Archiving the Internet

Brewster Kahle
Internet Archive
11/4/96

Bold efforts to record the entire Internet are expected to lead to new services.
Submitted to Scientific American for March 1997 Issue

The early manuscripts at the Library of Alexandria were burned, much of early printing was not saved, and many early films were recycled for their silver content. While the Internet's World Wide Web is unprecedented in spreading the popular voice of millions that would never have been published before, no one recorded these documents and images from 1 year ago. The history of early materials of each medium is one of loss and eventual partial reconstruction through fragments. A group of entrepreneurs and engineers have determined to not let this happen to the early Internet.

Even though the documents on the Internet are the easy documents to collect and archive, the average lifetime of a document is 75 days and then it is gone. While the changing nature of the Internet brings a freshness and vitality, it also creates problems for historians and users alike. A visiting professor at MIT, Carl Malamud, wanted to write a book citing some documents that were only available on the Internet's World Wide Web system, but was concerned that future readers would get a familiar error message "404 Document not found" by the time the book was published. He asked if the Internet was "too unreliable" for scholarly citation.

Where libraries serve this role for books and periodicals that are no longer sold or easily accessible, no such equivalent yet exists for digital information. With the rise of the importance of digital information to the running of our society and culture, accompanied by the drop in costs for digital storage and access, these new digital libraries will soon take shape.

The Internet Archive is such a new organization that is collecting the public materials on the Internet to construct a digital library. The first step is to preserve the contents of this new medium. This collection will include all publicly accessible World Wide Web pages, the Gopher hierarchy, the Netnews bulletin board system, and downloadable software.

If the example of paper libraries is a guide, this new resource will offer insights into human endeavor and lead to the creation of new services. Never before has this rich a cultural artifact been so easily available for research. Where historians have scattered club newsletters and fliers, physical diaries and letters, from past epochs, the World Wide Web offers a substantial

collection that is easy to gather, store, and sift through when compared to its paper antecedents. Furthermore, as the Internet becomes a serious publishing system, then these archives and similar ones will also be available to serve documents that are no longer "in print".

Apart from historical and scholarly research uses, these digital archives might be able to help with some common infrastructure complaints:

- Internet seems unreliable: "Document not found"
- Information lacks context: "Where am I? Can I trust this information?"
- Navigation: "Where should I go next?"

When working with books, libraries help with some of these issues, with "the stacks" of books, links to other libraries and librarians to help patrons.

Preservation of our Digital History

Where we can read the 400 year-old books printed by Gutenberg, it is often difficult to read a 15 year-old computer disk. The Commission for Preservation and Access in Washington DC has been researching the thorny problems faced trying to ensure the usability of the digital data over a period of decades. Where the Internet Archive will move the data to new media and new operating systems every 10 years, this only addresses part of the problem of preservation.

Using the saved files in the future may require conversion to new file formats. Text, images, audio, and video are undergoing changes at different rates. Since the World Wide Web currently has most of its textual and image content in only a few formats, we hope that it will be worth translating in the future, whereas we expect that the short lived or seldom used formats not be worth the future investment. Saving the software to read discarded formats often poses problems of preserving or simulating the machines that they ran on.

The physical security of the data must also be considered. Natural and political forces can destroy the data collected. Political ideologies change over time making what was once legal becomes illegal. We are looking for partners in other geographic and national locations to provide a robust archive system over time. To give some level of security from commercial forces that might want exclusive access to this archive, the data is donated to a special non-profit trust for long-term care taking. This non-profit organization is endowed with enough money to perform the necessary maintenance on the storage media over the years.

Packaging enough meta-data (information about the information) is necessary to inform future users. Since we do not know what future researchers will be interested in, we are documenting the methods of collection and attempt to be complete in those collections. As researchers start to use these data, the methods and data recorded can be refined.

Technical Issues of Gathering Data

Building the Internet Archive involves gathering, storing, and serving the terabytes of information that at some point were publicly accessible on the Internet.

Gathering these distributed files requires computers to constantly probe the servers looking for new or updated files. The Internet has several different subsystems to make information available such as the World Wide Web (WWW), File Transfer Protocol (FTP), Gopher, and Netnews. New systems for three-dimensional environments, chat facilities, and distributed software require new efforts to gather these files. Each of these systems requires special programs to probe and download appropriate files. Estimating the current size, turnover, and growth of the public Internet has proven tricky because of the dynamic nature of the systems being probed.

Protocol Number of Sites Total Data Change rate

WWW 400,000 1,500GB 600GB/month

Gopher 5,000 100GB declining (from Veronica Index)

FTP 10,000 5,000GB not known

Netnews 20,000 discussions 240GB 16GB/month

The World Wide Web is vast, growing rapidly, and filled with transient information. Estimated at 50 million pages with the average page online for only 75 days, the turnover is considerable. Furthermore, the number of pages is reported to be doubling every year. Using the average web page size of 30 kilobytes (including graphics) brings the current size of the Web to 1.5 terabytes (or million megabytes).

To gather the World Wide Web requires computers specifically programmed to "crawl" the net by downloading a web page, then finding the links to graphics and other pages on it, and then downloading those and continuing the process. This is the technique that the search engines, such as Altavista, use to create their indices to the World Wide Web. The Internet Archive currently holds 600GB of information of all types. In 1997 we will have collected a snapshot of the documents and images.

The information collected by these "crawlers" is not, unfortunately, all the information that can be seen on the Internet. Much of the data is restricted by the publisher, or stored in databases that are accessible through the World Wide Web but are not available to the simple crawlers. Other documents might have been inappropriate to collect in the first place, so authors can mark files or sites to indicate that crawlers are not welcome. Thus the collected Web will be able to give a feel of what the web looked like at a particular time, but will not simulate the full online environment.

While the current sizes are large, the Internet is continuing to grow rapidly. When it is common to connect one’s home camcorder to the upcoming high bandwidth Internet, it will not be practical to archive it all. At some point we will have to become more select what data will be of the most value in the future, but currently we can be afford to gather it all.

Storing Terabytes of Data Cost Effectively

Crucial to archiving the Internet, and digital libraries in general, is the cost effective storage of terabytes of data while still allowing timely access. Since the costs of storage has been dropping rapidly, the archiving cost is dropping. The flip side, of course, is that people are making more information available.

To stay ahead of this onslaught of text, images, and soon video information we believe we have to store the information for much less money than the original producers paid for their storage. It would be impractical to spend as much on our storage as everyone else combined.

Storage Technologies Cost per GigaByte Random access time

Memory (RAM) $12,000/GB 70nanoSeconds

Hard Disk $200/GB 15miliSeconds

Optical Disk Jukebox $140/GB 10seconds

Tape Jukebox $20/GB 4minutes

Tapes on shelf $2/GB human assistance required

(1 GigaByte = 1000 MegaBytes, 1TeraByte = 1000GigaBytes. A GigaByte is roughly enough to store 1000 books or 1 hour of compressed video)

With these prices, we chose hard disk storage for a small amount of the frequently accessed data combined with tape jukeboxes. In most applications we expect a small amount of information to be accessed much more frequently than the rest, leveraging the use of the faster disk technology rather than the tape jukebox.

Providing Access and New Services

After gathering and storing the public contents of the Internet, what services would then be of greatest value with such a repository? While it is impossible to be certain, digital versions of paper services might prove useful.

For instance, we can provide a "reliability service" for documents that are no longer available from the original publisher. This is similar to one of the roles of a library. In this way, one document can refer, through a hypertext link, to a document on another server and a reader will be able to follow that link even if the original is gone. We see this as an important piece of infrastructure if the global hypertext system is to become a medium for scholarly publishing.

Another application for a central archive would be to store an "official copy of record" of public information. These records are often of legal interest, helping to determine what was said or known at a particular time.

Historians have already found the material useful. David Allison of the Smithsonian Institution has used the materials for an exhibit on Presidential Election websites, which he thinks might be the equivalent to saving videotapes of early TV campaign advertisements. David Eddy Spicer of Harvard's Kennedy School of Government has used the materials for their "case studies" in much the same way they collect old newspapers articles to capture a point in time.

With copies of the Internet over time and cross correlation of data from multiple sources, new services might help users understand what they are reading, when it was created, and what other people thought of it. With these services, people might be able to give a context to the information they are seeing and therefore know if they can trust it. Furthermore, the coordination of this meta-information and usage data can help build services for navigating the sea of data that is available.

Companies are also interested in saving similar information and building similar services based on their internal information to help employees effectively learn from the experiences of others.

The technologies and the services that will grow out of building digital archives and digital libraries could lead towards building a reliable system of information interchange based on electrons rather than paper. Using the "library" might be done many times a day to use documents that are no longer available on the Internet.

Legal and Social Issues

Creating an archive of informal and personal information has many difficult legal and social issues even if the material was intended to be publicly accessible at some point. Such a collection treads into the murky area intellectual property in the digital era. What can be done with the digital works that are collected gets into the area of copyright, privacy, import/export restrictions, and possession of stolen property.

To give a few examples: what if a college student made a web page that had pictures of her then-current boyfriend, but later wanted to take it down and "tear it up", yet it lived on in digital archives (whether accessible or not). Should she have the right to remove that document? Should a candidate for political office be able to go back 15 years to erase his postings to public bulletin boards that have been saved in the Archive? What if a software program that is legal to publish in Denmark, but illegal in the United States is collected by an archive: should this program be removed and hidden even from historians and scholars? The legal and social issues raised by the construction of the Archive are not easily resolved.

By allowing authors to exclude their information from the Archive we hope to avoid some of the immediate issues, and allow enough time to pass to understand the larger issues at hand.

The Internet Archive might be able to help resolve some of these issues by publicly drawing the issues out and by participating in the debates. While many of these questions will take years to resolve, we feel it is important to proceed with the collection of the material since it can never be recovered in the future.

Where does it go from here?

The new technologies and services currently being created might be useful in all digital libraries and help make the Internet more robust and useful.

Through an archive of what millions of people are interested in making public, we might be able to detect new trends and patterns. Since these materials are in computer readable form, searching them, analyzing them, and distributing them has never been easier. A variety of services built on top of large data sets will allow us to connect people and ideas in new ways.

For instance, Firefly Inc. is using the individual tastes in music and movies to help suggest other CD's and videos based on finding "similar" people. They have even found that people are interested in communicating with the other "similar" people directly thus forming communities based on similar interests. This kind of computer matchmaking which is based on detailed portraits of people's preferences suggests similar services based on reading habits.

Trends in academic fields might be able to be detected more easily by studying gross statistics of the communications in the field. The hypertext links of the World Wide Web form an informal citation system similar to the footnote system already in use. Studying the topography of these links and their evolution might provide insights into what any given community thought was important.

If archiving cultural and personal histories become useful commercially, then the efforts can be expanded to record radio and video broadcasts. These systems might allow us to study these effects and influences on our lives.

Current terabyte technologies (storage hardware and management software) are relatively rare and specialized because of their costs, but as the costs drop we might see new applications that have traditionally used non-computer media. For instance,

- A video store holds about 5,000 video titles, or about 7 terabytes of compressed data.
- A music radio station holds about 10,000 LP's and CD's or about 5 terabytes of uncompressed data.
- The Library of Congress contain about 20 million volumes, or about 20 terabytes text if typed into a computer.
- A semester of classroom lectures of a small college is about 18 terabytes of compressed data.

Therefore the continued reduction in price of data storage, and also data transmission, could lead to interesting applications as all the text of a library, music of a radio station, and video of a video store become cost effective to store and later transmitted in digital form.

In the end, our goal is to help people answer hard questions. Not "what is my bank balance?", or "where can I buy the cheapest shoes", or "where is my friend Bill?" - these will be answered by smaller commercial services. Rather, answer the hard questions like: "Should I go back to

graduate school?" or "How should I raise my children?" or "What book should I read next?". Questions such as these can be informed by the experiences of others. Can machines and digital libraries really help in answering such questions? In the long term, we believe yes, but perhaps in new ways which would have importance in education and day-to-day life.

Further Reading:

Preserving Digital Objects: Recurrent Needs and Challenges, December 1995 presentation at 2nd NPO conference on Multimedia Preservation, Brisbane, Australia.

The Vanished Library, Luciano Canfora. University of Berkeley Press, 1990.

Biography:

Brewster Kahle is a founder of the Internet Archive in April 1996. Before that, he was the inventor of the Wide Area Information Servers (WAIS) system in 1989 and founded WAIS Inc in 1992. WAIS helped bring commercial and government agencies onto the Internet by selling Internet publishing tools and production services to companies such as Encyclopaedia Britannica, New York Times, and the Government Printing Office.

Schooled at MIT (BSEE ’82), Brewster designed super computers in the 80's at Thinking Machines Corporation.

http://www.archive.org/sciam_article.html