PRESERVING THE INTERNET

An archive of the Internet may prove to be a vital record for historians, businesses and governments

by Brewster Kahle

Manuscripts from the library of Alexandria in ancient Egypt disappeared in a fire. The early printed books decayed into unrecognizable shreds. Many of the oldest cinematic films were recycled for their silver content. Unfortunately, history may repeat itself in the evolution of the Internet—and its World Wide Web.

No one has tried to capture a comprehensive record of the text and images contained in the documents that appear on the Web. The history of print and film is a story of loss and partial reconstruction. But this scenario need not be repeated for the Web, which has increasingly evolved into a storehouse of valuable scientific, cultural and historical information.

The dropping costs of digital storage mean that a permanent record of the Web and the rest of the Internet can be preserved by a small group of technical professionals equipped with a modest complement of computer workstations and data storage devices. A year ago I and a few others set out to realize this vision as part of a venture known as the Internet Archive.

By the time this article is published, we will have taken a snapshot of all parts of the Web freely and technically accessible to us. This collection of data will measure perhaps as much as two trillion bytes (two terabytes) of data, ranging from text to video to audio recording. In comparison, the Library of Congress contains about 20 terabytes of text information. In the coming months, our computers and storage media will make records of other areas of the Internet, including the Gopher information system and the Usenet bulletin boards. The material gathered so far has already proved a useful resource to historians. In the future, it may provide the raw material for a carefully indexed, searchable library.

The logistics of taking a snapshot of the Web are relatively simple. Our Internet Archive operates with a staff of 10 people from offices located in a converted military base—the Presidio—in downtown San Francisco; it also runs an information-gathering computer in the San Diego Supercomputer Center at the University of California at San Diego.

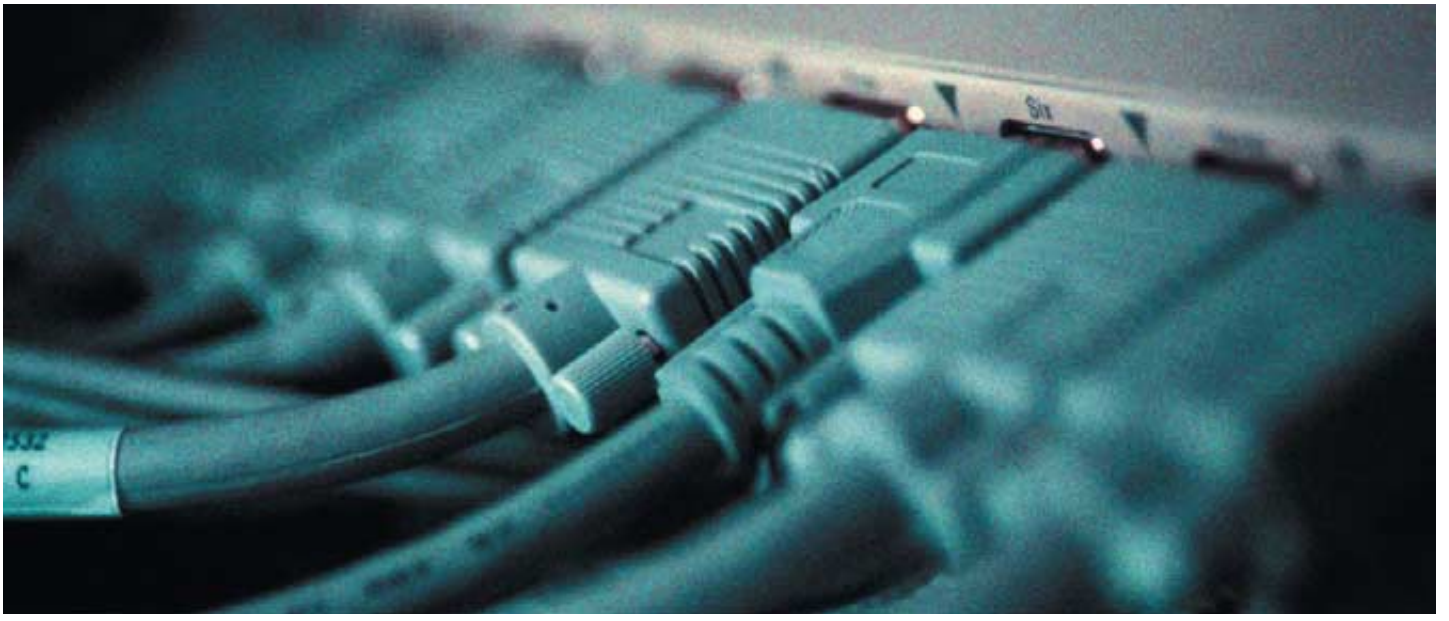
The software on our computers "crawls" the Net—downloading documents, called pages, from one site after another. Once a page is captured, the software looks for cross references, or links, to other pages. It uses the Web's hyperlinks—addresses embedded within a document page—to move to other pages. The software then makes copies again and seeks additional links contained in the new pages. The crawler avoids downloading duplicate copies of pages by checking the identification names, called uniform resource locators (URLs), against a database. Programs such as Digital Equipment Corporation's AltaVista also employ crawler software for indexing Web sites.

What makes this experiment possible is the dropping cost of data storage. The price of a gigabyte (a billion bytes) of hard-disk space is $200, whereas tape storage using an automated mounting device costs $20 a gigabyte. We chose hard-disk storage for a small amount of data that users of the archive are likely to access frequently and a robotic device that mounts and reads tapes automatically for less used information. A disk drive accesses data in an average of 15 milliseconds, whereas tapes require four minutes. Frequently accessed information might be historical documents or a set of URLs no longer in use.

We plan to update the information gathered at least every few months. The first full record required nearly a year to compile. In future passes through the Web, we will be able to update only the information that has changed since our last perusal.

The text, graphics, audio clips and other data collected from the Web will never be comprehensive, because the crawler software cannot gain access to many of the hundreds of thousands of sites. Publishers restrict access to data or store documents in a format inaccessible to simple crawler programs. Still, the archive gives a feel of what the Web looks like during a given period of time even though it does not constitute a full record.

After gathering and storing the public contents of the Internet, what services will the archive provide? We possess the capability of supplying documents that are no longer available from the origi-



nal publisher, an important function if the Web's hypertext system is to become a medium for scholarly publishing. Such a service could also prove worthwhile for business research. And the archival data might serve as a "copy of record" for the government or other institutions with publicly available documents. So, over time, the archive would come to resemble a digital library.

Keeping Missing Links

Historians have already found the material useful. David Allison of the Smithsonian Institution has tapped into the archive for a presidential election Web site exhibit at the museum, a project he compares to saving videotapes of early television campaign advertisements. Many of the links for these Web sites, such as those for Texas Senator Phil Gramm's campaign, have already disappeared from the Internet.

Creating an archive touches on an array of issues, from privacy to copyright. What if a college student created a Web page that had pictures of her then current boyfriend? What if she later wanted to "tear them up," so to speak, yet they lived on in the archive? Should she have the right to remove them? In contrast, should a public figure—a U.S. senator, for instance—be able to erase data posted from his or her college years? Does collecting information made available to the public violate the "fair use" provisions of the copyright law? The issues are not easily resolved.

To address these worries, we let authors exclude their works from the archive. We are also considering allowing researchers to obtain broad censuses of the archive data instead of individual documents—one could count the total number of references to pachyderms on the Web, for instance, but not look at a specific elephant home page. These measures, we hope, will suffice to allay immediate concerns about privacy and intellectual-property rights. Over time, the issues addressed in setting up the Internet Archive might help resolve the larger policy debates on intellectual property and privacy by testing concepts such as fair use on the Internet.

The Internet Archive complements other projects intended to ensure the longevity of information on the Internet. The Commission on Preservation and Access in Washington, D.C., researches how to ensure that data are not lost as the standard formats for digital storage media change over the years. In another effort, the Internet Engineering Task Force and other groups have labored on technical standards that give a unique identification name to digital documents. These uniform resource names (URNs), as they are called, could supplement the URLs that currently access Web documents. Giving a document a URN attempts to ensure that it can be traced after a link disappears, because estimates put the average lifetime for a URL at 44 days. The URN would be able to locate other URLs that still provided access to the desired documents.

Other, more limited attempts to archive parts of the Internet have also begun. DejaNews keeps a record of messages on the Usenet bulletin boards, and InReference archives Internet mailing lists. Both support themselves with revenue from advertisers, a possible funding source for the Internet Archive as well. Until now, I have funded the project with money I received from the sale of an Internet software and services company. Major computer companies have also donated equipment.

It will take many years before an infrastructure that assures Internet preservation becomes well established—and for questions involving intellectual-property issues to resolve themselves. For our part, we feel that it is important to proceed with the collection of the archival material because it can never be recovered in the future. And the opportunity to capture a record of the birth of a new medium will then be lost.

The Author

BREWSTER KAHLE founded the Internet Archive in April 1996. He invented the Wide Area Information Servers (WAIS) system in 1989 and started a company, WAIS, Inc., in 1992 to commercialize this Internet publishing software. The company helped to bring commercial and government agencies onto the Internet by selling publishing tools and production services. Kahle also served as a principal designer of the Connection Machine, a supercomputer produced by Thinking Machines. He received a bachelor's degree from the Massachusetts Institute of Technology in 1982.