresh redundant copies of works that are really under the custodialship of other organizations.

Current intellectual property laws inhibit an archive or library from preserving information in digital form, particularly since much of the digital information they acquire is licensed rather than owned. A recent study on Copyright by the National Academy of Science (National Research Council 2000) strongly recommended that intellectual property laws be changed to permit these institutions to legally preserve information in digital form, and that significant funding be allocated to digital preservation. We need to continue to monitor changes in intellectual property law (Besser Copyright website) and press for the changes that will allow us to engage in digital preservation without facing criminal penalties.

We need more experience in the two competing strategies for digital preservation -- *emulation* and *migration* (see Sidebar). The emulation approach is highly experimental, and we need to monitor the two experimental international studies that have recently begun to explore this area: NEDLIB sponsored by the European Community (Networked European Deposit Library website), and the CEDARS Project (CURL Exemplars in Digital Archives website) sponsored by Britain's Joint Information Systems Committee and the US National Science Foundation.

What we as a community can do

While no one has yet solved the broad set of problems around digital longevity, there are a number of particular actions we can take that will improve the likelihood that a work we seek to save will remain accessible over a prolonged period of time. There are also a series of actions that our community as a whole must begin to grapple with in order to reduce this immense problem.

Our community needs to insist upon clearly readable standardized ways for a digital object to self-identify its format and the applications needed to view it. With a standard for embedding the name of the viewing application in a particular place within an image header, 22nd century archeologists discovering today's files will at least be able to discover what applications they need to look for in order to view this file. Work on this and a number of related problems for longevity of digital images was begun as part of a Spring 1999 invitational meeting sponsored by the Commission on Preservation & Access, the National Information Standards Organization, and the Research Libraries Group (Besser 1999).

Our community needs to better understand how information relates to other information (Besser and Gilliland-Swetland 1999). In particular, we need further clarity about what constitutes the "boundaries" of information objects. When we are trying to save something (particularly a hypertext or hypermedia object), we need to know what pieces we really need to save.

Finally, our community needs to develop a concrete set of guidelines that can be used by people and organizations wishing to make information persist. In a sense, this chapter is the first attempt at struggling with what might be in such guidelines.

In deciding to digitally preserve a group of works, the institution must first understand the special needs of the types of works contained in that collection. This means understanding how reformatting these into another format may affect the understandability and the usability of those works. This means understanding the boundaries of this work and which pieces must be saved (perhaps even including contextual pieces). And (as we saw with the example of a digitized book) this also means saving the "behaviors" of a work, not simply its "contents".

The role of Metadata

At this point in time, extensive metadata is our best way of minimizing the risks of a digital object becoming inaccessible. Properly used, metadata can:

- identify the name of the work, who created it, who reformatted it, and other descriptive information
- provide unique identification and links to organizations, files, or databases which have more extensive descriptive metadata about this work (this is particularly important in the likely event that the digital file and its external metadata become separated)
- explain the technical environment needed to view the work, including applications and versions numbers needed, decompression schemes, other files that need to be linked to it, etc.

There are various types of metadata which appear unimportant today, but may prove critical for properly viewing these files in the future. (For example, saved information about a particular scanner's color profile will be critical for future color management systems to account for display device differences and to properly display colors on a particular device.) A good rule of thumb is to save any metadata that is cheap/easy to

capture, or that someone has indicated might eventually be important.

Further Information

Those involved in planning for digital longevity should read the key texts that have scoped out the problems for our fields: the Commission on Preservation and Access report (Task Force 1996), the Getty's Time & Bits conference on digital preservation (MacLean & Davis 1998), and other items referenced on the Sunsite Longevity Page (Besser, Digital Longevity website). They should also continuously monitor the Sunsite Longevity Page (Besser, Digital Longevity website), publications of the Commission on Preservation & Access (Commission on Preservation & Access website), and the work of the Internet Archive (Internet Archive website).

References

- Besser, Howard. (1987). *The Changing Museum*, in Ching-chih Chen (ed), **Information: The Transformation of Society** (Proceedings of the 50th Annual Meeting of the American Society for Information Science), Medford, NJ: Learned Information, Inc, pages 14-19.
- Besser, Howard. *Copyright* (website), <http://www.gseis.ucla.edu/~howard/Copyright/>
- Besser, Howard. *Digital Longevity* (website), <http://sunsite.berkeley.edu/Longevity/>
- Besser, Howard (1999). Image Metadata: meeting summary, <http://www>
- Besser, Howard and Anne Gilliland-Swetland (1999). *Multimedia: Issues in using Visual Material in Cultural Heritage Organizations* (Spring 1999 class and website) <http://www.sims.berkeley.edu/impact/s99/>
- Commission on Preservation & Access (website) <http://www.clir.org/programs/cpa/cpa.html>
- CURL Exemplars in Digital Archives (website) <http://www.leeds.ac.uk/cedars/>
- Internet Archive (website). <http://www.archive.org>
- Lyman, Peter and Howard Besser (1998). *Defining the Problem of our Vanishing Memory: Background, Current Status, Models for Resolution* in Magaret MacLean and Ben H. Davis (eds.) **Time & Bits: Managing Digital Continuity**, Los Angeles: J. Paul Getty Trust, pages 11-20.
- MacLean, Magaret and Ben H. Davis (eds.) (1998). **Time & Bits: Managing Digital Continuity**, Los Angeles: J. Paul Getty Trust.
- Making of America II White Paper (1998). <http://sunsite.berkeley.edu/moa2/>
- National Research Council (2000). **The Digital Dilemma**,
- Networked European Deposit Library (website). <http://www.konbib.nl/nedlib/>
- Sanders, Terry (1997). **Into the Future: Preservation of Information in the Electronic Age**, Santa Monica: American Film Foundation (16 mm film, 60 minutes)
- Task Force on Archiving of Digital Information (1996). *Preserving Digital Information*, Commission on Preservation and Access and Research Libraries Group (<http://www.rlg.org/ArchTF/tfadi.index.htm>)

Last modified: 4/20/2001