COMPUTERWORLD

Fending off the digital dark ages: The archival storage issue

We're churning out vast quantities of digital data that we aren't equipped to preserve.

Lamont Wood

August 26, 2010 (Computerworld)

Ready for a digital dark age?

Consider the BBC Domesday Project, undertaken in 1986 to mark the 900th anniversary of the original Domesday Book, a land-use survey of England commissioned by William the Conqueror in 1086. For the latter-day survey of the island, thousands of Britons contributed text, photos and video that were published on two custom laser disks.

"We could be facing a digital dark age 50 years from now, and future scholars will not be able to understand our culture," says Andy Maltz, director of the science and technology council of the Academy of Motion Picture Arts and Sciences.

But just 15 years later, it was impossible to access those disks without lots of custom hardware and extensive software emulation. Currently the Centre for Computing History in Haverhill, England, has a functional emulation and hopes to post the contents to the Web.

Meanwhile, the original Domesday Book, handwritten on sheepskin, remains in the British archives, usable after nine centuries by anyone literate in Latin.

Anyone with data stored on 5.25-inch floppies or text in WordStar format faces a problem similar to the one that befell the BBC Domesday Project. The digital data we are generating wholesale will very likely become unusable within our lifetimes unless we take steps to preserve it.

The situation cannot be blamed entirely on the computer industry's treadmill of planned obsolescence. In essence, digital storage technology has inherent drawbacks that make paper look immortal.

Data mortality

A hard drive removed from a computer and left on an office shelf will eventually become unusable just because of daily temperature changes, explains Tom Coughlin, a data storage consultant in San Jose. The thermal energy fed into the media will gradually trigger spontaneous reversals of the magnetic particles that store the information, until the original data is lost, he explains. However, the loss should not be a problem for the first 10 years, he adds. After that, it's anyone's guess as to when the data becomes unusable.

Magnetic tapes have the same problem, but take decades to lose data to thermal erasure because they have a lower bit density than hard drives, Coughlin says. On the other hand, tapes have a different problem: delamination, which is what happens when the magnetic media separates from the tape or is attacked by fungus. Tapes sometimes have to have their media re-affixed with a baking process to allow the one final read needed to move their contents to another medium, Coughlin says.

USB memory sticks are also subject to thermal erasure and face added risk because they typically have the cheapest available controllers. "I would not use them for archival purposes," he says. Continued use of USB memory would also require that USB ports still be in use decades from now, and it's anybody's guess what laptops will look like in 20 years, let alone 50.

As for DVDs and CDs, Bill LeFurgy, a project manager at the Library of Congress, reports that his organization has done accelerated aging tests on them using ovens and has found enormous variability among discs -- even among those that are the same brand. "Some will last a decade and others much less," he says. "Beyond five years, I would be nervous."

Other storage professionals complain that the throughput of DVDs is too slow for archival use. DVD throughput is typically less than a quarter of tape throughput, plus DVDs have to be changed every few gigabytes.

And as is the case with other storage formats, there remains the issue of whether any CD or DVD readers will be around decades from now.

Online survivability

What about online storage? These are hard drives that are turned on and ready for immediate access. Here, the data can be constantly checked for integrity and easily replicated. But it can also be corrupted quickly, and the long-term reliability necessary for archiving is not on the horizon, complains David S.H. Rosenthal, chief scientist for the "Lots of Copies Keep Stuff Safe" (LOCKSS) program, a Stanford University Libraries initiative.

Rosenthal has investigated what would be required for a petabyte stored online to have a 50% chance of being usable after a century. Analyzing the drive maintenance data published by various

Archival standards in the making

An often-cited example of a group doing archival standardization work is the Storage Networking Industry Association (SNIA) in San Francisco. Wayne Adams, SNIA's

storage farms, he found that to reach the petabyte-century goal, the reliability of online storage has to be improved by a factor of 10^9 (i.e., 1 billion).

But even if we could honestly achieve a billion-fold improvement in online storage reliability, there would be no realistic way to test such a system short of plugging it in and waiting 100 years, he points out.

With the odds of digital survival being so low, and with so much information originating in digital form, "we could be facing a digital dark age 50 years from now, and future scholars will not be able to understand our culture," says Andy Maltz, director of the science and technology council of the Academy of Motion Picture Arts and Sciences -- the group that awards the Oscars -- in Beverly Hills, Calif.

chairman and a senior technologist at storage vendor EMC Corp. in Hopkinton, Mass., says the association has developed the following three standards to address the issue:

- **XAM (Extensible Access Method):** Adams says that this standard separates the application from the data and "lets you manage the data in its own right and not worry about the forward migration of the application. Otherwise, to use the data 15 years from now you'd have to put a whole system in a time capsule." According to SNIA, XAM contains metadata definitions to help archived data achieve application interoperability and to make it more searchable. SNIA's Web site lists XAM-based products or services from 13 different organizations.
- **SIRF (Self-contained Information Retention Format):** This standard should make it possible for future users to query archived data without having to use the original application. SNIA literature calls it "a specification that defines a logical container format appropriate for the long-term storage of digital information."
- **CDMI (Cloud Data Management Interface):** This standard also defines metadata and other storage parameters and is therefore applicable to archiving, according to Adams.

Preservation standards

With awareness of the problem growing, various organizations have been working on approaches to the archiving problem, focusing primarily on ways to reduce the danger of format obsolescence. (See sidebar, above.)

Preventing obsolescence usually involves developing dictionaries of metadata -- information about a file that is stored with a file. That way, future users won't be stuck like the scientists in 1999 who were unable to make any sense of magnetic tapes containing NASA's Mars probe data from 1975. (After finding some printouts, the scientists were able to analyze about one-third of the data. To learn more, see "The lost NASA tapes: Restoring lunar images after 40 years in the vault.")

Beyond standards, there is also a more subtle management issue. "Most organizations could not tell you how long certain electronic content needs to be kept, and only 5% to 10% are tagging the content with metadata in sufficient detail" for employees to know how long to keep the data, says Donald Post, a SNIA spokesman and a partner at Imerge Consulting, a Chicago-based firm specializing in records management. "Meanwhile, 80% of what they are trying to keep are duplicates, but they are not taking the time to discard the duplicates. And 95% think that making a routine backup is [sufficient] protection."

Enterprise IT managers aren't pushing for commercial solutions to the problem, and therefore vendors aren't rushing to offer any, says Post, but he also expects the situation to change within the next three years as vendors realize the commercial potential for digital preservation products.

Keeping bits alive

Of course, there are some organizations that are successfully dealing with the challenge of digital archiving.

"Most countries have this problem of digital preservation," notes Dyung Le, director of systems engineering for the Electronic Records Archive initiative of the U.S. National Archives and Records Administration in College Park, Md. There, archived tapes are recopied every 10 years, and the National Archives tries to have at least three copies of everything, with at least one copy being off-site. The agency manages more than 400 terabytes of data, he estimates.

Since there's no telling what computer applications will be in use centuries from now, text-based material is typically converted to XML, which is based on ASCII. Various forms of metadata are preserved in the file, including descriptive data that could be used as a search aid. Le said that the XML files store the metadata using an extension of PREMIS (Preservation Metadata: Implementation Strategies), a digital preservation standard also based on XML and ASCII and created by the Online Computer Library Center.

The long-term reliability necessary for archiving is not on the horizon, complains David S.H. Rosenthal, chief scientist for the "Lots of Copies Keep Stuff Safe" program of the Stanford University Libraries.

There's no intermediate format like XML for non-text data, Le said. Therefore, the best an organization that wants to archive material can do is note what format the material is in and plan to eventually migrate it to whatever application format is dominant in the future -- but it must do that at a time when systems for converting from the original format are still available, Le says. In other words, organizations must take their best guess about what formats will be used in the future and convert to them while they still can.

An archivist must also be able to certify that material being saved is an authentic copy, he explains. That's done by creating a hash key for each file; the hash keys travel with the file. When copies are supplied, the archivist must also certify that no characteristics of the file have been changed that would change the meaning of the material. For that reason, text must sometimes be preserved in its original format, since the formatting is deemed essential to the meaning, Le adds.

Other federal government agencies, state archives and libraries, and sometimes even private individuals, are also facing the problem of digital preservation. For them the Library of Congress, at the direction of Congress, has set up the National Digital Information Infrastructure and Preservation Program (NDIIPP), says LeFurgy.

NDIIPP officials are working with about 170 stakeholders, including trade organizations and foreign governments, and they publish an inventory of tools and services at DigitalPreservation.gov.

The Library of Congress itself has archived about 167 terabytes of digital content, including Web sites involved in national elections, and information about major events like Hurricane Katrina. Like the National Archives, the Library of Congress keeps multiple copies and is on the lookout to avoid format obsolescence, says LeFurgy.

Thanks to its ongoing satellite surveys, the U.S. Geological Survey (USGS) adds about 50TB per month to its archives, and now manages about 4.5PB (counting copies) says John Faundeen, an archivist at the USGS Earth Resources Observation and Science Center in Sioux Falls, S.D.

The center has a three-copy storage strategy: the first copy is online, the second is near-line and the is third off-site. (This mirrors the storage strategy known as information life-cycle management that many enterprise IT shops are adopting.) The Earth Resources Observation and Science Center tries to migrate to new media every three to five years. And it tries to track all of the media it's using by date, to avoid situations where it uses something that a vendor no longer supports, Faundeen explains. Every other year, the center undertakes a study of the off-line media industry to see what's on the market.

Troubled Oscars and libraries

The movie industry got a nasty shock when it became evident that digital data is an impermanent medium. Before Hollywood adopted digital technology, it relied on celluloid film, and movies archived on that medium have lasted a century, according to Maltz of the Academy of Motion Picture Arts and Sciences. A 2007 study by the Academy found that the long-term cost of archiving the master material of a commercial movie on film is $1,059 per year. In digital format, the cost is 11 times higher -- $12,514 per year.

With digital technology, "you have to migrate your data formats and storage media -- your technology infrastructure -- every three to five years, or your data may be unrecoverable," he says.

The Academy has undertaken several projects to try to address the problem. For example, it has launched an effort to develop image file interchange conversion formats and metadata standards that would work for the movie industry. It also built an experimental digital preservation system. "I can say that it turned out to be way more complicated than we understood when we began," Maltz said of Hollywood's digital initiatives.

"Paper libraries are under attack" by people who take books or magazines, and that's why it's essential to maintain multiple digital copies, says Vicky Reich, head of the LOCKSS Program at the Stanford University Libraries.

Digital impermanence has also been a problem for libraries, says Vicky Reich, head of the LOCKSS program at Stanford University Libraries. Not only can material disappear in a twinkling, but troublemakers can tamper with things without leaving any evidence.

"Paper libraries are under attack a lot," she says, explaining that the challenges librarians face include people who remove books or magazine articles on topics they don't approve of. But with printed publications, there are usually multiple copies in libraries scattered across a particular jurisdiction, so it's unlikely that a crusade to eliminate a specific piece of material could be completely successful.

The LOCKSS project takes the same decentralized approach in the digital domain. Participating libraries (currently about 200, predominately at universities) first set up a PC to devote to the archiving project; the machine must have an Internet connection and at least two terabytes of storage and be equipped with open-source LOCKSS software. Each library then chooses material from a list of about 420 publishers that have granted

permission to archive their publications, or a library can get permission elsewhere on its own. The machines then crawl the sources and copy their material. The library machines act as a proxies for the original sites, serving clicks when the original sites can't.

LOCKSS machines with the same originals will compare their content and repair it as necessary. There's no tape backup -- the machines back each other up, Reich says. The "magic number" needed to ensure preservation appears to be six or seven, she adds, and it results from random overlapping among the preservation choices made by the participating libraries

Looking to the future

All in all, those responsible for the stewardship of digital archives don't sound upbeat about the future.

John Faundeen, an archivist at the U.S. Geological Survey, says his agency tries to migrate to new media every three to five years, and it tracks all media it's using by date to avoid situations where it's using something that a vendor no longer supports.

"There is no answer to the core technology issue at the moment, which is that our infrastructure does not take the need for long-term preservation into account," says Maltz.

"The word is *vigilance*," says Faundeen at the USGS. "Preservation efforts must be ongoing. You cannot rest on past work. You must continuously look forward."

Says Le at the National Archives: "It's a never-ending process, and if anything the situation is getting worse." The number of data formats keeps proliferating, and the volume of data arriving at the National Archives could at any moment become overwhelming. Nonetheless, he says, "I am confident that the things we process will be preserved."

The last word, for now, goes to Coughlin. "If you want data to last, you can't just let it sit there," he says. "It has to be active. You have to care for it, or it may eventually get lost."

***Wood** is a freelance writer in San Antonio.*