Proving Web History: How to Use the Internet Archive

By Beryl A. Howell

Showing what content a Web site *previously* contained (as opposed to what is currently on the site) may help answer questions that attorneys confront in a myriad of cases, ranging from copyright and trademark infringement to business torts and defamation. Showing that a particular Web site is currently using copyrighted text or images or protected marks may be all that is needed in a case, but documenting prior versions of the Web site can be critical to establish the scope or extent of the illegal or tortious conduct, the amount of the damages, or the requisite *mens rea.*[1] When the Web site at issue or the offending content on it has been removed or modified, the most effective, if not the only, way to document the content is to review prior versions stored in online archives of Internet sites.

Specifically, in trade secret and misappropriation cases, showing that the same information claimed to be secret or confidential has previously been made publicly available by the claimant, such as on the claimant's Web site, can be probative if not case dispositive. Diligently searching archived versions of the claimant's Web site for such evidence can be worthwhile.

Similarly, capturing evidence from archived Web sites is helpful in intellectual property infringement cases as well. For example, in cybersquatting and typo-squatting cases, where a trademarked name or a slightly misspelled trademarked name (*e.g.*, *mcrosoft.com*) has been registered as a domain name and used as an online address for a Web site, evidence from archived versions of the Web site may establish the period of time the offending Web site has been operational and the types of goods or services offered on the site over time. This evidence can help establish intent and harm. In cyberstuffing cases, where popular trademarked names are repeatedly embedded in hidden metatags and transparent text on a Web site, search engines will pick up on the trademarked names and push the infringer's Web page to the top of search engine results, diverting business from the trademark owner's site. Even if the Web site is modified after the infringer is notified of the claim, documenting the cyberstuffing activity on archived versions of the Web site can establish the nature of the offending activity, its scope, and duration.

No matter the legal context, gathering evidence of prior versions of Web sites should be performed in a careful forensic manner with cognizance of the underlying technology used in the archiving process. This article will review strategies and methods for capturing prior versions of Web sites from the most popular of the archives and considerations that counsel should be prepared to address in authenticating this evidence.

"MAP" OF ARCHIVES

At the outset, archived versions of Web sites are available for free at multiple sites. The federal government, in particular, archives government Web sites and makes those archives accessible online. For example, the US Government Printing Office, in partnership with the University of North Texas, provides online access to federal Web sites that have ceased operation on a site called the CyberCemetery.[2] The archived deceased Web sites include Access America, Advisory Commission on Intergovernmental Relations, Office of Technology Assessment, and others. Similarly, the Electronic Research Collection (ERC),[3] which is a partnership between the United States Department of State and the Federal Depository Library at the Richard J. Daley Library, University of Illinois at Chicago (UIC), makes available the US Department of State Web pages archived from 1998 through January 2001. In addition, the National Archives and Records Administration harvested all of the federal agency public Web sites as they existed at the end of the presidential term on January 20, 2005, and makes these archives available at the 2004 Presidential Term Web Harvest site.[4]

Archives of Web sites that are not associated with the federal government are available at several sites. The Library of Congress sponsors a project called Minerva (Mapping the INternet Electronic Resources Virtual Archive), which harvests Web sites based on subject matter and then provides the collections as an archive, rather than try to harvest every Web site. The collections currently available include: Election 2002 Web Archive (July

Beryl A. Howell is a Partner and heads the Washington, DC, office of Stroz Friedberg, LLC, a computer forensics and electronic discovery consulting and technical services firm with offices also in New York City, Minneapolis, and Los Angeles. She formerly served as a New York federal prosecutor and General Counsel of the US Senate Committee on the Judiciary. Research assistance for this article was provided by Donald Allison and Jessica Reust, computer forensic examiners, and George McLean, Evidence Technician at Stroz Friedberg, LLC.

1, 2002-Nov 30, 2002);[5] September 11, 2001 (September 11, 2001-December 1, 2001);[6] and Election 2000 (August 1, 2000-January 21, 2001).[7]

Certain Internet search engines, such as Google and Yahoo, also make archived Web sites available. The Google archive provides access to the last cached version of a Web site, but not to prior versions. These cached Web sites are a backup in case the original page is unavailable and are useful since they show the date and time stamps for when each page on the site was retrieved by Google. Google and other search engines often index a Web site about once a month, but Google explains that the "cache is the snapshot that we took of the page as we crawled the web" and cautions that "[t]he page may have changed since that time" or "[t]his cached page may reference images which are no longer available." Google states that many factors affect how often it indexes a site, but a 2003 survey showed that Google revisited most sites within one month.[8] Therefore, unless a page is defunct, a Google cached site often will be 30 days old or less. To look farther back in time, the Internet Archive is probably a better bet. Sites may not be cached if they have not been indexed or if the owners have requested that the content not be cached. The date-time stamps on the Google archive may be helpful in establishing, for example, when a site stopped operating within the last six months. If a site is no longer available online, a visit to the Google cache may indicate the date when the site was last indexed.

Yahoo has recently added the ability to view cached pages by clicking on a link entitled "cached." As with Google, clicking on "cached" brings up a copy of the Web page as it appeared when it was last crawled by the search engine. By contrast to the Google cached sites, however, the Yahoo archive does not date-stamp the version of the cached site but simply notes the following: "It's a snapshot of the page taken as our search engine crawled the Web. The Web site itself may have changed." To check the previous versions of the Web site, Yahoo directs users to the Internet Archive. As discussed in more detail below, the Internet Archive contains the most extensive archive of Web sites in terms of period covered, number of Web sites and pages archived, and the number of prior versions of Web sites archived.

Other search engines that provide cached Web sites include *search.msn.com* (MSN), *ask.com* and *teoma.com* (both from Ask Jeeves), *clusty.com* (from meta-search engine Vivisimo), and *Gigablast.com*. Of these, Gigablast may be the most helpful in researching historic Web sites because its search engine results include the date that the Web page was last modified, as well as the date that the page was last indexed by Gigablast. Gigablast also provides links to the cached site, a stripped version of the site without graphics, and a link to "older copies" found on *archive.org*.

THE INTERNET ARCHIVE AND THE WAYBACK MACHINE

The Internet Archive[9] is a free online resource that was created in 1996 to build a digital library of Web pages and other cultural artifacts in digital form with the purpose of offering permanent and free access to researchers, historians, scholars, and the general public.[10] Internet Archive provides not only an archive of websites but also of open source movies, feature films, cartoons, historic newsreels, and news video and music.

Five years after its creation, in October 2001, the Internet Archive launched the Wayback Machine, which provides the public with a free online service to search for and access archived Web sites. The name of the search service is derived from the Rocky and Bullwinkle cartoon in which the characters of a bow-tied dog, Mr. Peabody, and his boy assistant, Sherman, used a time machine called the WABAC Machine to travel back in time to famous events in history.

The Web pages are collected for the Internet Archive using a search engine technology called Alexa Crawl that traverses the Internet taking snapshots of Web sites. The Alexa Crawl currently captures about 1.6 terabytes (1600 gigabytes) of Web content per day and takes about two months to complete a snapshot of the more than 16 million Web sites accessible online.[11] This search-and-copy engine is owned and operated by Alexa Internet, a for-profit company that offers a free toolbar and a number of statistical services to subscribers based upon the Web content and usage information collected. The company donates a copy of each crawl of the Web to the Internet Archive, which may make the crawl results available after six months. Thus, there is a six- to 12-month lag between the date that a site is crawled and when it appears for free use in the archives of the Wayback Machine.[12] Alexa Internet is now offering a fee-based service to access its crawl results data before it goes to the Internet Archive.[13]

The Alexa Crawl does not purport to capture all Web sites accessible on the Internet, but instead prioritizes the Web sites and pages to copy based on the number of times that a Web site is requested through the Alexa search engine. Thus, not every Web site has an equal chance of being copied at all or copied in full. Alexa Internet uses a rating system for content at all that will be captured. Content that is not popular may be deliberately omitted if not visited often. This is related to Alexa's business model for selling databases of frequently visited sites to customers. The result is that the Wayback Machine does not hold archived versions of all Web sites of copies of every page for the Web sites that are archived.

In addition, sites may not be archived if they are password-protected, the site owners have requested exclusion from the Wayback Machine, or the crawler is blocked by use of a technical flag installed by the site owner called robots.txt, or the site is otherwise inaccessible. When the site is blocked by request or use of a robots.txt flag, the Wayback Machine search engine will indicate this with an error message, such as "blocked site error" or "robots.txt query exclusion error."

At the inception of the Wayback Machine, the Internet Archive contained 100 terabytes of data that was growing at a rate of 10 terabytes per month. By 2005, the amount of data stored in it is more than a petabyte, with a growth rate of 20 terabytes per month, making the Internet Archive the largest data archive in the world. All of this data is stored in huge server farms in the Presidio of San Francisco.

The archived Web sites are stored across multiple servers. A version of a particular Web site that is shown as indexed on the Wayback Machine may not be available at the time when a user wants to access it. A replica of the Internet Archive is stored at the Bibliotheca Alexandrina in Egypt.[14] If a version of a particular Web site cannot be accessed on the Internet Archives' primary site, the replica site can be checked.

The replica on the Bibliotheca Alexandrina Web site is not updated frequently, however, and it does not contain as much content. Test searches conducted on *archive.org* reveals many Web sites that do not appear on Alexandrina's Web site. For example, a search for *cnn.com* yields results for pages from July 2000-September 18, 2001, on the Alexandrina's Web site, while the *archive.org* site has version from November 26, 2004.

To use the Wayback Machine, users simply go to the *archive.org* Web site, and type in the Internet address[15] in the provided search box. Any versions of the Web site corresponding to the Internet address that are archived on the servers of the Internet Archive will pop up in a chronological list. A user can review this list and select the version or date for review by clicking on the selected date. The archived version of the Web site for the date selected will then appear and can be reviewed.

The nature of the legal dispute may require analysis of multiple archived versions of a particular Web site in order to establish whether and how content changed. For example, in a contract dispute, the question of whether a party offered services or items in violation of terms in the license at issue may require documenting changes in a party's advertised offerings on its Web site during and after expiration of the license term. Critical text may simply be eyeballed as part of this analysis to document changes over time. In addition, the Wayback Machine notes changes in an archived Web site with an asterisk. This asterisk system alerts only to changes in text or graphics and not to modifications in internal or external links and or in the source code for the Web site. This may become critical if, for example, the archived Web site is cited as evidence that it was used to link to an offending site. The link to the offending site in the archive version may not, in fact, have existed or existed in the same form at the time that version of the Web site was copied for the archive.

The Wayback Machine also offers a free service of comparing any two versions of an archived site using a technology called DocuComp, which is a patented algorithm licensed by Advanced Software for use in the Wayback Machine. The comparison can show how the contents, including text, images, and links, have changed over time and between any two versions being compared.

"MISSING" ARCHIVED WEB SITES

When a search for an archived Web site has negative results, this does not mean that the Web site does not exist, is not archived, or is only of current vintage. The Web site may have been excluded from the archiving process or in fact, the Web site may be archived but review of the archived versions is blocked. The Internet Archive takes steps to avoid archiving web sites for which the owner has indicated a preference to be excluded. A universal technical standard that indicates an exclusion preference is called the standard for robot exclusion (SRE). A file called robots.txt can be added to the header information on a Web site or specific Web page by an owner, and a denial or disallow command within that file can serve as a flag that the owner does not want the entire Web site or particular Web pages copied or scanned by a Web crawler. In other words, the directions in the robots.txt file can be set to allow full or partial copying or copying exclusion. The Alexa crawler respects this preference and will not copy those sites or pages with a robots.txt file embedded.[16] Alexa Internet and Internet Archive take this respectful technology a step further: When robots.txt is added to a Web site, Alexa will exclude the site from being copied by its crawler, and the Internet Archive will go back into archived sites to remove content already captured.[17]

In addition, intellectual property owners who believe that infringing activity is occurring on a Web site may contact the Internet Archive and request exclusion of the offending Web site. The Internet Archive provides specific directions to copyright and trademark owners seeking to have third-party Web sites containing infringing works removed from the archive. These owners must specifically identify the work allegedly being infringed and where it is located within the Internet Archive collections, contact information, and a statement made under penalty of perjury that use of the work is unauthorized by the copyright owner, along with an electronic or physical signature.[18]

The Internet Archives' respect for the exclusion preference of Web site owners and compliance with its own stated policy to remove Web sites with robots.txt flags is the subject of a recent suit in the Eastern District of Pennsylvania brought by Healthcare Advocates against the Internet Archive for, *inter alia*, breach of contract and misrepresentation due to a failure to block access to the plaintiff's archived Web sites.[19] The plaintiff operates a Web site that describes the services of the company, including helping the public get reimbursements for health care expenses, reporting on medical research, providing doctor referrals and information on discount prescriptions and healthcare plans. The company claims copyright in all of the Web site content. In mid-2003, the plaintiff installed the denial text string in the robots.txt file on the computer server hosting its Web site with the expectation that the Internet Archive would prevent users of the Wayback Machine from gaining access to the archived versions of its Web site.

Nevertheless, in another case brought by Healthcare Advocates against a competitor for misuse of proprietary and trade secret information, the defendant's counsel was able to access the archived versions of the plaintiff's Web site on the Wayback Machine by successfully circumventing the security offered by the denial text string in the robots.txt file. This circumvention was apparently facilitated by the fact that "the mechanism preventing www.archive.org from searching a particular web site's host computer server for a denial text string in the robots.txt file more than once per day was 'broken.'" In other words, when the Wayback Machine receives a query for an archived version of a Web site, the Web site is pinged for the presence of a robots.txt file denial string. If the string is found, the query is blocked, but apparently persistent queries will overcome the block. The defendant's counsel in the underlying lawsuit conceded that the plaintiff's archived Web sites on the Wayback Machine had been searched and accessed in connection with that underlying case. That counsel is now co-defendants with the Internet Archive in Healthcare Advocates' suit for copyright infringement and computer hacking.

This lawsuit will test the scope and merits not only of the claims at issue but also the indemnification provision of the Internet Archive's terms of use. Specifically, the terms governing the use of the collection of archived Web pages is predicated on the user's agreement "to indemnify and hold harmless the Internet Archive and its parents, subsidiaries, affiliates, agents, officers, directors, and employees from and against any and all liability, loss, claims, damages, costs, and/or actions (including attorneys' fees) arising from your use of the Archive's services, the site, or the Collections."[20]

CAPTURING ARCHIVED WEB SITES

Once an archived Web site has been located, the methods of capturing the virtual pages in a concrete form for use in court can vary. One method is to print each page that appears on the computer screen. The person performing or supervising the search and printing can attest to the date, time, and process used to obtain the printout. This method shows static pages of the Web site without any of the links that may remain active, other than any advertisements pushed to the site, even in the archived state. Similarly, screen-shots of each page viewed can be saved electronically for incorporation into expert reports or affidavits.

Importantly, Internet browsers and specialized tools used by computer forensic experts for downloading Web sites with metadata intact can be used to capture not only the graphical display of a Web page but also the underlying html code that is driving the display. Simply using the file save function on a browser can preserve code that may reveal who authored a contentious Web page. Saving underlying code in the same way may reveal a trademarked name written over and over again in white-on-white text, indicating that it was meant to be revealed to crawling search engines but hidden from a consumer's (or competitor's) naked eye. If two or more archived pages are linked to each other, download tools can provide a fuller layout of a Web site with its underlying code. At trial, this fuller layout can be presented to the judge or jury, and links and related pages can be navigated, much as an historic user might have surfed them.

In addition, specialized software tools are available that allow dynamic presentations, including demonstrations of any link that remains active on the Web site. One such software tool, called Camtasia, can be installed on the computer used to access the archived site to record every keystroke and screen shot appearing during review of the cached Web site. The recording of the review session is documented real time in video-like form that may be stored on a CDR or DVD for submission to court. For example, in a business diversion case, a recording of the cached version of the defendant's prior Web site may be able to show links that remain active and purportedly direct potential customers to the plaintiff's products, but the links instead actually channel users to the defendant's sites.

Beware when capturing an archived Web site that different browsers display Web sites with differing degrees of accuracy and completeness, and this holds true for archived Web sites and Web pages as well. There are a number of different reasons why some Web pages look different depending on which browser is used to view the page, including browser adherence to Web page standards, browser support of different technologies, and Web sites that do not use Web page standard code. The World Wide Web Consortium (WC3) develops the standard elements for Web site programming, which some browsers adhere to and some do not. For example Firefox and Mozilla adhere to the WC3 standards, while Internet Explorer supports additional non-standard Web-programming technologies. The resulting difference in the way that Web pages are displayed may be as minimal as the color of the scrollbar to as inconvenient as the navigation menus not working or the site content not being displayed at all.

A Web site that uses or requires a certain technology to be viewed will not be displayed correctly or completely by a browser that does not support that technology. For example, Firefox does not support ActiveX, which are software components from Microsoft that enable sound, Java applets, and animations to be integrated in a Web page.[21] For example, using a browser that supports ActiveX is necessary in order to access the Windows Update Web site, which otherwise will simply not be displayed but with an alert to the viewer that content is hidden from view. The fact that content is not being displayed or displayed in a different way from the original site is not always apparent.

The key to capturing an archived Web site as accurately and completely as possible is to examine the underlying code used to create and support the Web site to determine whether a browser is incompatible. This can be done by an examination of the source code for the initial page of the Web site. The entry point for the Web site usually includes language that will query and collect information from the browser and its computer system settings to determine the best method of providing the information from the site. For Web sites that use only standard html coding, the content and features of the site usually have the least variance across browsers. Where non-standard html coding is revealed, forensic experts capturing Web sites for litigation purposes may display the Web site with multiple browsers as a test to ensure that the display does not vary by browser and if variances are noted, capture the Web site with the browser that displays the most content.

ADMITTING INTERNET ARCHIVE DATA

Information obtained from reputable or government-sponsored online sources has generally been held admissible. For example, in *U.S. Equal Employment Opportunity Commission v. E.I. DuPont De Nemours & Co.*,[22] the defendant moved to exclude as an exhibit the printout of a table from the Web site of the US Census Bureau as inadmissible hearsay and lack of trustworthiness. The court denied the motion, stating that the hearsay exception for a public record applied. In addition, the court concluded that the printout was sufficiently authenticated under Federal Rules of Evidence 901(a) since it contained the "internet domain address from which the table was printed, and the date on which it was printed."[23] The court performed its own verification as well, noting that "[t]he Court has accessed the website using the domain address and has verified that the webpage printed exists at that location."[24] Similarly, printouts of data from other government-sponsored Web sites have been admitted over objection to the reliability of the information.[25]

Reported cases involving Web site captures from the Internet Archive are rare, even though *archive.org* is an important resource for litigators trying to establish prior representations or actions on Web sites. Significantly, in the few cases where challenges have been interposed to Internet Archive versions of Web pages, the evidence has been admitted over hearsay and authentication challenges.

The leading case for admission of archived Web sites from the Internet Archive is *Telewizja Polska USA, Inc. v. Echostar Satellite Corporation*.[26] The plaintiff in this case claimed that Echostar improperly had used the plaintiff's trademarks in "TV Polonia," a Polish-language television station, to sell subscriptions to the Dish Network satellite TV service after the contract allowing such marketing rights had expired in early 2001. Echostar argued that plaintiff had itself advertised that the Dish Network carried TV Polonia on its Web site after the marketing rights had expired and offered an exhibit of the plaintiff's Web site at various times in 2001 confirming this past Web site content. The plaintiff filed a motion *in limine* to bar Echostar from offering the exhibit on the grounds of double hearsay and lack of authentication. The court rejected these grounds and denied the motion, stating that "the contents of [plaintiff]'s website may be considered an admission of a party-opponent and are not barred by the hearsay rule."[27] In addition, the court relied on the affidavit of "Ms. Molly Davis, verifying that the Internet Archive Company retrieved copies of the websites as it appeared on the dates in question from its electronic archives."[28] The plaintiff "presented no evidence that the Internet Archive is unreliable or biased" or "denied that the exhibit represents the contents of its website on the dates in question" or otherwise "challenged the veracity of the exhibit."

AUTHENTICATION CONSIDERATIONS FOR ARCHIVED WEB SITES

The versions of Web sites and pages archived on Internet Archive can provide valuable and significant probative evidence in a variety of cases. To authenticate copies of prior versions of Web sites obtained from the Wayback Machine, a party proffering the evidence must show, under Federal Rules of Evidence 901(a) that the "matter in question is what its proponent claims." This can be done by producing the testimony, either orally or in written form, of the person who copied or supervised the copying of the archived Web site and the process followed to accomplish this task. In addition, the proponent must establish the general reliability of the copy.

The capture and use as evidence of archived Web site material must be approached with a full appreciation of three primary technical features and limitations that may affect the archived copy in order to respond to any challenges that may be raised to the completeness, reliability, and authenticity of the copy. For this reason, expertise in digital forensics, including the methods of forensic capture and documentation of the archived Web site proffered, may be recommended depending on the issue for which the archived Web site is being offered.

First, archived Web sites on the Internet Archive are compilations made over time. While the archived versions of Web sites are date- and time-stamped, the pages for each version of the Web site may not have been copied simultaneously. The Alexa crawler may take multiple passes at a Web site over the course of up to two days to try to capture the entire Web site. In short, due to bandwidth and storage constraints, all of the data on a Web site may not be captured at the same time. The Internet Archive explains that "Sites are usually crawled within 24 hours and no more than 48."[29]

Second, the archived versions of Web pages available through the Wayback Machine may not contain all of the content on each Web page that is captured. What you see is not always the complete story.

For example, when a Web site contains elements that require interaction with the originating host, copying that page for archiving breaks the necessary link with the original site, thereby reducing the functionality or eliminating entirely that particular element. The result is that the archived Web page or site has missing material, which may not be apparent or flagged for the viewer. Similarly, links originally enabled with a java script, which the Alexa crawl technology disables during the capture of the Web site or Web page, would no longer work.[30] The Internet Archive acknowledges: "Not all images are archived nor are retrievable from the original site. If they no longer exist on the original site then the images will not be available and not displayed within the archived pages."[31] Other types of coded content that the crawler technology does not capture include Flash enabled content, some photographic images, and some html coded content.

Moreover, content may not ever be captured if problem technology, such a password protected pages, or respectful technology, such as a robots.txt flag, is encountered. Additionally, even after the capture is completed, archived copies of Web sites may have content deleted if a robots.txt flag is added to the site or if a request for deletion is sent to the Internet Archive. Thus, the archived copy may show what was captured but not what was skipped or subsequently omitted.

Finally, depending on the technical sophistication of the Web site and its use of internal and outside linked material, the copy of the archived version of the Web site may not show links that existed on the Web site at the time of the original capture. Links that may have worked at the date of capture may be inactive because they simply no longer exist or are not in the archive library.

The links on archived Web sites may remain active but link to different material from that associated with the Web page at the time that it was archived. The linked material may be to current sites or to other stored link sites from a different time. Indeed, links may connect to current active sites and show *current* banner advertisements available at the site, rather than linking to sites as they existed at the date of capture. When the active links on archived Web sites pull information from the current site, the owner of the current Web site can track how many times the Wayback Machine is being queried for archived versions of the Web site. Logs of incoming IP addresses maintained by the server hosting the current Web site can reveal whether the incoming IP address originated with the Internet Archive.[32]

Alternatively, the working link may connect to sites or pages archived on the Wayback Machine around the time of the original Web site to which the link connected. The Internet Archive explains: "When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that are missing. In the event that we do not have the link archived at all, the Wayback Machine will look for the link on the live web and grab it if available."[33] In short, the process of copying a Web site for archiving may result in changes to the extent that the archived Web site may not show accurately the links that existed at the time shown for the Web site storage date. The Alexa Internet crawler technology rewrites the original link code in html to re-direct links to current or stored links.

Determining whether the content on a linked site is contemporaneous with the archived version of the site or dates from another time may be critical. For example, establishing that a linked promotion to a site containing infringing material persisted after notification from the copyright owner may be important to establish knowledge and intent in a copyright infringement suit. Each link must be checked for the date code embedded in the archived URL, or location within the Wayback Machine database, to verify whether the linked content is contemporaneous, current, earlier or later than the version of the archived Web site or page. The Internet Archive provides the following example: "in this url *http://web.archive.org/web/20000229123340/http://www.yahoo.com/* the date the site was crawled was Feb 29, 2000, at 12:33 and 40 seconds."[34]

Increasingly, documentation of offending activity that occurred on Web sites of opposing parties is relevant and, in some cases, dispositive of certain types of claims. Searching for, reviewing, and capturing archived copies of Web sites can be easily accomplished from the Internet Archive, but litigators should consider carefully the methods of capture and the issues surrounding the completeness, reliability, and authenticity of the Web site copies.

NOTES

[1] *See, e.g.*, Van Wetrienen v. Americontinental Collection Corp., 94 F. Supp. 2d 1087, 1109 (D. Or. 2000) (contents of defendant's Web site relevant to determination of whether defendant's conduct was so egregious as to merit an award of punitive damages).

[2] The CyberCemetery is located at *http://govinfo.library.unt.edu*.

[3] ERC is located at *http://dosfan.lib.uic.edu/ERC/*.

[4] The 2004 Presidential Term Web Harvest is located at *http://www.webharvest.gov/collections/peth04/*.

[5] *http://www.loc.gov/minerva/collect/elec2002/index.html*.

[6] *http://www.loc.gov/minerva/collect/sept11/index.html*.

[7] *http://www.loc.gov/minerva/collect/elec2000/index.html*.

[8] *See http://searchengineshowdown.com/stats/freshness.shtml*.

[9] The Internet Archive is located at *www.archive.org*.

[10] Kahle v. Ashcroft, 2004 U.S. Dist. LEXIS 24090, *5 (N.D. Cal. Nov. 19, 2004).

[11] *http://pages.alexa.com/company/technology.html*.

[12] *http://www.archive.org/about/faqs.php#The _Wayback_Machine*.

[13] *http://websearch.alexa.com/welcome.html*.

[14] *http://www.bibalex.org/english/initiatives/internetarchive/web.htm*; *see also http://en.wikipedia.org/wiki/Bibliotheca_Alexandrina*.

[15] The technical term for an Internet address is Universal Resource Locator or URL.

[16] Directions for removal of a Web site from the archive are found at *http://www.archive.org/about/exclude.php*.

[17] *http://www.archive.org/about/faqs.php#2* ("By placing a simple robots.txt file on your Web server, you can exclude your site from being crawled as well as exclude any historical pages from the Wayback Machine.").

[18] *Id.*

[19] Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, Civil Action (E.D. Pa., filed July 8, 2005), copy at *http://www.geocities.com/bledrydudenet/Healthcare_Advocates_v._Harding_Complaint__FINAL.pdf*. Healthcare Advocates, Inc. unsuccessfully moved to have the counts against the law firm for, *inter alia*, violations of the DMCA and the Computer Fraud and Abuse statute added to the underlying complaint, but that motion was denied. Flynn v. Health Advocate, Inc., 2004 U.S. Dist LEXIS 12536, *12 (E.D. Pa. July 8, 2004).

[20] *http://www.archive.org/about/terms.php/*.

[21] *See http://webmaster.lycos.co.uk/glossary*.

[22] U.S. Equal Employment Opportunity Commission v. E.I. DuPont De Nemours & Co., 2004 U.S. Dist. LEXIS 20753 (E.D. La. Oct. 18, 2004).

[23] *Id.* at *5.

[24] *Id.*

[25] *See* Chapman v. San Francisco Newspaper Agency, 2002 U.S. Dist. LEXIS 18012 at*2 (N.D. Cal. Sept. 20, 2002) (computer printout of page from US Postal Service Web site was sufficiently reliable to be admissible public record). *But see* St. Clair v. Johnny's Oyster & Shrimp, Inc., 76 F. Supp. 2d 773, 774 (S.D. Tex. 1999) (court deemed plaintiff's proffered data from the US Coast Guard's online vessel database insufficient since "any evidence procured off the Internet is adequate for almost nothing").

[26] Telewizja Polska USA, Inc. v. Echostar Satellite Corp., 2004 U.S. Dist. LEXIS 20845 (N.D. Ill). *See also* Attig v. DRG, Inc., 2005 U.S. Dist. LEXIS 5183, at *5, n.1 (E.D. Pa. Mar. 30, 2005) (in copyright infringement suit, parties agreed that copies of websites at issue obtained from *archive.org* are admissible evidence); Louis Vuitton Malletier v. Burlington Coat Factory Warehouse Corp., 42 F.3d 532, 535 (2d Cir. 2005) (in trademark infringement suit, evidence of defendant's Web site advertisements presented through *archive.org* capture of the site content at particular time).

[27] *Id.* at *16-17.

[28] *Id.*

[29] *http://www.archive.org/about/faqs.php#The _Wayback_Machine*.

[30] *Id.* ("javascript enabled links and actions are disabled in the comparison results to prevent errant scripts from being run").

[31] *Id.*

[32] This feature of the Wayback Machine is what alerted Healthcare Advocates in the pending lawsuit discussed *supra*, at n.20 that prior versions of its Web site had not been blocked as requested but instead were being accessed by the defendants.

[33] *http://www.archive.org/about/faqs.php#The _Wayback_Machine*.

[34] *Id.*