Digital Preservation: An Unsolved Problem

by Jonathan Shaw

GIVEN THE CONVENIENCE and potential cost saving of digital delivery for both libraries and users, combined with the power digitization offers to search within texts, why not embrace the digital future now? The issue of preservation is one of the main obstacles.

Speaking from her experience as head of collection care for the British Library, Helen Shenton explains that "the greatest risks to printed material are the environment, wear and tear, security, and custodial neglect." Facilities such as the Harvard Depository address most of those concerns, although wear and tear is an unavoidable consequence of use. On the digital side, on the other hand, *use* of the data is one of the best ways of preserving it, because "'bit rot' is one of the biggest risks." A book left on the shelf for a hundred years might be fine, Shenton says, but digital data must be read and checked constantly to ensure their integrity.

For digital preservationists, a prime concern is that data might be kept perfectly secure and complete, but *still* be unreadable by machines and programs in the future. A *New Yorker* cover depicting an alien, come to post-apocalyptic Earth, sitting amid the detritus of modern civilization—discarded CDs, tapes, and computers—illustrates the point: the alien is reading a book, the only thing that still "works." "You have to think about moving the content along as technology changes," explains Andrea Goethals, digital-preservation and repository-services manager. In order to make this feasible, librarians try to limit the number of file formats they make use of, and store detailed technical metadata with every object so that in 10 years or 100 "it can be rendered again in a usable way." Every few years, as the programs that created a text file or a PDF become obsolete, librarians must ensure that the contents of those files remain readable by the current generation of computers and software. But opening each file manually in order to save it in a current format is not feasible when there are millions of them. "Because of the enormous amount of digital material we hold, migrating content is done at scale, not one file at a time," Goethals explains. "We have to be able to do it on a whole class of objects"—all Microsoft Word files, for example. This content management strategy *should* work, "but because digital preservation is a young science, we don't have a lot of experience with it yet."

Objects that begin in an analog format—a book, a recording of a poetry reading, or a piece of music—are easiest to preserve digitally because librarians can choose the optimal file format for long-term access. Material that is born digital—e-mail, for example, which comes in many different, often proprietary, formats—is not so easy to preserve. This is proving a vexing problem for librarians charged with maintaining records of the University's intellectual life. Harvard archivists, for example, must figure out what to do with the president's e-mail—both a technical and *legal* challenge, because of privacy laws governing its handling.

For medical librarians like Isaac Kohane, the lack of case notes from the last 20 years is an equally grave concern. He recalls his excitement, shortly after becoming director of the Countway, at being able to read the notebooks of a surgeon who was about to conduct the first successful cadaveric kidney transplant, for which he later won the Nobel Prize. "You could see the foresight and the hypothesis he had, as he made that leap. So it is stunning to realize that we are in a period of unprecedented lack of documentation of academic output." Most notes, images, and communications today are stored as e-mail. "I'm worried that, 30 years hence, people will say, 'Well, what was happening during the great genetics revolution at Harvard University?' They'll have no idea."