FEATURE ARTICLE

WEEKLY FEATURE

Eternal Bits

How can we preserve digital files and save our collective memory?

By MacKenzie Smith

It took two centuries to fill the U.S. Library of Congress in Washington, D.C., with more than 29 million books and periodicals, 2.7 million recordings, 12 million photographs, 4.8 million maps, and 57 million manuscripts. Today it takes about 15 minutes for the world to churn out an equivalent amount of new digital information. It does so about 100 times every day, for a grand total of five exabytes annually. That's an amount equal to all the words ever spoken by humans, according to Roy Williams, who heads the Center for Advanced Computing Research at the California Institute of Technology, in Pasadena.

While this stunning proliferation of information underscores the ease with which we can create digital data, our capacity to make all these bits accessible in 200 or even 20 years remains a work in progress.

In an era when the ability to read a document, watch a video, or run a simulation could depend on having a particular version of a program installed on a specific computer platform, the usable life span of a piece of digital content can be less than 10 years. That's a recipe for disaster when you consider how much we rely on stored information to maintain our scholarly, legal, and cultural record and to help us with, and profit from, our digital labor. Indeed, the ephemeral nature of both data formats and storage media threatens our very ability to maintain scientific, legal, and cultural continuity, not on the scale of centuries, but considering the unrelenting pace of technological change, from one decade to the next.

At the Massachusetts Institute of Technology Libraries, in Cambridge, where I am associate

director for technology, we are attacking the problem of maintaining and sharing digital content over the long haul with a project called DSpace, which we embarked on with Hewlett-Packard Co., in Palo Alto, Calif., in 2000. For this digital repository we have built a simple, open-source software application that not only accepts digital materials and makes them available on the Web but also puts them into a data-management regime that helps to preserve them for generations to come.

The problem of digital preservation confronts everyone, even people who don't suspect that it could ever affect them. For instance, people who created their Ph.D. dissertations in WordStar in the mid-1980s can no longer read them. And there's the person who posted this question to the Obsolete Computer Museum's Web site: "I have years of letters to family and friends written by my deceased mother that are on [1983 IBM] DisplayWriter 8-inch disks. Can anyone tell me how to transfer them to a PC?"

Now view the challenge on the scale of centuries and the vast amounts of data being generated today that could be useful to tomorrow's researchers, executives, and public officials. To avoid handicapping future generations with huge blanks in the historical record, digital archivists are already fighting a pitched battle against a problem called bit rot—the well-documented degradation of data on magnetic storage media due to physical factors, such as alpha particles emitted from computer chips.

But saving raw data solves only part of the preservation problem. We also want to be able to read, play, or watch these bits when we need to. Then there are pesky legal obligations, which demand that we be able to guarantee that certain records haven't been altered by human hands or computer malfunction.

Awareness of the problem is growing rapidly, however, especially in large organizations. As official government and corporate records become entirely digital, certain obligations to keep these records around for future scrutiny must be met. In the United States, for example, the Sarbanes-Oxley Act of 2002 requires that business records, including electronic records and e-mail, be saved for "not less than five years." In some industries, such as pharmaceuticals, the regulations for record retention are much longer—30 years or more.

In Europe, the Council of Europe Convention on Cybercrime, an agreement that addresses problems posed by criminals such as identity thieves, pedophiles, and terrorists, contains provisions for governments to compel Internet service providers to preserve data that could be used as evidence in a court of law.

These new requirements, along with an increasing dependence on digital content across the board, have spurred companies, governments, and universities to devise or acquire ways to preserve just about everything stored as bits. The potential magnitude of the problem is staggering, encompassing literally every means people use to record and store information. That includes books, journals, maps, music, movies, e-mail, corporate records, government documents, course materials, data sets, and databases. It also covers scientific models, lab notebooks, parish records, family histories, and global weather data—even last night's news or Red Sox game broadcasts.

DSpace is storing and preserving materials just like these at MIT and 100 other organizations worldwide. Among them are Cornell University; the University of Toronto; the University of Cambridge, in England; the Australian National University; and the Hong Kong University of Science and Technology. Like Linux and other open-source software projects, DSpace has a growing group of committed programmers distributed across the globe who continually maintain and improve it.

CREATING A DIGITAL ARCHIVE that will last through the centuries depends not only on technology but also on administrative skills and policies. Before we choose a particular method of preserving digital documents, images, audio files, or data sets, we need to consider the ultimate goal. Is our objective to be able to read, hear, or play something a hundred years from now? Or is it to prove provenance in a court of law—for instance, that certain clinical test data really are the original ones we based a particular drug on?

The former is hard, but the latter is really hard. When you want to be able to read something in a hundred years, you probably don't mind if the background color and the fonts are a bit different as long as the words themselves haven't changed. In that case, a simple migration strategy should suffice: keep copying the file into newer formats that can be interpreted without any substantive changes by modern software.

But if you also need to know that what you're reading is authentic, in the sense that the billionaire's will you're reading is the real thing, then you need data essentially to be stamped with a digital signature to ensure that nothing's been changed since the item was stored. You'll also need to have the item tagged with a digital "paper trail" of all the changes the billionaire made to the will before finally storing it. DSpace supports both bit migration—transferring bits to new formats—and attaching the requisite tags to the file about what happens to it over time to prove legal provenance.

Like most difficult challenges, data preservation is really a mix of the simple and the complex challenges. At one end of the preservation continuum is a simple item, like an ASCII text document. Preserve the data by keeping the file on current media and provide some way to view it and you're pretty much done. We'll call this the "save the bits" approach.

At the other end lie the harder cases, like these:

- A compiled software program written in a custom-built programming language for which neither the language documentation nor the compiler has survived.
- A complex geospatial data set developed for the U.S. Geological Survey in a proprietary system made by a company that went out of business 20 years ago.
- A Hollywood movie created with state-of-the-art encryption to prevent piracy, for which the decryption keys were lost.

For these three items, we don't hold out much hope of being able to preserve the content forever. For the software program and the geospatial data set, the digital archeologists of the future probably won't have enough information about how the software and data set were created or the language they were created in—no Rosetta Stone, as it were, to translate the bits from lost languages to modern ones.

As for that encrypted movie, our archeologists might have read old reviews that raved about the special effects in Sin City, but this cinematic achievement will remain locked away until someone pays a lot of money to a master of ancient cryptology to crack the key.

Fortunately, many content types fall between these difficult cases and ASCII text. Usually, saving the bits using standard, well-documented data, video, and image formats, such as XML, MPEG, and TIFF, gets you halfway to an enduring digital archive. Put another way, the goal is to avoid formats that require proprietary software, such as AutoCAD or QuarkXPress, to play or render the data.

In some cases, even files created with proprietary software might survive. Take Microsoft Word. Even though Microsoft does not guarantee backward compatibility over time—leaving it to you to resave documents into the latest version of Word—the program is so popular that we can expect such migration programs to emerge from third parties to help companies and governments salvage their valuable information assets.

Today, organizations looking to convert vast stores of Microsoft documents can turn to companies like ConverterTechnology, in North Sydney, Australia. The company uses a proprietary program to batch-process documents from one version of Microsoft Office to another and from other programs like Lotus 1-2-3 to Microsoft Excel [see sidebar, "Resurrecting Software"].

AT THE MIT LIBRARIES, our DSpace archiving process encourages contributors to submit

content in standard formats. It also automatically prompts them to provide the information to help preserve those formats. As we build the service up, everything from scholarly papers and books to lecture notes, videos, photos, simulations, and tests will flow into our collection daily.

Because an archive by its very nature grows, it needs an expandable hardware setup. While almost any enterprise information technology system can be adapted to run DSpace, at MIT we run the system on two new Hewlett-Packard ProLiant servers with Intel Xeon 2.8-gigahertz processors and a 10-terabyte storage area network consisting of forty-two 250-gigabyte hard drives, also from HP.

Hardware is an important consideration, of course, but the real heart of a DSpace archive lies in the software. DSpace is an open-source system written in Java that runs on any computer platform, but typically on top of Unix and Unix-based operating systems, such as Linux. Each DSpace archive is divided into communities, each of which generally corresponds to a laboratory, research center, or department. Communities contain collections—that is, groupings of related content. Items, such as documents, video and audio clips, and class notes, are considered the basic elements of the archive and populate each collection.

Items are further subdivided into bit streams, continuous series of bits transmitted over the Internet, which when captured and stored on a hard disk compose ordinary computer files, such as a document or video. Closely related bit streams—for example, HTML files and images that compose a single HTML document—are organized into bundles. These bundles fall into three categories: the bundle with the original deposited bit streams; thumbnails of any image bit streams; and text extracted from the original bit streams, to be used for indexing [see diagram, "Working in DSpace"].

Once an institution decides which data formats its archive will support, it starts running DSpace on its storage area network servers and users start uploading files. To ensure that people actively contribute to the archive, we made the DSpace input process simple.

Suppose a faculty member decides to deposit her latest research article, which is in Adobe's Portable Document Format (PDF), into one of MIT's digital archive communities. After connecting to the submission interface, she clicks through a series of screens that ask her for various pieces of information about the article. Some of that information will be used as metadata—data about the data—which search engines, both on the Web and in the archive, will use to find the article. The archive's curator will also use the information to help preserve the article: the faculty member's name, the article's title, the publisher, the abstract, some keywords, and so on. Toward the end, the program prompts her to upload the article.

Next, DSpace processes the file to detect its format, in this case PDF, and to verify that it really is a PDF, is virus-free, and is not encrypted. DSpace also makes sure the file doesn't use images to represent foreign characters or any other features that are legal in the PDF standard but would make future conversions of the document difficult. If the file doesn't pass the validation step, it gets kicked back to the depositor for correction.

The program also detects some of the PDF's other physical properties, such as its size in bytes, which it records as technical documentation about the file. Then the program generates a "checksum" for the file by assigning a numerical value based on the number of bits in the file. DSpace uses that value over time to verify that the article hasn't been changed unintentionally or corrupted. When DSpace has finished this series of automated checks, it asks the researcher if the information generated by the program about her file is correct and if she'd like to supply a label for it—for example, "Preprint Version."

Finally, the researcher clicks through a license that grants DSpace the right to store, preserve, and redistribute the article, and if she retained the copyright to the article, asks whether she wants to assign a Creative Commons License to it. This license gives other researchers the right, among other things, to include the article as part of their course readings or quote it in their own scholarly writings, without asking for her explicit permission.

After the depositor submits the article to DSpace, it goes into a review and approval process, or workflow. These workflows vary but usually consist of a couple of steps to verify that the submission meets the standards of the community. For instance, was the article written by a member of the department and accepted for publication? Were the supplied metadata correct? Designated community members perform these checks, and each time a workflow's status

changes—for instance, when a reviewer accepts the submission—DSpace adds a provenance statement to the metadata, allowing the curator to track how the item has changed since a user submitted it.

Upon successful completion of the workflow process, normally within a day or two of submission, the program converts the submission into a full-fledged archived item in DSpace. Among other things, it assigns a "date.available" value to the metadata record of the item, storing and indexing the metadata in a database and making the article available on the DSpace Web site. An automatically generated e-mail message notifies community subscribers of the new item's addition. A few days later, the article will start to appear in scholarly indexes and Web search engines like Google.

The task of keeping the author's article available for future generations falls to the DSpace curator—and the curator's successors in the years to come. The author's valid PDF file appears on the list of supported formats, ensuring that over the coming years the curator will be monitoring the PDF standard and the support available for it. Are there tools to read and display it? Is the standard still available in case we need to write a conversion program? Are there legal problems that might make us want to avoid keeping content in that format?

TO ENSURE THAT WE DON'T WIND UP WITH A DIGITAL TOWER OF BABEL, WE NEED TO AGREE TO USE OPEN, PUBLISHED STANDARDS, SUCH AS XML, TIFF, PDF, AND MPEG

The curator also develops a preservation strategy for each supported format, specifying the steps needed to minimize the risk of losing items. For our PDF example, the curator might ask DSpace to make a second copy of the article in Adobe PostScript and a third in plain ASCII text, using currently available software tools to do the format conversions. These new versions are then stored in the archive as backups, along with the original PDF.

Now imagine that a few years have passed, and Adobe announces that it has developed a new format, "PDG," and will no longer sell tools to process PDF documents. Two years after that, the market for tools that read or process PDFs has dried up, and all existing PDF documents are at risk of no longer being readable. The curator then runs a query in DSpace to find all the PDF files in the archive, acquires or creates a program to automatically convert PDFs into PDGs, and runs the conversion.

Both PDF and PDG versions are then stored in the archive in case someone questions the conversion and wants to see the original PDF bits. The PDG version is now the version that appears first in the DSpace Web interface for access purposes, and the researchers looking at the article never need to know that the article has been converted from one format to another—it looks exactly as it used to, thanks to DSpace and its curators.

THERE'S NO ONE RIGHT WAY to preserve digital content. Just as biodiversity is good for the natural environment, different digital preservation policies and strategies are good for the preservation environment [see sidebar, "Preservation Societies"]. But to ensure that we don't wind up with a digital Tower of Babel, we need to agree to use open, published standards, such as XML, TIFF, PDF, and MPEG.

And that's true not just for the obvious items like images, documents, and audio files, but also for scientific images, genomics data sets, and multimedia presentations and simulations. In the scientific research community, standards are emerging here and there—HDF (Hierarchical Data Format), NetCDF (network Common Data Form), FITS (Flexible Image Transport System)—but much work remains to be done to define a common cyberinfrastructure.

MIT is working closely with the University of Cambridge to develop a preservation strategy for each of the formats that the DSpace project intends to support. We plan to share these as widely as possible for peer review. As a start, we are tackling the most commonly deposited formats: PDF, HTML, and a couple of the Microsoft formats, Excel and Word. We will work our way down the list over time, in order of popularity and ease of preservation, and we'll also

publish guidelines for the MIT community about which formats should be used when possible to make archiving easier.

We hope and expect that the worldwide community of digital archivists will begin to divide and conquer so that one group by itself doesn't have to address the tens of thousands of file formats that are out there. Toward that end, efforts are under way to build new collaborative services like the proposed Global Digital Format Registry, which we can all add to and use as an authoritative source of information about digital formats and the tools for processing them.

Individuals can help, too. Document what you create, when you created it, in what format, on what computer, with what parameters, and so on. Also try to tag documents with metadata. By the time archivists get digital items, they're often unmoored from their originator, so sometimes archivists don't even know what the items are or who made them, much less whether the institution has the right to archive them.

Digital preservationists know that metadata tagging is a lot to ask of people and that we need to make doing the right thing much, much easier. Until we accomplish that goal, back up your hard disk tonight, and maybe print out your most important documents, just in case.

TO PROBE FURTHER

For more information on creating an institutional depository with DSpace, go to http://www.dspace.org.

To browse through 40 billion Web pages archived from 1996 to a few months ago, check out the Internet Archive's WayBack Machine at http://www.archive.org/web/web.php.

More than 80 institutions are participating in the LOCKSS program to preserve digital content at http://lockss.stanford.edu/index.html.

ABOUT THE AUTHOR

MACKENZIE SMITH is associate director for technology at the Massachusetts Institute of Technology Libraries, in Cambridge.

IMAGE: JONATHAN BARKAT

URL: http://www.spectrum.ieee.org (Modified: 30 June 2005)