TOOLBOX

DIGITAL FORENSICS IN THE LIBRARY

Archivists are borrowing and adapting techniques used in criminal investigations to access data and files created in now-obsolete systems.

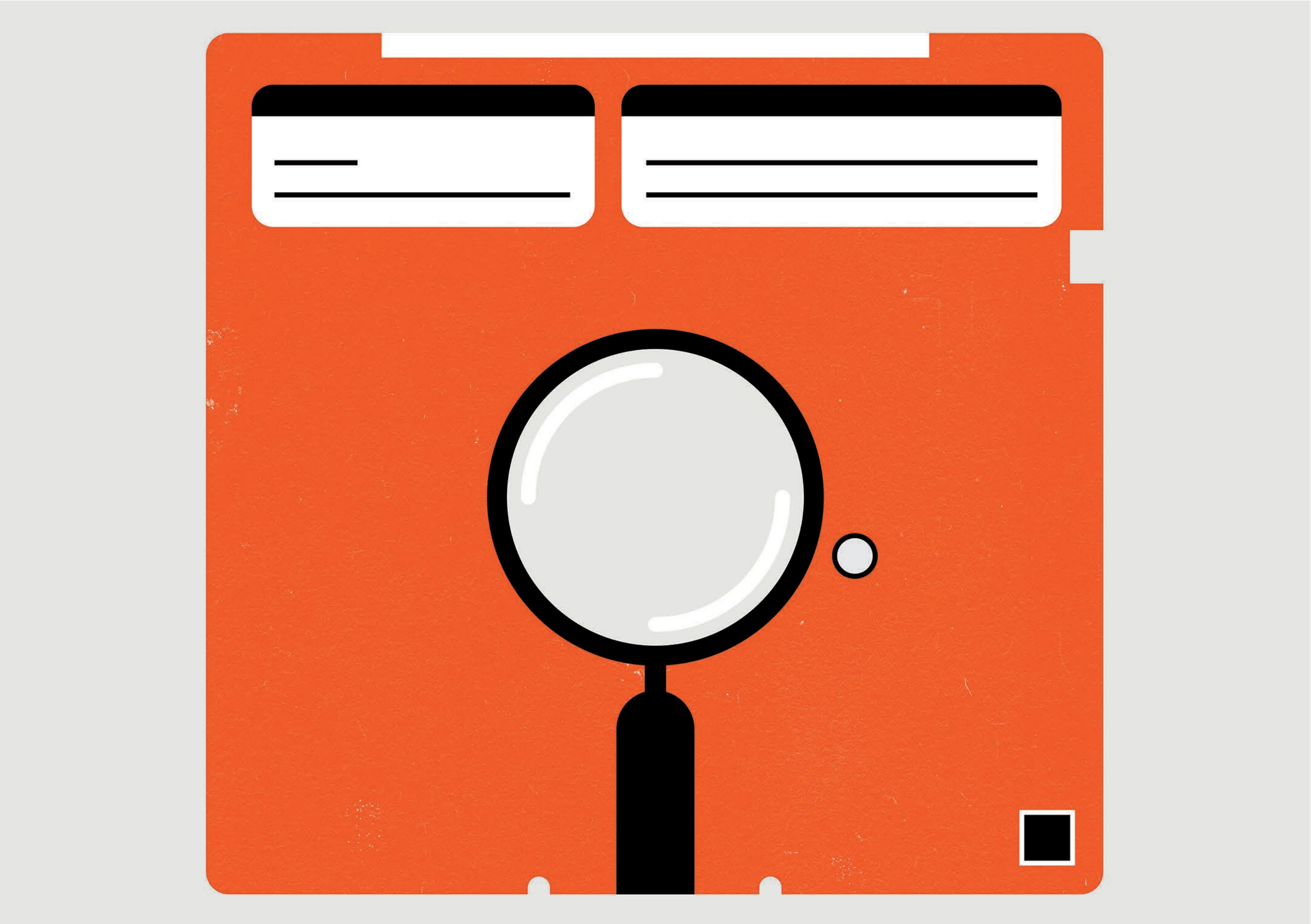

ILLUSTRATION BY THE PROJECT TWINS

BY MARK WOLVERTON

When archivists at California's Stanford University received the collected papers of the late palaeontologist Stephen Jay Gould in 2004, they knew right away they had a problem. Many of the 'papers' were actually on computer disks of various kinds, in the form of 52 megabytes of data spread across more than 1,100 files — all from long-outdated systems.

"It was a large collection, as you can imagine," says Michael Olson, service manager for the Born Digital/Forensics Lab at Stanford University Libraries. "He used a lot of early word processing for his writing, lots of disks and diskettes in different formats."

After considerable effort the Stanford archivists did get Gould's papers into order — first by finding hardware that could read the obsolete disks, and then by deciphering what they found there. "We had some challenges finding old applications to figure out what word processor he used, that sort of thing," says Olson.

The Gould papers were an early indication of an issue that's been rapidly worsening: four decades after the personal-computer revolution brought word processing and number crunching to the desktop, the first generation of early adopters is retiring or dying. So how do archivists recover and preserve what's left behind?

"People around the world have information stored on disks that are less readable with every passing day," says Christopher ▶

Lee, a researcher in the School of Information and Library Science at the University of North Carolina (UNC) in Chapel Hill. "This includes floppies, Zip disks, CDs, DVDs, flash drives, hard drives and a variety of other media." Many files can be accessed only with long-obsolete hardware, and all are subject to physical deterioration that will ultimately make them unreadable by any means. By now, many libraries, archives and museums have accumulated shelves full of such material, stashed away in the hope that if it's ever needed, somebody, somewhere will be able to figure out how to access it.

DIGITAL INSPIRATION

Increasingly, archivists are finding inspiration in the field of digital forensics: the art of extracting evidence about illicit activity from computer drives, smartphones, tablets or even GPS devices. "It turned out that law-enforcement and computer-security people were dealing with essentially the same problems of stabilizing and recovering data from digital media," says Matthew Kirschenbaum at the University of Maryland in College Park. And many of their solutions were directly applicable to the archivists' needs.

In law enforcement, for example, a top priority is to preserve material in its original form. This is often harder than it sounds: almost anything done on a computer, even something as innocuous as plugging in a USB drive, leaves a faint digital trace. So digital-forensics practitioners have developed techniques for creating an artefact-free 'disk image' that duplicates everything, down to the unused and hidden disk space. They can then preserve the integrity of the original for evidentiary purposes in court while doing all their forensic analysis on a perfect copy.

Institutions working to decipher collections have the same need, although in their case, the object is to maintain the provenance of the original for future researchers. Creating forensic copies of the data was a relatively fringe idea 8 or 10 years ago, Lee says. "It's now quite common in library and archive settings."

Unfortunately for archivists, however, disk imaging is usually done through commercial software packages such as the Forensic Toolkit made by Access Data in Lindon, Utah, or by EnCase, which is developed by Guidance Software in Pasadena, California. Because these packages are designed for criminal investigators, they include tools for file carving (assembling complete files from fragmentary data); cracking passwords; accessing encrypted files; advanced searching; and generating reports for use in court — tasks that tend to be less important for archival purposes. These packages also come with licensing costs in the thousands of dollars, which would strain the budget of many collecting institutions.

So in 2011, Lee and his colleagues launched BitCurator, a platform designed for the archival field, with funding from the Andrew W. Mellon Foundation, and with continued support from a consortium that currently encompasses 25 member institutions, including Harvard University, the Massachusetts Institute of Technology, Stanford University, Emory University and the British Library. BitCurator has the advantage of being open source and freely available for download (wiki.bitcurator.net). "It's a combination of third party open-source tools and our own work," says Kam Woods, a research scientist at UNC's School of Information and Library Science and co-principal investigator with Lee on the project. On the basis of the turnout at training sessions and other BitCurator events, Lee estimates that several dozen institutions now use the package actively, and several hundred more use it at least occasionally.

> ***"People around the world have information stored on disks that are less readable with every passing day."***

BitCurator not only handles disk imaging, but a number of other issues that criminal investigators don't have to worry about. One example is redaction: editing out confidential material before publication. That's an alien concept in the criminal investigations, says Olson. "Why would you ever want to redact evidence from a case? But from an archival or library standpoint, you wouldn't want to make somebody's health records available." So BitCurator has to have methods for access control that don't really exist in the forensics field.

Another speciality of BitCurator is its ability to read long-outdated disks — an essential tool for archivists who are faced with stacks of old floppies or even reels of magnetic tape. Although digital-forensics investigators usually deal with newer-generation systems, their techniques can still be quite useful for recovery, says Lee. "Taking a forensic approach, you can still create a safe copy of the data, even if you don't know what the file system is or you can't read it," he says. "As long as you can attach a drive and get the bits off of it, you can create an image." Archivists can then experiment on different ways to retrieve the files, safe in the knowledge that the original is not in danger.

Some advantages to the forensics-based approach transcend technical considerations, says Olson. With the Gould archives, for example, "you can get timestamps from different word-processing files to see how he actually wrote something, a particular order that he wrote, a way that he edited. That's really nifty if you're a researcher that wants to know how his mind worked."

SEARCH AND RESCUE

The same techniques can be used for other purposes besides archiving. At Stanford, Olson's lab is increasingly helping faculty members and students who need to access work that was born on now-outdated computer systems. "I had a graduate student about a year ago that came to us with an astrophysics data set on a Zip disk," he says. "It was something that their professor had created, that they weren't able to read and needed to get to because it was part of their research. And nobody had really shepherded that to a new modern system." The library was able to help the student do just that.

Another recent example is Stanford's long-running ME310 engineering course, which had a server full of design studies, presentation slides and videos that students had completed over the years as part of their graduate work. "The people running the programme wanted to preserve all the data from these projects," says Olson, "but they needed help to recover the data, organize it and also get permission from the students to actually make this available."

Data are already being lost to science at a rapid rate. One study, for example, found that as little as 20% of data for ecology papers published in the early 1990s is still available (T. H. Vines *et al. Curr. Biol.* **6,** 94–97; 2014). Co-author Tim Vines, who now runs a peer-review service called Axios Review in Vancouver, Canada, says that the best way for scientists to preserve their data for future generations is to upload it into library-maintained archives or open online repositories, such as Dryad or Figshare.

"Putting it into the hands of an organization committed to preserving it is far better than putting it on a shelf", he says. ■

MORE ONLINE

STAY CURRENT

- Scientists write their papers online — together go.nature.com/mcw1ow
- Researchers move their reference libraries to the clouds go.nature.com/i5lhwp
- Computer scientists clean up 'link rot' go.nature.com/b9agxg

IN THE NEWS

Librarians and researchers are racing to cope with a flood of open data go.nature.com/r5k6tw

WILL KIRK/JHU HOMEWOOD PHOTOGRAPHY