Ending Digital Obsolescence

Ken Quick and Mike Maxwell
ACS
Dallas, Texas, USA

Abstract

Digital Obsolescence is the eventuality that computer hardware, media, and file formats become obsolete before the useful life of the data that is stored on and in them ends. The digital obsolescence problem is comprised of issues from each of the above elements: Computer Hardware, Storage Media and File Formats.

The problem is international. Special groups from many countries have formed (mostly in the government and library/museum communities), to pursue a solution and to identify alternatives. As yet, no practical solution has evolved that could be commercialized for widespread use.

Digital obsolescence cannot be avoided through a no-migration policy as such a policy only delays the inevitable realization that the data has been lost.

Digital obsolescence cannot be avoided as long as new hardware is developed that obsoletes older hardware.

Therefore, the answer to avoiding format obsolescence is to archive the formatted data, the creating application, and the operating system it requires to execute. This poster presentation will discuss the use of 2-D Barcode and microfilm (stable and standardized technologies) to create a storage method that can sustain any digital file for hundreds of years. Any digital file, audio, video or document can be encoded, preserved, decoded and restored to its original operational state in the future.

The poster session will be an interactive demonstration with multi media examples. This process has been reviewed by many organizations, including NARA, EMC and Kodak, and found to have merit.

Obsolescence Problem

The digital obsolescence problem affects 3 primary technologies: Computer Hardware, Storage Hardware/Media, and File Formats.

Computer Hardware becomes obsolete when new generations of processors replace current ones. The increased capabilities of the new computer will often require the obsolescence of older fundamental foundations, much like 32-bit processors obsolete 16-bit processors and the computers, accessories and software that was designed to run on them.

Storage Hardware and Media become obsolete as their successors become faster and denser. There are many types of storage. Some are wholly contained like common Hard Disk Drives. Some are removable media like tape drives and optical drives. Most are available as internal or external devices. All storage media and their supporting hardware are destined for obsolescence as newer devices build upon the advantages they have brought to the market.

File Format obsolescence applies to both industry standard formats and vendor specific formats. In a competitive environment, a standardized and non-proprietary format will always lack the newer, proprietary features of a vendor specific format. Industry standard formats are perpetually obsolete and today's vendor specific formats will become obsolete as they are replaced by newer formats with more competitive features.

The digital obsolescence problem is international. Special groups from many countries have formed (mostly in the government and library/museum communities), to pursue a solution and to identify alternatives. Considerable sums of money have been expended in the last decade for research and theorizing solutions. As yet, no practical solution has evolved that could be commercialized for widespread use.

Computer Obsolescence

Computers evolve. We expect computers to get faster and more capable with every model. And the computer vendors have not disappointed us. They have not simply added speed. Features have been added along the way that inadvertently obsolete some fundamental characteristics of our older computer hardware. Features like the ability to address more than 1 megabyte of memory may prevent a program from running if it was written before computers had this ability.

Current generation computers are typically 32-bit or 64-bit data word lengths. Often they have dropped their ability to execute programs written for 8-bit or 16-bit computers.

Some computer manufactures have ceased production. Those computers become instantly obsolete. The long list of obsolete computers includes the TI-99, Commodore 64 and Amiga, and the Sanyo MBC-550 to name but a few.

Computers will become obsolete. In fact, all computers available for purchase today will be obsolete in 5 years.

Storage Hardware and Media Obsolescence

Every new generation of computer hardware and storage media advances the data density and access speed of its predecessor. Moore's Law can be paraphrased in storage terms as …*computer storage devices will double in capacity and half their access times every 18 months…*

Computer hardware and storage media obsolescence is a direct consequence of the relentless advances of new and more capable storage devices.

End users are left to deal with the incompatibilities between the old and new devices. Frequently the new hardware is not backward compatible with the older hardware's media for more than one generation.

Storage medias lose bits over time. Even if data migration is flawless across intermediate media with repeated reading and writing using differing technologies (very unlikely), the time spent idle between migrations will take its toll on the stored data bits. Data migration success rates are never 100% and successive storage/migration cycles accumulate failures and expose the data to corruption and loss. And there isn't any way to repair the damage.

Because storage and migration is not 100% reliable, trying to avoid digital obsolescence through a strategy of store and migrate is a plan to lose data. Storage Hardware and Media obsolescence cannot be avoided as long as new hardware is developed that obsoletes older hardware.

Format Obsolescence

Formats eventually die. Wildly popular in their prime, formats quietly slip into disuse and die as they age and are supplanted by new, more capable formats. Some formats die suddenly when their parent companies cease business. The majority of formats are proprietary and the only action possible to maintain the data object is to migrate it to a new format, typically from the same single-source vendor.

Format migration is even more perilous than data migration. Competitive applications necessarily differ in features to achieve differentiation. A vendor will revise the use of some features from version to version to show improvement. Format migration necessarily involves the machine interpretation of the human use of features, so migration utilities are notorious for incomplete or inaccurate format migrations. Some characteristics of the migrated format are altered to use similar features in the new format or the feature is simply ignored. The alterations may be slight or so severe that the migrated file is unusable.

Format obsolescence can only be avoided if the data is only used by the application that created it.

No Migration Strategies

Avoiding digital obsolescence through a no-migration strategy is unwise. A no-migration strategy freezes in time a set of hardware and software as an independent unit – never upgrading anything – with plans to continue to use the hardware, media, software and its formats forever.

But hardware fails. It is not a matter of *if* it will fail, just *when*. Power surges, dirt/dust accumulation, heat, cold, foreign debris, and mechanical fatigue all take their toll and one will eventually claim the life of any hardware device. Suddenly the data becomes trapped on severely obsolete and irreplaceable hardware.

Digital obsolescence cannot be avoided through a no-migration policy as such a policy only delays the inevitable realization that the data has been lost.

Ending Computer Obsolescence

Today's software engineers require the object computer hardware to develop the software that is to run on the object computer. Frequently, the object computer hardware is not available at the time the software engineers begin programming for it. Virtual hardware, or emulation, is common-ly used to substitute for the unavailable object computer hardware to allow both engineering tasks to be developed simultaneously.

Hardware emulation is a common alternative to using the actual hardware. Emulation is a special type of software application that creates a virtual object computer on a currently available host computer. The virtual hardware can be used to run and test software before the actual hardware is available. Although emulation is most commonly used to emulate future hardware designs, it is also used – and arguably more easily – to create virtual historic or obsolete computers.

Historic computers like the Commodore Amiga and C64 have an avid following (see www.amigaforever.com and www.c64.com). Often their treasured software is executed on modern computers using emulators that they have written themselves. Emulation has achieved the immortality that Commodore could not.

Intel-based computers (IBM-PC architecture) are significantly different from Motorola-based computers (Apple Macintosh) at the binary level. Emulators for either hardware platform are available for the other hardware platform and they allow software written for both hardware platforms to be executed on the same computer.

Emulation ends computer obsolescence. Just as we emulate the venerable DEC PDP circa-1970 computers on circa-2005 computers, future generations will emulate our circa-2005 computers on computers from their era.

Ending Storage Hardware and Media Obsolescence

Write-able digital media is necessarily entropically high energy and unstable. Nature tries to equalize energy levels, and over time, it succeeds at erasing the data that is stored on most digital medias. To store data for a long time, the media the data is stored on must be nearly entropically equalized to start with so that further deterioration will take centuries to happen.

To avoid storage hardware dependence, the storage media must not be tied to any one manufacturer and the retrieval device must be able to be crafted from simple parts. Future generations must be able to construct a device to read the media with minimal understanding of what is on the media and with simple, readily available parts.

No magnetic media is even close to being entropically equalized. Few optical medias approach the entropically equalized requirement. Neither magnetic nor optical medias meet the simple device requirement. But not all storage medias are magnetic or optical.

Some non-traditional medias that do meet the entropically equalized and simple retrieval requirements are clay or stone tablets, laser engraved non-corroding metals, low lignin paper, and microfilm. Only microfilm and acid-free (low lignin) paper are in common use today.

Microfilm and acid-free paper are storage medias that do not become obsolete and therefore avoid storage hardware and media obsolescence. Microfilm is particularly attractive as a hardware independent media that provides compact storage, organization, and proven archival properties. Microfilm, adapted to hold binary data, ends digital hardware and media obsolescence.

Ending Format Obsolescence

Format obsolescence is difficult to overcome. Migrating to a new format is very challenging, expensive, and requires full visual and operational verification. To avoid format obsolescence the creating application must be kept from becoming obsolete or it must be made available to interact with the obsolete format.

Format obsolescence is caused by its native application becoming obsolete. Therefore, if a format is to be retrievable in the future exactly as it is in the present, the creating application must also be saved for the future. And if the application is to be executable in the future, the operating environment (operating system) upon which it is dependent must be saved as well.

Therefore, it is necessary to retain the formatted data, the creating application, and the operating system to avoid format obsolescence. If multiple files were created by a single application, that application need only be saved once. By the same logic, the operating system need only be saved once for all of the applications that depend on it.

Ending Digital Obsolescence

Computer hardware obsolescence is prevented through emulation. Selecting a suitable media prevents Storage Hardware and Media obsolescence. And saving the application and operating system along with the data prevents Format obsolescence.

The Binary Foundation

All digital data, regardless of what it is to a higher level of intelligence, is just a series of 1s and 0s - binary data - to the computer.

Black and white and color images are binary data, as are animations, videos, and audio files. Spreadsheets, CAD/CAM drawings, and the rover's programs that are collecting data on the surface of Mars are all binary files. All computer files are binary files. What a file appears to be to other programs is a matter of interpretation by the other programs, but at their fundamental structure level, all computer files are simply binary files.

The missing link is the ability to store and retrieve binary data on a suitable long-life, obsolescence resistant media. Digital medias simply do not last very long and contribute to the problem, not the solution.

What is needed is the ability to store binary data on microfilm. Microfilm carries with it all of the properties necessary to end digital obsolescence for storage hardware and media and formats. Microfilm is hardware independent, lasts for 500+ years, and adapted to hold binary data, can carry the formatted data, its application, and the application's operating system into the future.

Once stored, the data, application, and operating system form a complete binary record of the object data, and the ability to return it to active use in the future. The only requirement left for future generations is to write emulators that create a circa-2005 virtual computer.

Putting Binary Files on Microfilm

Microfilm lasts for 500+ years, eliminating media to media migrations for centuries.

Microfilm retrieval devices are simple. Little more than a light and lens are required to view microfilm and capturing an electronic image can be accomplished using several different technologies.

Microfilm stores images. Writing human readable data (Computer Output to Microfilm) or TIFF images to film has existed for decades. Converting binary data into images allows microfilm to hold binary data. The binary data images must also be machine readable, so computers can read them and assist in retrieving the binary data back from the binary microfilm images.

Barcode technology, especially 2-dimensional (2-D) barcode technology, can be used to encode binary data. Although individual barcodes can only store small amounts of binary data, linking a series of these barcodes together can store binary files of almost limitless capacity.

The ability to distinguish each encoded binary file from all other encoded binary files with its particular identifying characteristics is important and that meta-data information must also be stored with the binary file. This information is also binary data, so it is also encoded into the 2-D barcodes.

After gathering all of the meta-data about a file that is to be archived on microfilm, segment all of the meta-data and binary data into blocks of data small enough to be encoded into the 2-D barcode symbols. As the symbols are created, assemble them into images that are compatible with a microfilm writer device. Once the entire file has been encoded into 2-D symbols, the images can be written to microfilm. Process and handle the microfilm to LE-500 standards and the file and its meta-data will be secure for 500+ years.

Retrieving Archived Binary Files

In order to archive binary files, a method to retrieve them must be in place. Any available microfilm scanner that can capture a clear image of the 2-D barcode symbol on the microfilm can return the image to a digital image in a computer. If no microfilm scanner is readily available, the construction of one is not a major feat. Finally, as a fail-safe measure, a human can key the barcodes into a computer array structure and the computer can decode them (even if the human makes some mistakes!). Tedious, but a functional fail-safe capture method.

Computer programs can read the 2-dimensional barcodes from the images. The output would be a block of binary data from each barcode.

The sequence number that was embedded with each encoded block facilitates the reassembly of the binary data blocks back into a binary file.

The original file's name, placement, date, attributes, etc., are applied to the reconstructed (cloned) binary file, which then becomes indistinguishable from the original file that was archived to microfilm.

Retrieving a file stored on microfilm is a straightforward process. Simply scan the microfilm images that contain the file segments from the desired file. Decode each 2-D barcode and save the contents using the segments' sequence number for identification. Once all segments have been decoded, reassemble them in the sequence order. Apply the file's meta-data that is retrieved from the meta-data symbols and the file has been fully restored.

Microfilm Artifacts

Microfilm contains artifacts. Artifacts can best be described as distortions and defects in the clarity of the microfilm that create random black specs in the image retrieved from microfilm. Depending on the size of the elements of the 2-D barcode that is written to microfilm, some of the artifacts will approach or exceed the element size. When this happens, a formerly clear element may appear black when read back from microfilm.

The selection of the 2-D barcode used to encode the binary data is critical. The 2-D barcode must have some internal form of error detection and correction capability.

Data matrix (ECC-200) barcodes utilize Reed-Solomon error correction code. Some larger data matrix symbols hold over 1000 bytes of binary data and can correct microfilm artifact errors that affect up to 40% of the symbol.

The data matrix symbol was invented by RVSI and placed in the public domain for all to use. It is an approved ISO standard (ISO 16022). Although the data matrix symbol is not the only 2-D barcode that can be utilized in this approach to ending digital obsolescence, it is particularly well suited to use in the microfilm environment.

Rosetta Stone Images

A black basalt slab with strange inscriptions on it, the Rosetta Stone was unearthed in July 1799 by Napoleon's army in Rosetta (Rashid), Egypt, and kept as a souvenir by one of the troops through generations of the soldier's family. Eventually it ended up in a flea market and was spotted by an Oxford professor of Egyptology on vacation in France who recognized hieroglyphics on a portion of the stone. It was quickly discovered that the Rosetta Stone contained the same text in 3 languages and one was Greek. It was the key to decoding the there-to-fore undecipherable hieroglyphics, the archived writings of the ancient Egyptians.

Likewise, any archived encoded data needs an equivalent of the Rosetta Stone.

A series of images that are not encoded, but are human-readable standard text images with illustrations, tables, formulae, and example computer source code should be stored on microfilm along with the coded binary data. These images will convey to future generations exactly how to decode the barcode symbols and reconstruct the data blocks back into binary files.

Collections

Rarely will a single file be archived (unless it is a self-extracting archive file of a collection itself). A collection is a group of files that are interdependent. Modern examples are all of the files in the 9/11 hearings or all of the files associated with a new drug's development and testing.

Through collections, future generations of digital archaeologists will discover all of the relevant pieces of their quest in a single place.

Index Database

Whenever multiple files are archived using this technique, it is an aid to future retrieval efforts to include an index of where each file is located.

Along with the Rosetta Stone images, in both human-readable text and barcode encoded forms, an optional index database can be stored. The index contains the locations of where each file of a collection can be found on the microfilm. The index is a simple importable format flat-file text database (comma delimited format) that can be imported into any current or future search engine. Retrieval of the database would not be necessary if all files in the collection are to be retrieved, however it would be a great help if only a single file is to be retrieved.

Processor Emulation

Emulation of circa 2000 processing hardware is all that is required for future generations to use the software that is stored on the film - including the operating system, the application and the file objects. Emulators are the only piece that the future generations will have to create. Emulation of hardware peripherals is not required for restoration of files from Datasurance.

Emulators create virtual machines that can be used in the future to allow older programs to run on newer hardware with minimal development effort. Hardware processor emulation does not require a rewrite of the operating system. It runs the original operating system (that is stored on the film) on a virtual hardware computer. Hardware emulation is not very complicated. It is a software program that reads the instructions that were originally written for a circa 2000 computer and translates them into instructions suitable for the hardware of a circa 2500 computer.

As mentioned earlier, emulation exists in many forms - PC In Mac and Mac in PC emulators; emulators exist for circa 1990 *Commodore* and *Amiga* computers to run their games on today's PCs. JAVA requires an emulator of a standard machine to run. Both commercial and consumer demands will cause continued development of emulators.

A significant benefit demonstrated is that the emulated program will run much faster on the future computer than it did on the original computer. As history shows, the speed improvements of processors will continue and it is certain that future computers will run emulations of circa 2000 computers faster than the original circa 2000 computers ran. A nice, free benefit! If speed is a problem in the emulation, either the future computer can be clocked slower or sleep/wait instructions can be added to the emulator to slow the execution to circa 2000 computer speeds.

Pros and Cons

Pros:

- uses the most stable proven media available today
- uses simplistic, non proprietary ISO format
- applies to any digital file
- high trustworthiness of file integrity
- no multiple migrations or refresh required
- can include human readable files
- relies only on emulation as a future computer science
- does not rely on un-invented technology to prevail

Cons:

- packing density is lower vs. other digital options
- accessibility is slower then other options
- higher front end costs vs. other options

Conclusion

Much has been said and written about the impending probability that ours will become the lost generation because we saved everything in unintelligible formats on irretrievable and unstable media.

Today, digital obsolescence has been overcome. By adhering to the guidelines presented in this paper, your archived digital data will survive for 500 or more years.

Biographies

Michael C. Maxwell is a member of AIIM and ARMA, is a frequent speaker at ARMA conferences. Mike has an MBA degree from the University of Rochester Simon School of Business Administration. He has over 35 years of imaging and storage industry experience with Eastman Kodak and Document Strategies, Inc.

Ken Quick received a Bachelor of Computer Science degree from Georgia State University in 1991. Ken joined ACS in 1987 and has primarily developed production image pro-cessing systems. He designed the system that wrote 570 million images of the 2000 Census to microfilm.

Affiliated Computer Services (ACS) of Dallas, TX has applied for a patent on this binary data archive technology. Contact Mike Maxwell at 585-781-0376 or Jerry Karpa at 303-790-8190 for more information.