Ephemeral to Enduring: The Internet Archive and Its Role in Preserving Digital Media

Eli Edwards

The Internet was first created as a communications network, but it has produced myriad effects in many sectors of technology, business, and society. Among these important changes are the growth of electronic publishing and the increasing use of the Internet as a medium for disseminating diverse documents, literature, art, journalism, commentary, and miscellaneous expressions of information and thought. Digital media, and the rise of Internet distribution of it, has caused profound changes in librarianship as well.

In spring 2001, it was estimated that more than 93 percent of all new information is "born digital"—originating in digital form.[1] But just as the Internet, particularly the World Wide Web, is noted for its growing capacity and diversity of material, it is also notorious for its impermanence. The average life span of a Web site is approximately six weeks.[2] Lyman cites that 44 percent of Web sites online in 1998 were not found the following year.[3]

How to preserve digital media has been a growing concern of librarians, archivists, computer scientists, and scholars of various backgrounds. Digital libraries have been formed—often as cooperative efforts between similar brick-and-mortar libraries—and libraries have developed protocols to add digital media and bibliographic data to their current collections. According to Feldman, "these archiving activities—important and revolutionary as they are to the future of information service—still focus [on] reproducing print products in a digital venue."[4]

In addition, there is increasing pressure to produce and distribute information in solely digital formats in many sectors of publishing. However, there are grave concerns that governmental and commercial interest in preserving digital media has been less than enthusiastic, and thus the results have been less than rigorous or impressive.

Archiving the Internet

There still remains the challenge of archiving the Internet itself. There is a significant growth in Internet-only material, including multimedia (images, audio, and video), journalistic accounts, fiction, personal journals, academic and business documents (papers, presentations, and proceedings), and government documents. Scholars view the Internet as an important artifact that warrants archiving and preservation, but efforts to archive the Internet have been piecemeal. Right now, there is no international standard, no agreed-upon protocol, and no single consortium that has developed a way to archive the Internet.

In the absence of a centralized effort, a number of Internet-archival projects have arisen. Most of these projects are in the purview of national libraries, such as the British Library's new project to archive Internet Web sites, focusing on sites of British interest.[5] As in the case of the U.K. effort, most government and national libraries have strategized to preserve national material—both official government Web sites as well as Web pages—that is registered under each nation's domain.

Among these projects is one that is easily set apart from the others. The most well-known of the Internet-archiving projects and, in some ways, the most controversial, it is also one of the earliest projects of its kind and it has the broadest mission. The Internet Archive (IA) has proven to be both an anomaly and a vanguard in its attempt to archive as much of the Internet as it can via its technological processes. What IA collects, how it collects, and how it gives access to its collections are important considerations in defining and evaluating the role of Internet-archival projects, particularly as a major activity of digital librarianship.

IA was founded as a nonprofit organization in 1996 by Brewster Kahle, founder and CEO of Alexa Internet. IA is the noncommercial sister organization of Alexa Internet (now a wholly-owned subsidiary of Amazon.com), and the two companies share technology. Alexa's Web crawlers search the Internet and accumulate Web pages and related information to those pages. After a six-month period, the pages themselves go to IA, while the information is used by Alexa to add such functionalities as listings of related sites based on users' patterns and such bibliographic information as the site's domain and what organization is running it, how long the site has been running, user ratings, and basic subject cataloging for the Web site.

What Is in the IA?

As of January 2004, IA is in excess of 300 terabytes of data, which includes more than 1,200 short films in MPEG-2 and MPEG-4 formats and more than 30 billion Web pages. Right now, IA has the following collections:

Web Sites

IA has several collections of Web pages, including the following:

Eli Edwards (eedwards@slis.sjsu.edu) is an MLIS student at the School of Library and Information Science, San Jose State University.

- a general collection of Web sites (now accessible via the Wayback Machine interface);
- two collections of Web sites from the 1996 and 2000 presidential elections;
- Web pages containing content related to the terrorist attacks on the World Trade Center and the Pentagon on September 11, 2001 (news, reactions, commentary); and
- a Web Pioneers collection highlighting early commercial and organizational sites on the Web.

Multimedia

IA's multimedia collections have historical and such current material as live and animated images. The collections include the following:

- a movie archive of more than 1,200 ephemeral short films (which have limited copyright or are part of the public domain) donated by the Prelinger Archives; and
- a new sister site (www.televisionarchive.org) that has worldwide broadcast news from September 11, 2001.

Digitized Content

IA also offers access to digitized text archives, including the following:

- The ARPANET collection, commonly referred to as the precursor of the Internet, includes "memoranda, interview notes, periodicals, papers, and other materials documenting the development of the Advanced Research Projects Agency Network (ARPANET) of the U.S. Department of Defense"[7]
- The Million Book Project
- Project Gutenberg
- Children's Library, the entire collection of the International Children's Digital Library (see the following paragraphs for more information on this project)

While IA has always proclaimed free access to researchers and the public for its collections, the methods for accessing the growing collection of Web pages was a barrier to most casual users of IA. Researchers interested in accessing any of the Internet collections had to submit proposals and were provided with Unix shell accounts on the IA's servers; downloading content required some knowledge of Unix programming as well as account maintenance.

This changed in October 2001, when IA introduced the Wayback Machine, a custom-search engine that allowed users to enter a URL and receive—from that URL—links to pages that have been archived, whether the Web site is currently active or dead. One feature of the Wayback Machine is the ability to search for video, audio, and images in addition to text, binary, and PDF files.[8] The interface is designed solely to work with the collection of Web sites; access to FTP and Usenet sites is still limited to the former procedure.

IA's one hundred terabytes of data are stored on digital linear tape (DLT) and Integrated Drive Electronics (IDE) hard drives in archive format—100-megabyte ARC files consisting of multiple individual files.[9] To aid in long-term preservation, the collections are copied on multiple sites, and IA plans to migrate the data from current to new DLT tape at least every ten years; it also is collecting software and emulators for current data formats that may become obsolete in the future.[10]

IA has taken as its mission to preserve and to offer free unlimited access to its collections, with the implicit philosophy that the Internet is not only a medium of cultural artifacts, it is itself an artifact that warrants the same consideration and efforts toward preservation and access that books, newspapers, periodicals, films, and paintings currently have.

Gaps in IA

Despite the technological efforts, IA has major gaps in its basic collection. The Web became part of the infrastructure of the Internet in 1992; IA is missing material from the first four years of the Web. In addition, while there is a small amount of non-Web material from the Internet (such as the aforementioned Usenet and FTP archives), IA is lacking in content from other prominent pre-Web protocols, especially Gopher.

In addition, the technology is unable to retrieve or archive restricted-access Internet material (whether via password protection or Internet protocol [IP] recognition), as well as Web sites on secure servers. This excludes material that is part of fee-based subscriber services (online databases and journals), as well as free content that requires registration for access. There is also the problem of *orphan pages*, Web sites that exist in isolation with no links to or from their locations to other sites. Since Alexa's/IA's technology works by following links from one site to another, a site with no links to the rest of the Web is inaccessible and will go unarchived (although Alexa does have a feature for Web-site owners to submit their sites for archiving). The Web crawlers are limited by how many times they may visit a domain to collect new or modified pages; while the intervals of new archival material for a site have been drastically reduced, there are still sizeable gaps in content that is lost between each archived page.

Another technology hurdle IA faces is a problem with certain types of dynamic content on Web pages. Cascading style sheets (CSS), PHP programming, and other XML protocols that render Web pages in standard, localized HTML

can be archived and displayed as intended from IA. Other types of dynamic content, such as JavaScript, will not work as originally programmed if the site is archived. For any dynamic content that requires access to the originating host (not the server that the page itself is stored on), the links to the originating host are broken in the automated archiving process. Thus, the archival copies of the site will have reduced functionality in comparison to the original site or may not function at all.

Moreover, there are ways for content providers to, intentionally or unintentionally, exclude their material from IA. IA abides by the standard for robot exclusion—its Web crawlers will not archive any site that has a robots.txt file in the site's server directories.[11] It also maintains the policy of removing material from IA at the request of the Web-site owner. Since September 11, 2001, a significant number of federal agencies have removed or revised their Web sites due to concerns regarding sensitive data being too accessible to potential terrorists. The U.S. Nuclear Regulatory Commission (NRC) asked IA to remove copies of the NRC Web site after the agency deleted information about the 103 nuclear reactors currently operating in the United States. IA complied with the request, and the original pages are no longer available.[12]

Copyright Issues

IA has been conceived as not only a repository of Internet material, but also as an Internet library, specially formed to preserve, store, and provide access to artifacts in digital format. The directors of IA consider its activities, and its very mission, as a function of scholarship and preservation of historical and cultural material; thus, its archival activities are a fair-use exception to laws regarding access of copyrighted material.[13] Some content providers disagree.

In a letter to the *Chronicle of Higher Education,* Professor Stephen R. Brown of American University complained that "[t]he Internet Archive is nothing more than an enormous copyright violation disguised as a library."[14] Newspapers, especially those that have used their digital and microfilm archives as profitable ancillary revenue streams, have also complained that IA is undermining their archival collections by offering the same material for free, without the express permission of the papers. However, since IA does not provide daily or weekly archival copies of newspaper or journal Web sites and is willing to remove material on request, the complaints have been somewhat muted.

How IA handles such copyright issues has significant affect on its collections. As a private nonprofit, it relies on philanthropic contributions not only for its funds but also for its collections. Its technical and operational infrastructure is based on automated processes; it does not seek permissions or special access to any Internet material. Therefore, IA is limited to the public, or visible, Web. However, it cannot penetrate, and thus cannot preserve, the invisible Web: the sites (newspapers, databases, subscription-based journals, fee-based sites) and information sources that comprise a large part of the Web but are not accessible to every user.

The IA Is Not a Library

However, even though IA bills itself as a library of the Internet, it lacks many basic functions and controls exercised by libraries. Currently, IA does not index or catalog its collections; classifications are based solely on the Web site's URL (the Wayback Machine does have an advanced feature that recognizes multiple aliases for a single URL). There is only one access point for searching IA—the URL. Currently, neither IA nor its interface has any keyword search capability and cannot provide bibliographic descriptions of the material.

And unlike libraries, there is a very limited notion of collection development for IA. Its goal is to archive as much of the Internet as its technology will allow. There is no judgment on the veracity, quality, or appropriateness of the content being archived. And there are no filters to prevent age-inappropriate material or content censored in another jurisdiction from being accessed by anyone who knows the URL for such sites. IA's general collection includes such controversial Web sites as www.rotten.com (a site that collects highly graphic and explicit photos from autopsies, accidents, and crime scenes), neo-Nazi and racial-supremacist sites, and anti-abortion sites that have been accused of provoking violence against women's clinics and medical personnel (such as www.christiangallery.com/atrocity or the "Nuremberg Files" Web site).

IA is interested not only in preserving "born-digital" items; it sees a role for itself in carrying out digital preservation of print material. To accomplish this, IA is expanding its partnerships with other organizations, including traditional libraries, to provide digitalization and electronic dissemination of bibliographic resources. It has announced two high-profile projects in this regard:

- With the University of Maryland's Human-Computer Interaction Lab, IA will establish and maintain the International Children's Digital Library, "a large-scale digital archive of literature for children ages three to thirteen."[15] The project, with assistance from the Library of Congress, will begin with a collection of 225 digitized children's books and will eventually grow to ten thousand titles from more than 100 countries.
- IA has donated a copy of its entire collection to the new Bibliotheca Alexandrina (Alexandria Library), as well as two thousand hours of Egyptian and American television broadcasts. IA also donated a

book-scanning facility to digitize books in Arabic that would be available via interlibrary loan. The digitized holdings will also be available via IA.

According to Kahle, the Internet is a medium for artifacts that are considered to be ephemeral, both bibliographic and nonbibliographic.[16] IA wants to make those artifacts enduring for current and future scholars and the public. It provides resources for cultural and historical scholarship, the study of technology's effects on society (cultural, legal, economic), and "snapshots in time" of historical events. IA also extends the functionality of the Internet. Even very popular Web sites will only exist as long as the Web-site owner is willing and able to continue and maintain the site. The current recession and its extremely destabilizing effect on high-tech companies have especially affected new media and Internet-only content providers. This has led to sudden removals and disappearances of popular niche and mainstream Web sites. However, archival copies of these dead sites allow for continued access to their content.

Library-Based Internet Archiving Projects

Web and Internet archiving projects that are developed and implemented by libraries tend to use a different model and have a different focus. Such projects include:

- PANDORA (Preserving and Accessing Networked Documentary Resources of Australia)—National Library of Australia
- AOLA (Austrian On-Line Archive)—a joint project between the Austrian National Library and the Vienna University of Technology
- Our Digital Island—State Library of Tasmania
- NedLIB (Networked European Deposit Library)—an EU-funded consortium that includes libraries in the Netherlands, France, Norway, Finland, Germany, Portugal, Switzerland, and Italy
- Kulturarw[3]—The Swedish Archive[17]

All of these efforts, which include the Bibliothèque nationale de France and Die Deutsche Bibliothek (the German National Library), are based on a primary strategy to preserve national material on official government Web sites as well as Web pages that are registered under the nation's domain.[18] In addition, the vast majority of these projects are based in part or in whole on the manual selection, collection, and metadata treatment of Web sites to meet the criteria of national or cultural interest, with limited use of automatic *harvesting*—that is, the use of robot or spider software to search the Web for eligible material and capture it for archiving. The Swedish Archive is the only other archive besides Kahle's to draw all of its material via harvesting; however, what is harvested is still determined by the criterion of Swedish interest.[19]

There is another key distinction between the emerging library model of Internet archiving and the methods used by IA—permissions. As mentioned above, IA does not seek the permissions of domain owners or copyright holders when amassing content for preservation, though it will remove any page or site on the basis of a complaint by the content owner. The scope of IA makes any attempt to license proprietary information or otherwise negotiate permissions for desired content a technical and logistical impossibility at this time. The philosophical underpinnings of IA, as articulated by its founder and chief spokesperson, buttress this: Kahle perceives the Internet as a unique historical artifact that will be of invaluable use to current and future scholars. Thus, IA's attempt to store and provide access to digital content, whether proprietary or in the public domain, is inherently covered by the fair-use exception to copyright regulations.

The archival models used by national libraries have taken a much different approach. Most of the projects have made the negotiation of licensing and permission to access, preserve, and display proprietary digital content an initial and central task. Die Deutsche Bibliotek, for instance, is working with German publishers to create and maintain a digital deposit for their online journals, periodicals, and other Web-based material.[20]

There are weaknesses to this library model, one of which is the probability of underestimating the historical or scholarly value of material, particularly by favoring commercial, well-designed content over amateur content (such as personal Web pages). Rauber, Aschenbrenner, and Schmidt of AOLA mention a couple of assumptions that are used to argue against the archiving of personal or trivial content: "Who is interested in some fellow's homepage?" and "Important information is published in 'real' media anyway." They respond to these assumptions by arguing that "Letters [are] more interesting than books," and that "ads, posters, and snippets tell more about a society than 'high-quality information sources.'"[21] They then ask rhetorically, "What if only codices had been preserved [from previous eras]?" What may appear to be detritus or vanity projects today might provide important contextual clues to future historians and other social scientists. In this way, IA's holdings and overall collection policy may significantly counteract the contextual gaps in other archives of the Internet.

Advice from the Founder

However, IA is intent on the goal of being an Internet library, not simply an auxiliary repository of digital doc-

uments. Founding director Kahle regards electronic disseminators of information as digital librarians and has suggested a code of ethics for digital librarianship. The suggested code includes the following tenets:

- Don't give away user logs except for scholarly use.
- Take the job of information serving seriously.
- Count on wide use of the information served, for good uses and bad, so be proud of the information and the collection.
- Users learn as much from a question that has no answer as from the ones with answers. This requires a complete and up-to-date collection.
- Assume that the patron will not know your affiliations . . . do not tempt patrons to use a service they would regret if they knew more about you.
- Respect your patrons.[22]

Libraries and IA

Lyman asks the following question: "Which profession should develop digital archives—librarians or computer scientists?"[23] His answer is that both professions offer overlapping strategies that include unique functionalities based on the philosophies of each discipline. Moreover, according to Lyman, IA illustrates the computer scientists' model of Internet archiving, one that is meant to preserve not only the content and information on the Internet but also the structure and interconnections of the Internet as a technical object.[24]

The potential for IA's successes and failures in preserving our past and present is already apparent to many in the technology realm. Scholars and technology are increasingly concerned with the possibility of a Digital Dark Age, a period in the not-so-far future when the manuscripts and ephemera used by historians, social scientists, and others to examine the past and present will not exist in a significant accumulation to yield useful historical or social context. At a conference of special librarians in 2003, historian David McCullough and futurist Stewart Brand both addressed the growing disparity of digital content to analog or physical media and the dire consequences of not being able to understand our history. Brand referred to the Digital Dark Age directly, while McCollough lamented that future historians will not have the manuscripts and diaries that have been so instrumental to his own work. Both noted the importance of ephemera to provide context to events, locations, and even structures.

In the meantime, IA's potential to assist in staving off this Dark Age is already being recognized. California has a large, well-organized state library, which maintains a catalog of material that the state has published, with information on how it can be accessed by other agencies and the public. A recent report on current efforts of state governments to provide permanent public access (PPA) to electronic information found that the state of California has made only inconsistent efforts to preserve access to digital information, despite the attempts of the state library: "Due to limited staff and other resources, it is often difficult for the state library to capture many electronic government documents."[25]

In a May 2003 catalog of California state publications, there was an entry for an e-commerce report published and placed on the state government Web site in PDF format in 2000. According to the catalog entry, the publication is out of print, but the PDF file was still accessible on the Web via IA. However, the PDF has since been removed from IA's server.

In effect, the state of California is relying on a private nonprofit to provide public access to public information.

The issues that IA has with digital media are similar or identical to the concerns and issues of physical and digital libraries. Defining fair use for digital-based information is a major area of concern. Libraries and IA share issues of digital rights management, storage, preservation, and access. While IA's mode of operation does not follow the traditional library model of archiving and preservation, its intent in archiving digital media is a familiar one within librarianship: to preserve and provide current and future access to scholars and the general public. Many information scientists, archivists, and others concerned with the future of preserving and accessing digital artifacts agree that the IA model and the emerging library model of Internet archiving are complementary efforts that have already shown some measure of success and have the potential to develop significant collections of digital material.

IA is just one project in a growing, collaborative effort to preserve an immense information system. Unlike most of those projects, IA is not funded directly by tax dollars nor is it under the administrative restrictions of a government agency. IA has been able to develop and implement procedures for Web archiving that have saved billions of bytes of data from obsolescence and neglect. Unfortunately, it will not be able to archive the greater part of the Internet, such as the orphaned sites, dynamic Web pages, or sites that restrict access based on passwords, IP recognition, or other protocols, which increasingly comprise what is termed the Deep Web. However, IA is playing a key role in the preservation of digital online media. In order to extend the life of information on the Internet, libraries, archivists, and computer scientists will need to examine IA and its successes and failures and draw lessons from what has been done by IA and other projects, as well as what still needs to be done, to preserve our growing yet precarious digital heritage.

Author's note: Eli Edwards is in no way affiliated with, or, in any way, connected to IA, Alexa Internet, or any of its principals, staff, board members, or donors.

References and Notes

1. Richard Wiggins, "Digital Preservation: Paradox and Promise," *Library Journal netConnect* (spring 2001): 12–15.

2. Jim McCue, "Can You Archive the Net?" *The Times* (London), Apr. 29, 2002.

3. Peter Lyman, "Archiving the World Wide Web," *Building a National Strategy for Preservation: Issues in Digital Media Archiving* (Washington, D.C.: Council on Library and Information Resources, 2002). Accessed May 5, 2002, www.clir.org/pubs/reports/pub106/web.html.

4. Susan E. Feldman, "'It Was Here a Minute Ago!': Archiving on the Net," *Searcher* 5, no. 9 (Oct. 1997): 52.

5. McCue, "Can You Archive the Net?"

6. Internet Archive, "Internet Archive: Archive Collections," The Internet Archive, Apr. 26, 2002. Accessed Apr. 13, 2002, www.archive.org/index.php.

7. Internet Archive, "Arpanet," The Internet Archive, Apr. 26, 2002. Accessed Apr. 13, 2002, www.archive.org/texts/arpanet.php.

8. Greg R. Notess, "The Wayback Machine: The Web's Archive," On the Net, *Online* 26, no. 2 (Mar./Apr. 2002): 61.

9. Internet Archive, "Storage and Preservation," About the Internet Archive. Accessed Apr. 13, 2002, www.archive.org/about/about.php#storage.

10. Ibid, "Preservation."

11. Jeffery Selingo, "In Attempting to Archive the Entire Internet, a Scientist Develops a New Way to Search It," *Chronicle of Higher Education* 44, no. 26 (Mar. 6, 1998): A27–A28.

12. "Now You See It," Net Effects, *Foreign Policy* no. 128 (Jan./Feb. 2002). Accessed Mar. 12, 2002, www.foreignpolicy.com/issue_janfeb_2002/net_effect.html.

13. Internet Archive, "Future Libraries: How People Envision Using Internet Libraries," About the Internet Archive, Oct. 4, 2002. Accessed Apr. 13, 2002, www.archive.org/about/about.php#future.

14. Stephen R. Brown, "Is Online Archive Fair Use?" Letters to the Editor, *Chronicle of Higher Education* 44, no. 34 (May 1, 1998): B9.

15. "Executive Summary," International Children's Digital Library—Info for Grown-Ups. Accessed Apr. 13, 2002, www.icdlbooks.org/adults/exec.html.

16. Internet Archive, "Future Libraries."

17. PANDORA Archive. Accessed Aug. 18, 2002, http://pandora.nla.gov.au/index.html.

18. AOLA. Accessed Aug. 18, 2002, www.ifs.tuwien.ac.at/~aola.

19. Our Digital Island. Accessed Aug. 18, 2002, http://odi.statelibrary.tas.gov.au.

20. NedLIB. Accessed Aug. 18, 2002, www.kb.nl/coop/nedlib.

21. Kulturarw^3. Accessed June 24, 2003, www.kb.se/kw3/ENG/Default.htm.

22. Andreas Aschenbrenner, "Related Work," Long-Term Preservation of Digital Material—Building an Archive to Preserve Digital Cultural Heritage from the Internet, Dec. 2001. Accessed Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Related_Work.html.

23. Andreas Aschenbrenner, "Kulturarw^3—The Swedish Archive," Long-Term Preservation of Digital Material—Building an Archive to Preserve Digital Cultural Heritage from the Internet, Dec. 2001. Accessd Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Kulturarw^3.html.

24. Andreas Aschenbrenner, "Die Deutsche Bibliotek," Long-Term Preservation, Dec. 2001. Accessed Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Die_Deutsche.html.

25. Andreas Rauber, Andreas Aschenbrenner, Alfred Schmidt, "Motivation: The Invention of the Press: Archiving the Internet: Challenges, Projects, and the Austrian Perspective," Slide 4, talk at INST conference, "Knowledge Networking in Cultural Studies," in Reichenau, Austria, May 26, 2001. Accessed Aug. 19, 2002, www.ifs.tuwien.ac.at/~aola/publications/slides_reichenau_01.pdf (English version).

26. Brewster Kahle, "Ethics of Digital Librarianship," Feb. 1992. Accessed Apr. 26, 2002, www.archive.org/about/ethics_BK.php.

27. Lyman, "Archiving the World Wide Web."

28. Ibid.

29. Joan Allen-Hart, "California State Report," *State-by-State Report on Permanent Public Access to Electronic Government Information* (June 2003): 46. Accessed Aug. 17, 2003, www.ll.georgetown.edu/aallwash/State_report.pdf.