Businesses worry about long-term data losses
Betts, Mitch
Computerworld; Sep 20, 1999; 33, 38; ABI/INFORM Collection
pg. 22

Businesses Worry About Long-Term Data Losses

Will we access our saved data in 20 years?

BY MITCH BETTS

We know what Alexander Graham Bell said in the first telephone call: "Mr. Watson, come here; I want you." We don't know what Ray Tomlinson said in the first network e-mail in 1972. He doesn't remember exactly, and he didn't save it.

"The main reason it's lost to history is just that it didn't seem worth saving [at the time]," said Tomlinson, principal engineer at GTE Corp.'s BBN Technologies unit in Cambridge, Mass.

In retrospect, it would be nice to have that piece of history, Tomlinson said, but you can't save everything — or even recover it. "Even if backup tapes did exist, they might not be readable. They were just mag tapes, and after seven or eight years, the oxide starts falling off, especially from tapes from that era," he said.

There's the rub: Digital information — from the historically interesting to the economically vital — is at risk of disappearing or becoming inaccessible because of the deterioration of storage media like magnetic tapes (see chart). Other culprits include ever-changing data formats and the fact that software and hardware become obsolete quickly.

Major portions of the 1960 U.S. census, for example, could be read only with a Univac type II-A tape drive, which was a museum relic just 16 years after the census, according to a report by the Council on Library and Information Resources (CLIR) in Washington. It took a major data-rescue effort to copy the raw data to industry-standard tapes.

There's a reason it's called "machine-readable" data. Pre-1979 Landsat satellite data is inaccessible, for instance, because it was recorded on ancient Xerox Corp. computers that can no longer be operated, according to the journal *Science*. And just try finding a PC that can read a WordPerfect 4.0 file on a 5.25-in. diskette — and not lose the footnotes or formatting.

Historians, librarians and archivists — who have a natural affinity for really old stuff — are already alarmed about the loss of cultural and government records. But digital preservation is becoming a business issue, too, as certain industries find that they need to keep data longer and longer for regulatory or business reasons.

Pharmaceutical companies are a prime example. They need to keep records about new drugs for as long as the drug is on the market and records about clinical trials for the life span of the patients, said Jeff Rothenberg, a senior computer scientist at the Rand Corp., a think tank in Santa Monica, Calif.

"IT DIDN'T seem worth saving," says GTE's Ray Tomlinson, who sent the first e-mail in 1972

And now companies are spending millions to build data warehouses intended to hold cradle-to-grave records of customer purchases and health care. "We do have concerns that 10 years from now, 20 years from now, what is the likelihood that [the data] will even be retrievable," said Joe Bruscato, chief architect of data warehousing at Anthem Blue Cross and Blue Shield in Cincinnati.

"Companies aren't doing enough to make sure that data that was archived several years ago is retrievable. It's a valid concern," Bruscato said. "Just as companies have disaster recovery plans, they need data recovery plans."

Analysts say the quality of corporate data archives is all over the map, ranging from ideal setups of optical jukeboxes and tapes in climate-controlled vaults to haphazard, undocumented storage of reels in basements. And some companies don't archive at all, figuring employees will print out anything that's really important.

But short-term thinking and sloppy record keeping could lead to data disasters for corporations if they lose valuable information because the magnetic tape decomposed or because they no longer have the software or hardware required to retrieve it.

Besides proper stewardship of storage media, companies need to have a records-management team that maintains a central repository of metadata — sort of a catalog of the company's data and formats — and schedules the conversion to new media, said Wendy Anne Ailor, records manager at a major pharmaceutical company. The trick is to schedule conversions well before the particular storage medium is expected to deteriorate.

Managers should also log the software and hardware versions required to read or manipulate the data — and keep an eye out for discontinued models.

Ailor said her company had 1.5 million image files stored on a Sony Corp. jukebox, but Sony announced two years ago that it would discontinue support for the 12-in. optical platters after next year. So the company is copying the images to 5.25-in. platters from Hewlett-Packard Co.

Companies may need to keep an "application archive," said Kris Newton, an analyst at Strategic Research Corp. in Santa Barbara, Calif. She explained: Users who converted legacy mainframe systems to Unix or Windows NT — perhaps for year 2000 compliance — need to keep a copy of the old applications so they can read the legacy data. And they may need a hardware emulator, because they probably don't want to keep an old mainframe around just for that purpose.

The classic solution to the problem of media decay is to copy the data to newer storage media. But migration isn't perfect.

"As you move material from system to system, there are all sorts of chances for errors — accidental or deliberate — to crop up," said Anne Gilliland-Swetland, an assistant professor of information studies at the University of California in Los Angeles. She's also co-director of a new research project trying to formulate model policies, standards and strategies for ensuring that authentic electronic records can be preserved over long periods of time.

Eye to the Future

The key to designing an archival system is to keep an "eye to the future" with a long-term view, said Kermit Patton, a researcher at SRI Consulting in Menlo Park, Calif.

"Most companies tend to be short term in their perspective, so they're thinking only of getting [data] onto the next generation [of storage media]," Patton said. "If you design the system and data standards while thinking of multiple generations, you're in better shape."

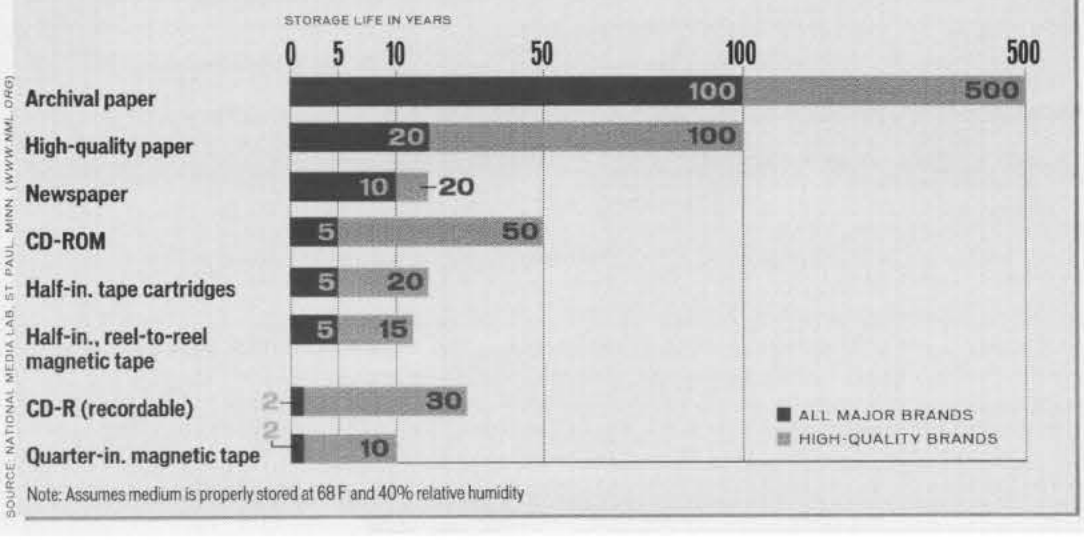
We won't really know how long today's storage media will reliably hold data until we let it age a decade or two. And we won't see whether the data is corrupted or missing until we try to read it.

That's what happened when technicians checked the magnetic tapes holding data from the 1976 Viking mission to Mars: They found that 10% to 20% of the tapes had significant errors, the CLIR report said. The technicians called magnetic tape "a disaster" as an archival storage medium.

"You can see books crumbling, but there's no way of looking at a magnetic tape and seeing errors on it — you have to run the tape. It's very labor-intensive," said Abby Smith, preservation expert at the CLIR. "So we won't know how big the problem is until the first time we go back and try to use that information."

Life Expectancy of Storage Media

Old newspapers last about 20 years; high-quality magnetic tapes can last up to 15 years



Note: Assumes medium is properly stored at 68 F and 40% relative humidity

SOURCE: NATIONAL MEDIA LAB, ST. PAUL, MINN. (WWW.NML.ORG)