The New York Times
nytimes.com

April 9, 2008

# In Storing 1's and 0's, the Question Is $

By JOHN SCHWARTZ

LISTEN. Do you hear it? The bits are dying.

The digital revolution has spawned billions upon billions of gigabytes of data, from the vast electronic archives of government and business to the humblest photo on a home PC. And the trove is growing — the International Data Corporation, a technology research and advisory firm, estimates that by 2011 the digital universe of ones and zeros will be 10 times the size it was in 2006.

But the downside is that much of this data is ephemeral, and society is headed toward a kind of digital Alzheimer's. What's on those old floppies stuck in a desk drawer? Can anything be read off that ancient mainframe's tape drive? Will today's hard disk be tomorrow's white elephant?

Data is "the natural resource for the Internet age," said Francine Berman, director of the San Diego Supercomputer Center at the University of California, San Diego, a national center for high-performance computing resources. But, she added, "digital data is enormously fragile." It can degrade as it is stored, copied and transferred between hard drives across data networks. The storage systems might not be around or accessible in the future — it is like putting precious information on eight-track tapes.

"It's very important that we have an awareness that digital preservation has to be a part of our infrastructure," Dr. Berman said. But as the problem has been studied over the years, researchers have found that "there's no one-size-fits-all model for preserving data in the digital age," she added. And there's an even bigger potential roadblock: how to pay for it. "Economic sustainability," Dr. Berman said, is "the gorilla in the room."

The National Science Foundation has begun a $100 million program over the next five years for an initiative, known as DataNet, that will help develop methods and technologies to keep the data we create. The goal is more than safeguarding the family's digital photo album: it's to preserve science and engineering data in ways that are "open, extensible and evolvable" — in other words, not just to make sure that bits aren't lost but also to make them accessible and usable far into the future.

At the same time, a second National Science Foundation-supported effort is finding ways to address the cost of saving digital memories. Dr. Berman leads this two-year task force with Brian Lavoie, a research scientist at the Online Computer Library Center, a nonprofit organization near Columbus, Ohio, that helps more than 60,000 libraries around the world find, share and preserve materials.

Dr. Lavoie said the task force would outline ways that digital preservation could be used in diverse situations, with an eye to economy and sustainability. "The common thing among all of them is that somebody has to pay for it," he said.

For all their qualities, electrons can seem awfully feeble when compared with a good old-fashioned book. “With the right kind of paper and the right kind of stewardship,” Dr. Berman said, “you can keep a book for 100 years or more.” The interface is as simple as it gets: open the book and look at the page.

By contrast, in the hundred years that a book might have spent on the shelf, technology might have gone through “dozens of generations of storage media,” she said.

No one is suggesting that we try to hold on to every bit of data lingering in every obsolete corner. Choices must be made about the kind of material that should be kept fresh and accessible for 5 years, or 50, or 1,000. Census data? Put it on the “forever” drive, please. To-do lists? A little less crucial.

Dr. Berman identifies collections like the Protein Data Bank, run by the Research Collaboratory for Structural Bioinformatics. A repository of information on protein structures, it represents a research investment of more than $80 billion, said Dr. Berman, whose supercomputer center is a collaborating institution. That kind of data, which could lead to new understanding of the body’s functions and to new drugs, is also a keeper.

Those in the digital-preservation field have talked for years about technologies that will achieve their goals. In a world where steeply falling hardware prices allow companies like Google to create vast server farms, preservationists have come up with ideas for electronic depositories, big and small, that businesses or government could build.

But the talk goes well beyond simple storage to true preservation, which ensures that information remains accessible. The plans therefore include making data retrievable with technologies like format migration, in which outdated files could be made readable in a more generic format, and computer emulation, in which one machine would pretend to be an older computer that could make sense of old files.

All that work is going on, Dr. Lavoie said, but “that misses the point” that the task force was formed to examine: ensuring that the various technologies make economic sense. “You can have the most elegant technological solution to the digital-preservation problem, but if there’s no economics underpinning it, then there’s no solution at all,” he said.

So while it is important to develop technologies that will make digital preservation simple and inexpensive, Dr. Lavoie said that the field was “not about picking winners and losers at all.”

He described an economic framework that would follow the course of the evolution of the computer-security market. In that case, private companies emerged to handle the needs of industry and government to protect against hacking, while others developed products and services that smaller organizations and even consumers could use. Some companies developed their own expertise and did the work in house as well.

The government spurred development, too, through tax breaks, monetary penalties for lax security practices in the financial industry and paying for security initiatives like the CERT center at Carnegie Mellon University, which monitors computer attacks.

“The question, I think, is articulating that full menu of models” so that the development of a preservation market will be encouraged, Dr. Lavoie said. “That’s something we’re really missing now.”

Margaret Hedstrom, associate professor at the School of Information at the University of Michigan, has been preaching preservation for 30 years, ever since she got a job organizing Wisconsin state records. But in a throwaway culture, it’s been a hard argument to make stick.

Now, she said, “one of the things that’s changing, finally, is people in places like the National Science Foundation are paying enough attention to this problem and understand its scale to start making investments that can make a difference.”

The concern is about more than losing any particular set of data, Dr. Hedstrom said. “The issue is about losing the ability, in a systematic way, of being able to preserve anything.”

Yes, people know how to keep data, she said, but she added that it was just as important “to preserve the right information” well enough to keep it meaningful and accessible.

“There might be 100 versions of a report on a company’s hard drive, but which one was the final draft?” Dr. Hestrom said. “How was the underlying data used? Which architectural drawings of the many versions generated for a project were actually used to erect the building, and what was the chain of decisions that led to the brick-and-mortar result?

“It’s not that the bits aren’t lying around,” she continued. “They may or may not be lying around. But being able to understand how they were collected,” and being able to ascertain how the underlying data was used, makes the information useful. People think that because the cost of storage is dropping “we can save everything,” she said. “But that’s based on a naïve view of what ‘everything’ actually is.”

Efforts like the Internet Archive (archive.org), which trolls and collects billions of pages from the World Wide Web, are laudable but merely represent “the surface Web,” she said. The underlying data that enriches understanding isn’t present in that kind of collection, she said. “The Internet Archive is great for what it is,” she said, “but we’re not going to solve the preservation problem with one small not-for-profit organization” and volunteers.

She said she was thrilled, therefore, to see serious projects coming from the National Science Foundation and heartened that many approaches were being considered. “If everybody’s doing the same thing, we might all be making the same mistake,” she said.

She added that she hoped the movement to hold on to our digital memories would finally succeed. “It’s taken longer than I would have liked,” she said, “but I think we’re getting there.”