The Cyberarchive: A Look at the Storage and Preservation of Web Sites

Carol Casey

Although librarians recognize the Internet as a resource for knowledge and information, they have yet to make a formal effort to collect and preserve the Web sites found there. This paper addresses not only the need to set up a cyberarchive but also some of the issues involved. With Web sites appearing and disappearing constantly from the Internet, the time to save them is now—before we lose a precious thread in our cultural and intellectual history.

The well-researched, neatly presented Web site that perfectly pulls together information in a manner not found on any other Web site or in a print source is a gem that librarians pounce on and bookmark, and then guard with their lives. Based on content alone, this site is a solid entity in the unstable chaos otherwise known as the Internet. Suddenly, the Web site is gone—with no forwarding address. Unable to comprehend how such a treasure of information could be removed without even a warning, frantic checks are made to see if it still exists, perhaps disguised with a new title. Although it is human nature to be a little slow on the uptake when it comes to preserving our contributions to the world record, we always are shocked when we actually witness this lack of understanding of the importance of our work and the impact it has on others. We lament our ancestors' lack of insight for not preserving things they should have recognized as important, and yet we suffer from this same affliction.

Given the number of broken links and search engine entries that lead to the vacant lots of the Internet, one gets the feeling that a significant percentage of Web sites have the life span of a housefly and about as much chance of meeting an untimely end. The Internet is a wildly uncontrolled, often inspired, outlet for the need to communicate ideas, research, creativity, and information to anyone who will listen. This abandonment of social, geographic, and cultural boundaries is what makes it dangerously unstable as a resource for information, business, and entertainment. There is no guarantee that a site will still be on the Web five minutes after it is visited or that its contents will remain the same. Ease of access and use makes human whim more than a minor player on the World Wide Web. Serendipitous behavior not only goes with the territory but is a wonderful characteristic of it. Unfortunately, a serendipitous world only works when inhabited by serendipitous minds, and we have been programmed from birth that even the appearance of control is better

Carol Casey is Head of the Catalog Department in Dupre Library at the University of Southwestern Louisiana; e-mail: casey@mindancer.com.

than no control at all. It may not be possible to control the Web sites that are on the Internet, but it is possible to control their fate.

For every page devoted to a pet iguana, there is a site dedicated to the communication of information and knowledge. From a librarian's point of view, this means that "in addition to evaluating the quality and utility of the information itself, we must develop methods of determining, for each resource, the optimal storage, delivery, and preservation mechanisms."[1] As far as the Internet goes, librarians are tackling the evaluating and delivery part of this declaration, and now it is time to develop storage and preservation measures for Internet resources.

Here Today, Gone Forever

It is not our place to predict what will be considered important when future generations sift through the remnants of our culture. However, as librarians, it is our place to preserve our contributions to the culture regardless of the medium those contributions manifest. Robert Skinner says it best in the conclusion of his article on the use of the Internet as a library resource:

> The Internet, for all its vagaries and faults, has strong roots in and parallels to traditional forms of knowledge transmission, and has a legitimate a place in our collections as any other type of library material.[2]

Allowing Web sites to flit in and out of existence is to let the same type of material that is painstakingly preserved in physical mediums to be lost forever. There is someone out there who probably has collected everything humankind has ever produced on the ownership of iguanas as pets, regardless of medium—book, audiocassette, periodical, or videotape. It makes no sense to think that this person does not see the pet iguana Web sites on the Internet as something else to add to his or her collection. Yet, does an archivist of Texas history see Texas-related cybernewsletters, sites such as www.traveltex.com, or scholarly papers on Texas published only on the Internet as something to add to their own collections alongside the paper and microform copies of newsletters, gazetteers, and monographs? Perhaps, but is anyone making a move to preserve these Web sites in a systematic way? "Just as with microforms and videos, the fact that information is in a non-book format is not an adequate reason to keep it out of the collecting purview."[3]

There is no guarantee that a site will still be on the Web five minutes after it is visited or that its contents will remain the same.

Although there are many Web sites we wish would just disappear, there are quite a few that form a strong core of knowledge on the Internet. After all, one must remember that "the Internet was conceived and built by the research and educational community."[4] Unfortunately, content does not dictate levels of stability. It is possible that the page devoted to pet iguanas has a solid hold in cyberspace, whereas the fate of a scholarly Web site on iguanas is uncertain. Even if the author of the Web site saves the computer files on a labeled disc and puts them in safe storage, the life span of a computer disc is relatively short and it does not take much to damage it. Just as an attic is not the ideal place to store things we want to preserve, a Web site on a computer disc in a box is not the best way to save things. Web sites need to be collected into archives with controlled storage and access if they are to survive the Internet age at all.

Web Site Authorship versus Web Space Ownership

More often than not, Web site stability relies on who "owns" the server space where a site resides. Ownership means that the author of the Web site has control over use of the space as long as the content is within the policies of the administration of the server. Control comes from paying for the Web space through an Internet service provider; signing up for it using a free Web space provider; or owning a server. The amount of space available depends on the Internet service provider and one's own budget. Given the nature of the Internet, it is not an unusual practice to have a Web site spread across many servers without a visitor realizing that each frame on a page may come from different servers on different continents. If one wonders why everyone is not signing up for the free Web space, there is a nonmonentary price to pay for space, usually a mandatory display of an advertising banner. For some, that is too high a price to pay.

More often than not, Web site stability relies on who "owns" the server space where a site resides.

Many folks in the academic world use the space allowed them on their university accounts to post Web pages. They cannot claim ownership of this space, just the right to borrow it for as long as they are associated with the institution or according to the Internet usage policy of the university. The irony is that many of these sites possess the content and quality that librarians want to preserve. Sometimes an author makes an effort to move the site to another address, but only if he or she has developed enough understanding of the Internet to do this and believes it is worth the effort to maintain.

One of the downsides of owning the space for a Web site is that of maintaining the site. Besides paying the Internet, and sometimes telephone bills, the author has to make all the changes, react to all the feedback about how the graphics are not displaying or how the text color is unreadable on certain browsers, and make sure the links to other Web sites are still valid. In the end, the author may conclude that too much time is spent on maintaining the site. The ideal situation is to have a Web master take care of all the technical details so that the author can concentrate on the content or go on to other projects.

Physical archives and libraries are not the only places that have to weed and withdraw items for lack of space for newer items. Sometimes an author wants to post a new Web site but does not have enough space for both the new site and an existing site. The difference between a physical library and the Internet is that the Web site author has the option to buy or borrow more Web space. More often than not, though, the easiest choice is to remove the old page, usually with the justification that it has been on there forever and it is time for something new.

Living in a cybervacuum, most Web authors have no idea whether anyone visits their Web site, much less finds it interesting or useful. Without a hit counter, sophisticated statistical program, guest book, chat page, or some other evidence of visitor activity, the page appears static and untouched. To visitors, it may be the only resource that covers a particular subject in a way that makes it a favorite reference source, but the Web site author may never know this. It is ironic that the people who have Web sites just because they want to, or who are obsessed devotees of some aspect of popular culture, are more adept at advertising their page, engaging visitor interaction, and keeping careful statistics than those with sites that offer stronger information to a grateful few. So when an author pulls a site off the Web, it is usually without the knowledge that half the reference departments in the country may have a bookmark to it and have it listed in their virtual reference guides.

For whatever given reason a Web site is removed from the Internet, the underlying reason is that Web site authors suffer from the same tunnel vision as the rest of the world. They cannot grasp the impact the Web site has on visitors and the importance of its place on the Internet to help future generations map the intellectual landscape of our time.

Physical Archives of Web Sites

All Web sites exist in a physical medium, whether a hard drive, mainframe, CD-ROM, or computer disc. A Web site, after all, is a set of computer files. Web sites can be preserved by collecting the computer files and storing them on the best possible storage medium, currently the CD-ROM. With the availability of affordable CD recorders and blank CDs just a few dollars each, in-house CD-ROM production is not only feasible, but economical. A CD-ROM can hold 127 megabytes of data. It is not inconceivable to have up to two thousand Web sites on a single disc if they have little or no graphics or have only one to two Web pages. Because CD-ROM production is relatively cheap, a second copy can be loaded onto local area networks (LANs), local PCs, or the Internet for easy access. An interesting difference between copies of Web sites and facsimiles (print and electronic) of special collections materials is that the user copy of the Web site is in the same format as the original, without any degradation of content or quality.

Access to the physical collections can follow traditional lines of cataloging Internet resources with some modifications to take into account that the Web sites no longer reside on the Internet. Depending on the sophistication of the online catalog and the overall technology at a given library, the catalog entry can point to both the physical and virtual location of the Web site.

Cyberarchives of Web Sites

At first glance, it seems logical to have an archive of Web sites housed on the Internet itself. After all, Web sites live in cyberspace. But the nature of the Internet creates interesting challenges for setting up and maintaining such an archive. As Charles B. Lowry points out, the Internet "is really just a large distributed computing system with a decentralized administration, which makes it enormously complex and difficult to make 'user-friendly.'"[5]

One of the first challenges is to communicate to the casual visitor that a Web site is a part of an archive. Confusion already exists with pages being pulled out of the context of a Web site by search engines and with multiple versions of the same Web site existing without indications to the visitor which is the most recent incarnation. Adding Web sites that may include obsolete or out-of-date information, or older versions of current "live" sites will just add to the chaos. It is like interlacing a book of trivia within a book of trivia and expecting to easily pick out what parts come from which book.

For whatever given reason a Web site is removed from the Internet, the underlying reason is that Web site authors suffer from the same tunnel vision as the rest of the world.

The road maps in cyberspace are only as good as the information contained on the individual Web sites. This means that a library must take over ownership of a copy of the Web site in the same way that it owns a copy of a book. The Web site must have a "property stamp" placed on it, and access must be controlled so that a visitor who finds it will be aware that it is a part of an archive and has a different set of values attached to it than other Web sites on the Internet. In addition, it must contain links to the cyberarchive main page where an access and search mechanism resides, along with information on the archive's scope and purpose.

A major problem to overcome is when an archived Web site is picked up by a search engine and the visitor is unable to distinguish the site from any other on the Web. There are several solutions to this problem. One is to add coding to the Web sites that masks them from the search engines spiders. Of course, this does not necessarily keep the Web sites off the search engines with databases that are compiled and indexed manually. Another solution is to have the search engines index intermediary pages that provide access to archived Web sites but not index the Web sites themselves. A final solution is to put the property stamp and perhaps a disclaimer in the MetaData summary statement that displays below the entries of a search engine hit list.

Eventually, the Internet community will get used to the existence of "old" archival copies of Web sites residing next to new ones on the Internet.

Eventually, the Internet community will get used to the existence of "old" archival copies of Web sites residing next to new ones on the Internet. After all, no one seems to have a problem with eighteen editions of the same book shelved together in a library, even though the older editions are only useful for historical purposes.

Collection Development

As with any collection, content is up to the collector. Although a solid collection development policy focuses the collector in the search for potential Web sites for the archive, finding every Web site worth saving is impossible. As with everything, it is hoped that the effort made is enough to save a good representation of the whole.

Along with unique, one-of-a-kind items, physical archives have books and journals and other items that have been produced in multiple copies. Following this example, the Web site archive does not have to be limited to sites that have been removed from the Internet. Realistically, all the Web sites are going to be removed or replaced eventually, so it is never too soon to put together a comprehensive collection. To keep from adding to the clutter of cyberspace, Web sites that are still alive and well on the Internet can be housed in the physical archive only and moved to the Web when the original has been removed.

The rapidly changing technology will dictate many aspects of Web site collection development, storage, and display. As with all other aspects of librarianship, keeping up with these changes must be as important as keeping up with the explosion of Web sites that need to be preserved.

Acquiring Web Sites

The hardest part of acquiring Web sites for an archive will be selling the idea of a cyberarchive to the Web site owner. After cyberarchives become more commonplace, it is likely that libraries and other collectors will be avalanched with inquiries and computer files for every possible type of Web site, much in the same way that libraries are swamped with gift books and materials. Until then, the collector will expend a lot of energy just explaining what the cyberarchive is all about.

Knowing human nature, some Web site owners probably will want monetary compensation or put restrictions on use of their sites. Issues such as these will be ironed out in collection development policies and will be a part of the never-ending, lively debate on whether a price can be put on a Web site and, of course, the ongoing dilemma of copyright. In the meantime, an archive can promise the best security for an insecure Web site owner, which is a promise that the site still will be around when all these issues are settled. The library will hold a copy of the Web site without any ulterior motive other than to preserve it.

The actual process of acquiring a Web site is simple enough. The collector either sees or is alerted to a potential site and contacts the owner or author. A prepared explanation of the cyberarchive should be a part of the initial contact. A decision is made as to whether to add the site to the collection while it is still on the Web or to wait until the Web site owner is ready to remove it from the Internet and archive the final version of it. A Web site owner may be approached by more than one cyberarchive. Having a Web site in more than one archive only adds to its chances of survival.

Adding Web Sites to the Collection

The idea of writing even a call number in a rare book sends chills up the spines of librarians, so the thought of tampering with a Web site while trying to preserve it in its original state requires a mental adjustment. The physical copy can remain unaltered, but any copy that is put on the Internet, or even a LAN, because a patron might not be able to distinguish between the two must contain marks of archival ownership.

The nature of the Internet forces the need to make necessary changes to certain elements of a Web site. To preserve the intent of the site and the integrity of the text and graphics, it is necessary to keep the links to other Web sites and e-mail addresses in place, but they must be disabled. If one thinks of a Web site as a snapshot in time, the disabled links are no more bothersome than an eighteenth-century travel book that mentions a quaint restaurant on a city corner where a sixteen-screen movie theater currently stands. Although it may be argued that someone with any sense will not use an eighteenth-century guidebook to find out what is in a city today, the same argument eventually may be made for an archived Web site. It is just a matter of getting used to the idea that just as there are collections of outdated books, there also are collections of outdated Web sites.

An identifying mark to indicate that the Web site belongs in an archive, along with links to the cyberarchive main page, should be placed so they are visible but do not interfere with the content of the site. MetaData containing keywords, summary statements, the archive property stamp, and a disclaimer or statement about the archive also should be added.

As the technology to create and access Web pages becomes more sophisticated, the need to physically "deface" a Web site with an identifying mark or even links to the main archival Web site will not be necessary. Currently, some browsers allow small pop-up windows to display simultaneously with the main browser window. These windows can contain graphics, text, and links, leaving the original page untouched. Unfortunately, not all browsers support this advancement and, more important, not every Internet user has the most recent browser upgrades, modems, or computers. Any method of putting identifying marks on a Web site must be readable by the oldest browser, the slowest modem, and the most ancient computer.

Conclusion

In terms of human culture, the Internet is still very much an infant, but someone forgot to tell this to the millions of people who have added their contributions to it. With the advent of the Graphical User Interface, the Internet took on a new persona known as the World Wide Web and captured the imagination of the world. The relative ease of use and sudden availability of access and Web space to anyone who wanted it created a communications explosion that keeps expanding.

The Internet has grown too large, too quickly, leaving librarians choking on the cyberdust. The disembodied nature of Web sites makes it difficult to remember that they deserve the same attention to collecting and preservation as their physical counterparts. It is up to the librarian to decide whether to let this information and cultural resource burn out and disappear or to try to preserve the wealth of creativity and knowledge found on the Internet.

Notes

1. Samuel Demas, Peter McDonald, and Gregory Lawrence, "The Internet and Collection Development: Mainstreaming Selection of Internet Resources," *Library Resources & Technical Services* 39 (July 1995): 275–90.

2. Robert Skinner, "Collecting Bits: The Internet As a Library Resource," *Collection Management* 21, no. 3/4 (1996): 121–37.

3. Susan K. Martin and Don L. Bosseau, "Organizing Collections within the Internet: A Vision for Access," *Journal of Academic Librarianship* 22 (July 1996): 291–92.

4. Paul Evan Peters, "How the Information Network Will Affect the Research and Education Communities," in *The Emerging Information Infrastructure: Players, Issues, Technology, and Strategies, Association of Research Libraries, Proceedings of the 123rd Meeting, Part I. Arlington, Virginia, October 20–22, 1993,* ed. Dru Mogge (Washington, D.C.: ARL, 1994), 47–51.

5. Charles B. Lowry, "Putting the Pieces Together—Essential Technologies for the Virtual Library," *Journal of Academic Librarianship* 21 (July 1995): 297–300.