

That which survives: The problem of data longevity

07 / 02 / 07 |

[The opinions expressed here are mine alone, and not those of Google, Inc. my employer.]



Commentary-- Recently in Tulsa, Oklahoma, a buried time capsule from 1957 was unearthed. Its contents largely consisted of a car, a 1957 Plymouth Belvedere. The creators of the capsule thoughtfully provided a separate can of petrol, assuming that the nuclear powered flying cars of 2007 would long since have moved beyond the need to burn fossil fuels to get around. In literature buried with the car it was billed as becoming "a priceless antique in 2,007!"

Antique, it was; priceless, not so much. Groundwater had seeped into the vault, and rust had completely destroyed the poor vehicle, tail-fins and all. It's easy for us sophisticates of 2007 to laugh at the naïveté of the capsule creators, but would we do much better in leaving readable records for our future historians? I'm not so sure we would.

Increasingly all our records are moving into electronic format. The rapid demise of photographic film is a good illustration on how quickly this can happen. Almost no one buys simple film cameras any more, nearly all of them are digital. These records are now stored on memory-sticks, or downloaded onto personal computers and stored on extremely fragile hard drives, usually without any form of back up. Even as a computer professional, I'm as lazy about backups as everyone else. One unfortunate accident and most of my precious memories would be irretrievably gone. You might hope that the professionals would do a better job. You'd be wrong.

Recently NASA discovered that the original "slow-scan" tapes containing the images from the very first Apollo 11 Moonwalk were [missing](#). The footage that everyone has seen is actually a copy of these tapes, created by pointing a 1969 television camera at the higher resolution video. In June 2007 as I write this the original tapes are still missing. According to NASA there are 2612 boxes containing tapes that may contain the original data but there are 13,000 additional tapes that are missing. I'm reminded of the government warehouse shown at the end of "Raiders of the Lost Ark". I'm sure that NASA has "top men" assigned to the job of finding them.

Even if the tapes are found, I wonder how many 1969-era tape playback machines will be in working order so the data can actually be read? How brittle will the magnetic tape have become? This is only 40 or so years ago. What will the problem be like in 100 years, or 500, or a 1,000? The tapes of the first moon landing are an historical treasure, more important than any records of human travel to anywhere on the Earth. They are a record of the first steps our species made to leave our initial birthplace. If we can't look after these, what hope do we have for less important records? With written records at least we can still read them after many thousands of years, so long as we still understand the language.

I'm not a Luddite, believing that we need to stop migrating our data and records online. The benefits of doing this far outweigh any possible disadvantages. Digitizing all books and museum records, for example, will make records available to scholars who have an Internet connection and who might never have had the chance to visit the physical objects. Being able to search through them all is pretty nifty too. Unexpected but important new discoveries like new meteorite impact craters have been made now that global satellite imagery is freely available to anyone on the Internet. Who knows what people will find as more and more human knowledge moves into the public space from its current state of needing physical access? I don't want to give this up. But it would be good to have some thought given to preserving the ability to access the important historical records of our time.

I think proprietary record formats will present a problem for historians. Perhaps not in the short-term, but certainly in the medium to long term (and remember I'm talking about hundreds if not thousands of years now). Imagine that some historian in 500 years time discovers Vice President Cheney's "undisclosed location" and finds his secret laptop computer. "Finally," the historian thinks, "we will know who advised this administration about energy policy!" as he swims back to the surface of the ocean above the Washington monument. Unfortunately it turns out the data was written in the "Word-mangler for Windows 2002" format, for which no specifications were ever published, and which was deliberately designed to be difficult for the competition to read.

Joking aside, proprietary record formats will increase the difficulty of preserving our culture, on top of the problems with obsolete hardware interfaces and the decay of storage media we think of as permanent. File data formats that are not published standards are just asking for trouble for long term data storage. Much though I prefer the OpenOffice Open Document Format (ODF) for documents, the Microsoft Office Open XML (OOXML) format is also a documented format (although without any other implementations as yet) so it shouldn't cause problems for long term storage. However, most of the world's documents in both governments and corporations are still in undocumented proprietary formats, and it sometimes ends up that the documents that people don't think are worth preserving are the ones historians are most excited to find.

I'm not too worried about the "doomsday scenarios" of a post-industrial, non-electrical civilization having lost all scientific knowledge by being unable to read our word-processor documents. I somehow suspect that a society in that state has more to worry about than whether they can access old corporate financial records. No, the thing that worries me is someone suddenly wanting access to an old video recording of the fall of the Berlin wall and finding that there are only copies of copies of copies, the original "truth" of the incident being lost centuries ago. Or even worse, being unable to determine which version of the video shows the real event.

In a book by my favorite cartoonist, ["2024"](#) by Ted Rall, a modern re-telling of Orwell's 1984, the hero idly edits the records of the Nobel Peace Prize winners and adds an obscure punk rock singer to the list. By the time the media picks up and re-broadcasts his changes he's forgotten he even did it and blindly accepts the new history along with everyone else.

Once an event has gone from living memory, and the only records are electronic and mutable, how can we ever know the truth of our past? Maybe a fragile DVD or other physical record that can be dated to an accurate time near a historical event will become more valuable than any

electronic data, or might serve as the ultimate arbiter of the alternate digital versions already available. Maybe the curse we will leave to future generations is so many versions of the same data it will be impossible to tell which one, if any, was the original truth. Write back and let me know what you think.

Finally I can't help apologizing for criticizing NASA above by pointing out that they are responsible for the only really long term data storage human beings have ever created. On the Pioneer and Voyager space probes are plaques containing data designed to be read by extra-terrestrial intelligences, showing the position of our solar system. The Voyager space craft also carry a disk containing audio and video from the planet Earth. A billion years from now, well outside the Solar system, it is hoped these messages will still be readable by any intelligence able to retrieve them. Now that's a time capsule worth celebrating.

Jeremy Allison is one of the lead developers on the Samba Team, a group of programmers developing an Open Source Windows compatible file and print server product for UNIX systems. Developed over the Internet in a distributed manner similar to the Linux system, Samba is used by all Linux distributions as well as many thousands of corporations worldwide. Jeremy handles the co-ordination of Samba development efforts and acts as a corporate liason to companies using the Samba code commercially. He works for Google, Inc. who fund him to work full-time on improving Samba and solving the problems of Windows and Linux interoperability.

That which survives: The problem of data longevity

By [Jeremy Allison](#), Special to ZDNet

Published on [ZDNet News](#): July 2, 2007, 6:31 AM PT, http://news.zdnet.com/2100-9584_22-6194495.html. July 3, 2007.