

TechnologyReview.com

Your *Opinion*
MATTERS

Take a quick survey.

Yes!

 QUESTIONMARKET.com

[Print](#) | [Forums](#)

The Fading Memory of the State

By David Talbot July 2005

The official repository of retired U.S. government records is a boxy white building tucked into the woods of suburban College Park, MD. The National Archives and Records Administration (NARA) is a subdued place, with researchers quietly thumbing through boxes of old census, diplomatic, or military records, and occasionally requesting a copy of one of the computer tapes that fill racks on the climate-controlled upper floors. Researchers generally don't come here to look for contemporary records, though. Those are increasingly digital, and still repose largely at the agencies that created them, or in temporary holding centers. It will take years, or decades, for them to reach NARA, which is charged with saving the retired records of the federal government (NARA preserves all White House records and around 2 percent of all other federal records; it also manages the libraries of 12 recent presidents). Unfortunately, NARA doesn't have decades to come up with ways to preserve this data. Electronic records rot much faster than paper ones, and NARA must either figure out how to save them permanently, or allow the nation to lose its grip on history.

One clear morning earlier this year, I walked into a fourth-floor office overlooking the woods. I was there to ask Allen Weinstein--sworn in as the new Archivist of the United States in February--how NARA will deal with what some have called the pending "tsunami" of digital records. Weinstein is a former professor of history at Smith College and Georgetown University and the author of *Perjury: The Hiss-Chambers Case* (1978) and coauthor of *The Story of America* (2002). He is 67, and freely admits to limited technical knowledge. But a personal experience he related illustrates quite well the challenges he faces. In 1972, Weinstein was a young historian suing for the release of old FBI files. FBI director J. Edgar Hoover--who oversaw a vast machine of domestic espionage--saw a Washington Post story about his efforts, wrote a memo to an aide, attached the Post article and penned into the newspaper's margin: "What do we know about Weinstein?" It was a telling note about the mind-set of the FBI director and of the federal bureaucracy of that era. And it was saved--Weinstein later found the clipping in his own FBI file.

But it's doubtful such a record would be preserved today, because it would likely be "born digital" and follow a convoluted electronic path. A modern-day J. Edgar Hoover might first use a Web browser to read an online version of the Washington Post. He'd follow a link to the Weinstein story. Then he'd send an e-mail containing the link to a subordinate, with a text note: "What do we know about Weinstein?" The subordinate might do a Google search and other electronic searches of Weinstein's life, then write and revise a memo in Microsoft Word 2003, and even create a multimedia PowerPoint presentation about his findings before sending both as attachments back to his boss.

▼ ADVERTISEMENT ▼



The Royal Bank of Scotland Group didn't become one of the largest banks in the world by sitting around talking about it.

[Click Here to Find Out More](#)

Make it happen

Definitions

Megabyte

1,024 kilobytes. The length of a short novel or the storage available on an average floppy disk.

Gigabyte

1,024 megabytes. Roughly 100 minutes of CD-quality stereo sound.

Terabyte

1,024 gigabytes. Half of the content in an academic research library.

Petabyte

1,024 terabytes. Half of the content in all U.S. academic research libraries.

Exabyte

1,024 petabytes. Half of all the information generated in 1999.

What steps in this process can be easily documented and reliably preserved over decades with today's technology? The short answer: none. "They're all hard problems," says Robert Chadduck, a research director and computer engineer at NARA. And they are symbolic of the challenge facing any organization that needs to retain electronic records for historical or business purposes.

Imagine losing all your tax records, your high school and college yearbooks, and your child's baby pictures and videos. Now multiply such a loss across every federal agency storing terabytes of information, much of which must be preserved by law. That's the disaster NARA is racing to prevent. It is confronting thousands of incompatible data formats cooked up by the computer industry over the past several decades, not to mention the limited lifespan of electronic storage media themselves. The most famous documents in NARA's possession--the Declaration of Independence, the Constitution, and the Bill of Rights--were written on durable calfskin parchment and can safely recline for decades behind glass in a bath of argon gas. It will take a technological miracle to make digital data last that long.

But NARA has hired two contractors--Harris Corporation and Lockheed Martin--to attempt that miracle. The companies are scheduled to submit competing preliminary designs next month for a permanent Electronic Records Archives (ERA). According to NARA's specifications, the system must ultimately be able to absorb any of the 16,000 other software formats believed to be in use throughout the federal bureaucracy--and, at the same time, cope with any future changes in file-reading software and storage hardware. It must ensure that stored records are authentic, available online, and impervious to hacker or terrorist attack. While Congress has authorized \$100 million and President Bush's 2006 budget proposes another \$36 million, the total price tag is unknown. NARA hopes to roll out the system in stages between 2007 and 2011. If all goes well, Weinstein says, the agency "will have achieved the start of a technological breakthrough equivalent in our field to major 'crash programs' of an earlier era--our Manhattan Project, if you will, or our moon shot."

Data Indigestion

NARA's crash data-preservation project is coming none too soon; today's history is born digital and dies young. Many observers have noted this, but perhaps none more eloquently than a U.S. Air Force historian named Eduard Mark. In a 2003 posting to a Michigan State University discussion group frequented by fellow historians, he wrote: "It will be impossible to write the history of recent diplomatic and military history as we have written about World War II and the early Cold War. Too many records are gone. Think of Villon's haunting refrain, 'Ou sont les neiges d'antan?' and weep....History as we have known it is dying, and with it the public accountability of government and rational public administration." Take the 1989 U.S. invasion of Panama, in which U.S. forces removed Manuel Noriega and 23 troops lost their lives, along with at least 200 Panamanian fighters and 300 civilians. Mark wrote (and recently stood by his comments) that he could not secure many basic records of the invasion, because a number were electronic and had not been kept. "The federal system for maintaining records has in many agencies--indeed in every agency with which I am familiar--collapsed utterly," Mark wrote.

Of course, managing growing data collections is already a crisis for many institutions, from hospitals to banks to universities. Tom Hawk, general manager for enterprise storage at IBM, says that in the next three years, humanity will generate more data--from websites to digital photos and video--than it generated in the previous 1,000 years. "It's a whole new set of challenges to IT organizations that have not been dealing with that level of data and complexity," Hawk says. In 1996, companies spent 11 percent of their IT budgets on storage, but that figure will likely double to 22 percent in 2007, according to International Technology Group of Los Altos, CA.

Still, NARA's problem stands out because of the sheer volume of the records the U.S. government produces and receives, and the diversity of digital technologies they represent. "We operate on the premise that somewhere in the government they are using every software program that has ever been sold, and some that were never sold because they were developed for the government," says Ken Thibodeau, director of the Archives' electronic-records program. The scope of the problem, he adds, is "unlimited, and it's open ended, because the formats keep changing."

The Archives faces more than a Babel of formats; the electronic records it will eventually inherit are piling up at an ever accelerating pace. A taste: the Pentagon generates tens of millions of images from personnel files each year; the Clinton White House generated 38 million e-mail messages (and the current Bush White House is expected to generate triple that number); and the 2000 census returns were converted into more than 600 million TIFF-format image files, some 40 terabytes of data. A single patent application can contain a million pages, plus complex files like 3-D models of proteins or CAD drawings of aircraft parts. All told, NARA expects to receive 347 petabytes (see "Definitions") of electronic records by 2022.

Currently, the Archives holds only a trivial number of electronic records. Stored on steel racks in NARA's 11-year-old facility in College Park, the digital collection adds up to just five terabytes. Most of it consists of magnetic tapes of varying ages, many of them holding a mere 200 megabytes apiece--about the size of 10 high-resolution digital photographs. (The electronic holdings include such historical gems as records of military psychological-operations squads in Vietnam from 1970 to 1973, and interviews, diaries, and testimony collected by the U.S. Department of Justice's Watergate Special Prosecution Force from 1973 to 1977.) From this modest collection, only a tiny number of visitors ever seek to copy data; little is available over the Internet.

Because the Archives has no good system for taking in more data, a tremendous backlog has built up. Census records, service records, Pentagon records of Iraq War decision-making, diplomatic messages--all sit in limbo at federal departments or in temporary record-holding centers around the country. A new avalanche of records from the Bush administration--the most electronic presidency yet--will descend in three and a half years, when the president leaves office. Leaving records sitting around at federal agencies for years, or decades, worked fine when everything was on paper, but data bits are nowhere near as reliable--and storing them means paying not just for the storage media, but for a sophisticated management system and extensive IT staff.

Data under the Desk

The good news is that at least some of the rocket science behind the Archives' "moon shot" is already being developed by industry, other U.S. government agencies, and foreign governments. For example, Hewlett-Packard, IBM, EMC, PolyServe, and other companies have developed "virtual storage" technologies that automatically spread terabytes of related data across many storage devices, often of different types. Virtualization frees up IT staff, balances loads when demand for the data spikes, and allows hardware upgrades to be carried out without downtime. Although the Archives will need technologies far beyond virtual storage, the commercial efforts form a practical foundation. The Archives may also benefit from the examples of digital archives set up in other nations, such as Australia, where archivists are using open-source software called XENA (for XML Electronic Normalizing of Archives) to convert records into a standardized format that will, theoretically, be readable by future technologies. NARA will also follow the lead of the U.S. Library of Congress, which in recent years has begun digitizing collections ranging from early American sheet music to immigration photographs and putting them online, as part of a \$100 million digital preservation program.

But to extend the technology beyond such commercial and government efforts, NARA and the National Science Foundation are funding research at places like the San Diego Supercomputer Center. There, researchers are, among other things, learning how to extract data from old formats rapidly and make them useful in modern ones. For example, San Diego researchers took a collection of data on airdrops during the Vietnam War--everything from the defoliant Agent Orange to pamphlets--and reformatted it so it could be displayed using nonproprietary versions of digital-mapping programs known as geographic information systems, or GIS (see ["Do Maps Have Morals?" Technology Review](#), June 2005). Similarly, they took lists of Vietnam War casualties and put them in a database that can show how they changed over the years, as names were added or removed. These are the kinds of problems NARA will face as it "ingests" digital collections, researchers say. "NARA's problem is they will be receiving massive amounts of digital information in the future, and they need technologies that will help them import that data into their ERA--hundreds of millions of items, hundreds of terabytes of data," says Reagan Moore, director of data-knowledge computing at the San Diego center.

Another hive of research activity on massive data repositories: MIT. Just as the government is losing its grip on administrative, military, and diplomatic history, institutions like MIT are losing their hold on research data--including the early studies and communications that led to the creation of the Internet itself. "MIT is a microcosm of the problems [NARA] has every day," says MacKenzie Smith, the associate director for technology at MIT Libraries. "The faculty members are keeping their research under their desks, on lots and lots of disks, and praying that nothing happens to it. We have a long way to go."

Now MIT is giving faculty another place to put that data. Researchers can log onto the Internet and upload information--whether text, audio, video, images, or experimental data sets--into DSpace, a storage system created in collaboration with Hewlett-Packard and launched in 2002 (see ["MIT's DSpace Explained"](#)). DSpace makes two identical copies of all data, catalogues relevant information about the data (what archivists call "metadata," such as the author and creation date), and gives each file a URL or Web address. This address won't change even if, say, the archivist later wants to put a given file into a newer format--exporting the contents of an old Word document into a PDF file, for instance. Indeed, an optional feature in DSpace will tell researchers which files are ready for such "migration."

Because the software behind DSpace is open source, it is available for other institutions to adapt to their own digital-archiving needs; scores have already done so. Researchers at MIT and elsewhere are working on improvements such as an auditing feature that would verify that a file hasn't been corrupted or tampered with, and a system that checks accuracy when a file migrates into a new format. Ann Wolpert, the director of MIT Libraries (and chair of *Technology Review's* board of directors), says DSpace is just a small step toward tackling MIT's problems, never mind

NARA's. "These changes have come to MIT and other institutions so rapidly that we didn't have the technology to deal with it," Wolpert says. "The technology solutions are still emerging." Robert Tansley, a Hewlett-Packard research scientist who worked on DSpace, says the system is a good start but cautions that "it is still quite new. It hasn't been tested or deployed at a massive scale, so there would need to be some work before it could support what the National Archives is looking at."

Digital Marginalia

But for all this promise, NARA faces many problems that researchers haven't even begun to think about. Consider Weinstein's discovery of the Hoover marginalia. How could such a tidbit be preserved today? And how can any organization that needs to track information--where it goes, who uses it, and how it's modified along the way--capture those bit streams and keep them as safe as older paper records? Saving the text of e-mail messages is technically easy; the challenge lies in managing a vast volume and saving only what's relevant. It's important, for example, to save the e-mails of major figures like cabinet members and White House personnel without also bequeathing to history trivial messages in which mid-level bureaucrats make lunch arrangements. The filtering problem gets harder as the e-mails pile up. "If you have 300 or 400 million of anything, the first thing you need is a rigorous technology that can deal with that volume and scale," says Chadduck. More and more e-mails come with attachments, so NARA will ultimately need a system that can handle any type of attached file.

Version tracking is another headache. In an earlier era, scribbled cross-outs and margin notes on draft speeches were a boon to understanding the thinking of presidents and other public officials. To see all the features of a given Microsoft Word document, such as tracked changes, it's best to open the document using the same version of Word that the document's creator used. This means that future researchers will need not only a new piece of metadata--what software version was used--but perhaps even the software itself, in order to re-create fonts and other formatting details faithfully. But saving the functionality of software--from desktop programs like Word to the software NASA used to test a virtual-reality model of the Mars Global Surveyor, for example--is a key research problem. And not all software keeps track of how it was actually used. Why might this matter? Consider the 1999 U.S. bombing of the Chinese embassy in Belgrade. U.S. officials blamed the error on outdated maps used in targeting. But how would a future historian probe a comparable matter--to check the official story, for example--when decision-making occurred in a digital context? Today's planners would open a map generated by GIS software, zoom in on a particular region, pan across to another site, run a calculation about the topography or other features, and make a targeting decision.

If a historian wanted to review these steps, he or she would need information on how the GIS map was used. But "currently there are no computer science tools that would allow you to reconstruct how computers were used in high-confidence decision-making scenarios," says Peter Bajcsy, a computer scientist at the University of Illinois at Urbana-Champaign. "You might or might not have the same hardware, okay, or the same version of the software in 10 or 20 years. But you would still like to know what data sets were viewed and processed, the methods used for processing, and what the decision was based on." That way, to stay with the Chinese embassy example, a future historian might be able to independently assess whether the database about the embassy was obsolete, or whether the fighter pilot who dropped the bomb had the right information before he took off. Producing such data is just a research proposal of Bajcsy's. NARA says that if such data is collected in the future, the agency will add it to the list of things needing preservation.

Data Curators

Even without tackling problems like this, NARA has its hands full. For three years, at NARA's request, a National Academy of Sciences panel has been advising the agency on its electronic-records program. The panel's chairman, computer scientist Robert F. Sproull of Sun Microsystems Laboratories in Burlington, MA, says he has urged NARA officials to scale back their ambitions for the ERA, at least at the start. "They are going to the all-singing, all-dancing solution rather than an incremental approach," Sproull says. "There are a few dozen formats that would cover most of what [NARA] has to do. They should get on with it. Make choices, encourage people submitting records to choose formats, and get on with it. If you become obsessed with getting the technical solution, you will never build an archive." Sproull counsels pragmatism above all. He points to Google as an example of how to deploy a workable solution that satisfies most information-gathering needs for most of the millions of people who use it. "What Google says is, 'We'll take all comers, and use best efforts. It means we won't find everything, but it does mean we can cope with all the data,'" Sproull says. Google is not an archive, he notes, but in the Google spirit, NARA should attack the problem in a practical manner. That would mean starting with the few dozen formats that are most common, using whatever off-the-shelf archiving technologies will likely emerge over the next few years. But this kind of preservation-by-triage may not be an option, says NARA's Thibodeau. "NARA does not have discretion to refuse to preserve a format," he says. "It is inconceivable to me that a court would approve of a decision not to preserve e-mail attachments, which often contain the main substance of the communication, because it's not in a format NARA chose to preserve."

Meanwhile, the data keep rolling in. After the 9/11 Commission issued its report on the attacks on the World Trade Center and the Pentagon, for example, it shut down and consigned all its records to NARA. A good deal of paper, along with 1.2 terabytes of digital information on computer hard disks and servers, was wheeled into NARA's College Park facility, where it sits behind a door monitored by a video camera and secured with a black combination lock. Most of the data, which consist largely of word-processing files and e-mails and their attachments, are sealed by law until January 2, 2009. They will probably survive that long without heroic preservation efforts. But "there's every reason to say that in 25 years, you won't be able to read this stuff," warns Thibodeau. "Our present will never become anybody's past."

It doesn't have to be that way. Projects like DSpace are already dealing with the problem. Industry will provide a growing range of partial solutions, and researchers will continue to fill in the blanks. But clearly, in the decades to come, archives such as NARA will need to be staffed by a new kind of professional, an expert with the historian's eye of an Allen Weinstein but a computer scientist's understanding of storage technologies and a librarian's fluency with metadata. "We will have to create a new profession of 'data curator'--a combination of scientist (or other data specialist), statistician, and information expert," says MacKenzie Smith of the MIT Libraries.

The nation's founding documents are preserved for the ages in their bath of argon gas. But in another 230 years or so, what of today's electronic records will survive? With any luck, the warnings from air force historian Mark and NARA's Thibodeau will be heeded. And historians and citizens alike will be able to go online and find that NARA made it to the moon, after all.

73.180803538605

CATEGORY_KEYWORDS:MEDIA

UID:- UL: 2 - IE:False/