TOOLBOX

DISKS BACK FROM THE DEAD

Getting data off an ancient floppy disk or computer tape isn't easy, but it can be done with the help of clever software and hardware.

PROJECT TWINS

BY MONYA BAKER

In 2012, William Parker and his colleagues went hunting for a data set that tracked the growth of more than 50,000 carefully planted white spruce trees over a decade and a 1,500-kilometre range. They found a reel of computer tape, a relatively modern 3.5-inch diskette and a box of older 5.25-inch diskettes. These contained data from field trials in the late 1970s, which were set up to improve yields of commercial timber. Parker, who is at the Ontario Forest Research Institute in Sault Saint Marie, Canada, needed ways to evaluate how strategies such as 'assisted migration' might preserve forests on a warming planet, and this long-term systematic study was just the thing, he says. "When we found it, it was like, 'Eureka! Hallelujah! We've finally got it!'."

Not so fast. Parker booted up an old computer, but it could not read the newest disk. No one had equipment to even try the others.

Parker's IT services referred him to a data-retrieval company. The older disks turned out to be 'flippies', double-sided disks written in formats that few drives can read. The specialists were ultimately able to read them using a carefully placed hole puncher, a bit of digital forensics and some programming that converted ancient software to modern spreadsheets.

Parker's experience encapsulates the problems that many researchers encounter. Retrieving information from defunct ▶

data-storage media is like unlocking a series of cages, says Bertram Lyons, an archivist with AVPreserve in Madison, Wisconsin. "Scientists have information trapped in older formats. Some are physical barriers, some are encoded structures. Both can go obsolete."

Scientists hoping to get data off old media first need to find a device that can read it and connect to a modern computer. But moving files to modern media is just the first step; the next is making sense of its contents, which requires another suite of tools.

GOING MODERN

When it comes to old hardware, a good place to start may be the local library. The Memory Lab at the Public Library in Washington DC offers a do-it-yourself station that allows people to transfer 3.5-inch floppies onto modern formats, for instance, and Stanford University Libraries offers a similar resource for 5.25-inch disks. Gavan McCarthy, director of the University of Melbourne's eScholarship Research Centre in Australia, has what he calls "the Museum of Redundant Technology", which can handle a range of formats. "If you have the tape, the disk and whatever it can fit into, we've got the plugs," he says.

For a few dollars per disk, conversion service firms, such as FloppyDisk in Lake Forest, California, and RetroFloppy in Cary, North Carolina, can help. So, too, can data-recovery services, which specialize in damaged media. DriveSavers, a data-recovery firm based in Novato, California, has around 20,000 storage devices, the oldest being a Shugart ST-506 hard-disk drive from 1980. Parker used CBL Data Recovery in Toronto, Ontario, which subcontracted with Muller Media Services (now George Blood Audio in Manhasset, New York), to recover his data and paid about US$3,000.

Success depends on the fragility of the media and how it was stored. 5.25-inch disks are easily damaged by oils and pressure, and Iomega Zip disks are unstable. But it's not just 'bitrot', or damage to media themselves, that makes old media unreadable says McCarthy. "The number of machines and the spare parts are falling off incredibly rapidly." Paper is, ironically, more stable.

People who have the old drives and power cables may be tempted to set up their own do-it-yourself stations, only to find that new computers no longer contain the boards and interfaces required to make the connection. Some old Zip drives, for instance, plugged into a 'parallel' (printer) port — an interface that has largely disappeared today. But there are a range of adapters, mainly used by archivists and video-game enthusiasts, that can help. At the top end is the KryoFlux device, developed by the Software Preservation Society, which can transfer floppy-disk data through a USB interface. The KryoFlux Preservation Technology Group in Maidstone, UK, charges private users about $100 for the hardware.

The operating systems on modern computers may also be unable to read files in old formats. Lori Emerson, director of the Media Archaeology Lab at the University of Colorado, Boulder, says that helping a local science museum to recover a mysterious file on a Zip disk depended on finding the right computer (a Power Macintosh 8100 from 1994 running OS 7) to read the file, which turned out to be a library from an old version of the citation manager EndNote.

Guido Pauli, a medicinal chemist at the University of Illinois in Chicago, suggests that the best defence against data decay is staying current. Pauli maintains the NAPRALERT database, which lets researchers search for natural products (such as botanical extracts) and reported biological activities. It began on index cards organized by Guido's PhD adviser, and has since moved through magnetic tape and various disk formats, and is now in the cloud and on hard drives on two continents. "I do have some of the old media, but I'm not dependent on reading them," Pauli says.

DIVING INTO DATA

The next challenge in recovering old data is making sense of the data files themselves. For digital archivists, the first step in preservation is capturing a disk image — a bit-for-bit copy of all the digital data on a device, including overwritten and hidden files. That's the remit of digital forensics technologies, but commercial licences for such tools can cost thousands of dollars. Plus, with their focus on legal applications, they neglect certain functions important to archivists, such as redacting sensitive information.

"The number of machines and the spare parts are falling off incredibly rapidly."

This led archivists to create BitCurator, an open-source 'virtual machine' that images a disk and guides people through the first steps in interpreting its contents, such as detecting how those bits and bytes are formatted into files readable by, for instance, the Windows NT operating system, Linux or DOS. The more obscure the format, the harder this is.

Chris Muller, who founded Muller Media, has written software to unlock ancient files, but human clues can sometimes be more valuable, he says. Muller asks clients to e-mail a photograph of the original media early on in a potential project. Sometimes a few squiggles from a Sharpie that are meaningless to his client are letters or digits that let Muller deduce what formatting and software might have been used.

The next step is accessing the files, explains Christopher Lee at the School of Information and Library Science at the University of North Carolina at Chapel Hill, and one of the main forces behind BitCurator. The files might be in an unrecognizable format, making it hard to know what program might open it, he says. "The software often is the barrier." Researchers can use computer programs known as Hex editors to show the raw binary content of such files. With luck, this might reveal what software a file was written in, or allow usable data to be extracted directly. BitCurator also interacts with the US National Institute of Standards and Technology's Software Reference Library (www.nsrl.nist.gov) to try to match files with the software that created them.

With a few clues, researchers can often identify modern programs to open files from similar, older software and convert them to newer formats. An alternative, assuming the original software is available, is emulation: recreating the older operating-system platform within a modern machine. The Internet Archive, for example, has emulators for platforms such as MS-DOS that can run over an Internet browser. Emulation is more cost-effective when software is set up for highly specific tasks or visual renderings and cannot be easily migrated to contemporary formats, says Klaus Rechert of the Institute for Computer Science at the University of Freiburg in Germany. He recently set up an emulator to recreate analyses from a natural-language study that had produced custom language maps in the typesetting program LaTeX.

Another option is 'digital archaeology', writing software to make old files intelligible. But that route is expensive, often futile, and usually requires a reasonable idea of what the file contains. In one relatively simple example, David Schmidt at RetroFloppy looked for sets of repeating codes that corresponded to letters in a client's name to craft a conversion matrix and recover data from an obscure IBM system stored on an 8-inch floppy. Firms such as George Blood and AVPreserve specialize in more-complex versions of these problems.

The biggest hurdle is sometimes not technological but human, digital archivists say. It's not enough to extract a file just to learn that it has 6 columns and 100,000 rows; researchers need to know what the numbers mean. Archivists led by Amy Pienta at the Interuniversity Consortium for Political and Social Research in Ann Arbor, Michigan, for instance, bought a refurbished punch-card reader to retrieve data from a large, longitudinal study of retirement from the 1950s. But after physical punches were converted to ASCII numeric codes, they needed preserved codebooks to know what the numbers referred to — did a code of '1' mean yes or no?

Parker's story has an interesting coda: the digital data contained only averaged values for groups of trees, but a lucky phone call revealed that paper records with measurements of individual trees had been kept. He took a few hours' drive to meet the original scientist and collect the data sheets.

Says Melbourne's McCarthy, "If you want to preserve something, you've got to move while those people are still around." ■