A WEBSITE OWNER'S PRACTICAL GUIDE TO THE WAYBACK MACHINE

HOLLY ANDERSEN*

INTRODUCTION

Almost everyone has something floating around in cyberspace that he or she wishes never existed. Whether it is an embarrassing photograph, an irrational blog rant, or a childish video, the Internet has a way of broadcasting our best (and worst) moments. Fortunately, many of these anecdotes can be removed and potentially never see the light of a computer screen again; however, in some cases, the old adage that "once on the Internet, always on the Internet" may ring true.

Most website owners are probably not aware of the Internet Archive or its incredible "Wayback Machine." While Internet Archive's mission to preserve our electronic history may be a noble one, it can cause serious legal consequences for website owners. Because the admissibility

[*] J.D. Candidate, 2013, University of Colorado Law School; B.A., 2008, Colby College. I would like to thank my wonderful family and friends for their love and support.

of screen shots obtained from the Wayback Machine has been judicially approved in a majority of federal jurisdictions, it has become even more important that website owners understand the possible implications of having their website content stored on the Wayback Machine as well as the limited options available to them for excluding their content. Unfortunately for some, the options to exclude content from the Wayback Machine are incredibly limited because the primary tool to prevent web crawling, the robots.txt file, is far from fool-proof. However, the Wayback Machine's capabilities are not entirely negative for website owners because they can use the service to their own advantage as well. By ensuring the archiving of infringers or potential infringers' websites, a website owner can monitor and potentially enforce his or her own intellectual property rights.

This note endeavors to briefly explain the admissibility of screen shots from the Wayback Machine in federal court and to expand the understanding of the benefits and limitations of the archive provided by the Internet Archive. Part II describes how the Wayback Machine works as well as how the federal courts have treated the use of its screen shots as evidence. The Internet Archive's process for obtaining the affidavit required by most courts in order to enter the screen shots as admissible evidence will also be described. Part III will present the potential benefits and problems with the Wayback Machine. While the Wayback Machine's archives can be helpful particularly in infringement of intellectual property lawsuits, the archive can be inconsistent and prove unhelpful in proving some key elements of a case. Finally, Part IV will address options to prevent access of web crawlers and to exclude a website's content from the WayBack Machine. Additionally, this note will provide recommendations to take full advantage of the potential benefits this service can offer.

I. THE WAYBACK MACHINE AND THE ADMISSIBILITY OF SCREEN SHOTS AS EVIDENCE

A. WayBack What?: How It Works

The WayBack Machine[1] is a service provided by the Internet Archive that allows people to visit archived versions of websites.[2] The Internet Archive is a non-profit entity that was founded to build an Internet library with permanent access to texts, audio, moving images,

[1] *The Wayback Machine*, INTERNET ARCHIVE, http://archive.org/web/web.php (last visited Feb. 10, 2013).

[2] *The Wayback Machine: Frequently Asked Questions*, INTERNET ARCHIVE, http://www.archive.org/about/faqs.php (last visited Feb. 10, 2013) [hereinafter *The Wayback Machine: FAQ*].

software, and archived web pages.[3] Its self-proclaimed mission is "to preserve society's cultural artifacts and to provide access to them. If libraries are to continue to foster education and scholarship in this era of digital technology, it's essential for them to extend those functions into the digital world."[4] "The Internet Archive is working to prevent the Internet – a new medium of major historical significance – and other 'born-digital' materials from disappearing into the past."[5] The Internet Archive also cites the importance of open and free access to writings that are considered "essential to education and to the maintenance of an open society."[6] The WayBack Machine is cited as a device that displays the Web on any given date, giving historians and others a literal window on the past.[7]

The WayBack Machine allows anyone to type in a Uniform Resource Locator (URL), select a date range, and begin surfing on an archived version of the desired web page.[8] For example, as of February 10, 2013, a simple search of "www.yahoo.com" shows that that website has been crawled 38,583 times, starting in October 1996.[9] An interested web surfer could then select a date and time to view the archive website. By way of illustration, a selection of June 15, 2011 at 3:28:34 shows that at 3:28 AM on June 15, 2011, www.yahoo.com had an article announcing that Hugh Hefner's wedding to Crystal Harris had been called off.[10] Another example would be to search "www.usmagazine.com" and select January 1, 2013 at 19:20:51 to see an article announcing that Hugh Hefner married Crystal Harris on New Years Eve.[11]

The way that the WayBack Machine is able to compile and store this information is fairly complicated. However, its process can be simplified somewhat. The Internet Archive has teamed up with Alexa Internet, Inc., which has designed a three dimensional index that allows

[3] *About the Internet Archive*, INTERNET ARCHIVE, http://www.archive.org/about/ (last visited Feb. 10, 2013).

[4] *Id.* (scroll to "Why the Archive is Building an 'Internet Library'").

[5] *Id.*

[6] *Id.*

[7] *Id.* (scroll to "Future Libraries – How People Envision Using Internet Libraries").

[8] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "What is the Wayback Machine?").

[9] YAHOO!, http://wayback.archive.org/web/*/http://www.yahoo.com (last accessed Feb. 10, 2013 by searching for "www.yahoo.com" in the Internet Archive's Wayback Machine index).

[10] YAHOO! (June 14, 2011, 03:28:34), http://web.archive.org/web/20110615032834/http://www.yahoo.com/ (accessed through the Internet's Archive's Wayback Machine index).

[11] US WKLY. MAG. (Jan. 1, 2013, 19:20:51), http://web.archive.org/web/20130101192051/http://www.usmagazine.com/ (accessed through the Internet's Archive's Wayback Machine index).

for the browsing of web documents.[12] Alexa Internet, which is an Amazon.com company, "created one of the largest Web crawls, and developed the infrastructure to process and serve massive amounts of data."[13] Since early 1996, Alexa has been crawling the web and "[a]s a service to future historians, scholars, and other interested parties, Alexa Internet donates a copy of each crawl of the web to the Internet Archive . . . ."[14]

Web crawlers are software programs "that surf the Web and automatically store copies of website files, preserving these files as they exist at the point of time of capture."[15] Another description provides: "A crawler or robot is an automated program that scours the Internet and takes pictures of every web page that it is instructed to visit."[16] Alexa Internet has developed such a web crawler and gathers approximately 1.6 terabytes (1,600 gigabytes) of web content per day.[17] Each snapshot of the web takes approximately two months to complete; however, since 1996, Alexa Internet has gathered shots of 4.5 billion web pages from over 16 million websites.[18]

There are various other sources for archived web pages available, including Gigablast,[19] Google's Googlebot,[20] etc.[21] However, the Internet is constantly developing and some of these sources have been discontinued or replaced with newer versions. For example, Yahoo! announced in February 2009 that its archive service, Yahoo! MyWeb, would be discontinued and replaced with Yahoo! Bookmarks and another service, Delicious.[22]

In contrast to the transient nature of some other archiving sites, the Internet Archive's WayBack Machine has been storing images for public

[12] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "How was the Wayback Machine Made?").

[13] *About Alexa Internet*, ALEXA INTERNET, INC., http://www.alexa.com/company (last visited Feb. 10, 2013).

[14] *Technology*, ALEXA INTERNET, INC., http://www.alexa.com/company/technology (last visited Feb. 10, 2013) [hereinafter *Alexa Technology*].

[15] *Standard Affidavit*, INTERNET ARCHIVE, http://www.archive.org/legal/affidavit.php (last visited Feb. 10, 2013).

[16] Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 631 (E.D. Pa. 2007).

[17] *Alexa Tech.*, *supra* note 14.

[18] *Id.*

[19] GIGABLAST.COM, http://www.gigablast.com/ (last visited Feb. 10, 2013).

[20] *Webmaster Tools Help: Googlebot*, GOOGLE.COM, http://www.google.com/support/webmasters/bin/answer.py?answer=182072 (last visited Feb. 10, 2013).

[21] *See* Matthew Fagan, *"Can You Do a Wayback on That?" The Legal Community's use of Cached Web Pages in and out of Trial*, 13 B.U. J. SCI. & TECH. L. 46, 51-56 (2007).

[22] Elinor Mills, *Yahoo MyWeb Bites the Dust*, CNET NEWS (Feb. 13, 2009, 1:41 PM), http://news.cnet.com/8301-1023_3-10163931-93 html; DELICIOUS, http://delicious.com/ (last visited Feb, 10, 2013).

use since 1996.[23] Many of the other services were started after 1996 and do not retain as many archived copies as the Wayback Machine.[24] For example, Gigablast was founded in 2000[25] and appears to only retain one cached or archived copy of each web page. A search for Yahoo!'s cached pages only results in a screen shot from November 6, 2011. Interestingly, a link next to the cached date is entitled "older copies" and links directly to the Wayback Machine's date specific search results.[26] Therefore, some of the other archiving services available on the web even rely on the Wayback Machine for more dated screen shots.

Additionally, the Internet Archive recently expanded its tracking capabilities when Google announced that its HTTP Archive has merged with the Internet Archive.[27] While the WayBack Machine tracks the content of the web, Google's HTTP Archive tracks how the content is built and served.[28] The HTTP Archive has therefore joined the Internet Archive's mission of recording history for future generations.[29] HTTP Archive's merger with the Internet Archive also highlights the importance of the WayBack Machine for the rest of the Internet industry.

The Internet Archive also recently released a new and improved version of the Wayback Machine.[30] Wayback Machine Beta includes more archived web pages than ever before and features a new calendar view that simplifies searches.[31] One information industry analyst

[23] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "Who was Involved in the Creation of the Internet Archive Wayback Machine?").

[24] *E.g.*, *About Us*, GIGABLAST.COM, http://www.gigablast.com/about html (last visited Feb. 10, 2013).

[25] *Id.*

[26] *See* YAHOO!, http://www.gigablast.com/search?k1l=453388&q=yahoo; http://wayback.archive.org/web/*/http://www.yahoo.com/ (accessed by searching for "www.yahoo.com" in the Gigablast index and choosing "cached," which in turn provided access to the Internet Archive's Wayback Machine index) (last visited Feb. 10, 2013).

[27] Cade Metz, *Google's HTTP Archive Merges with Internet Archive*, THE REGISTER (June 15, 2011, 19:42 GMT), http://www.theregister.co.uk/2011/06/15/http_archive_teams_with_new_relic/.

[28] *Id.*

[29] Steve Sounders, *Announcing the HTTP Archive*, STEVESOUNDERS.COM BLOG (Mar. 30, 2011, 5:44 PM), http://www.stevesouders.com/blog/2011/03/30/announcing-the-http-archive/; *About*, HTTP ARCHIVE.ORG, http://httparchive.org/about.php (last visited Feb. 10, 2013).

[30] Gary Price, *Internet Archive Releases New Version of the Wayback Machine*, INFORMATION TODAY, INC. (Jan. 31, 2011), http://newsbreaks.infotoday.com/NewsBreaks/Internet-Archive-Releases-New-Version-of-The-Wayback-Machine-73492.asp; Jay Hathaway, *Internet Archive Wayback Machine Introduces New Beta Version with Calendar View*, SWITCHED (Jan. 21, 2011, 8:10 PM), http://downloadsquad.switched.com/2011/01/21/internet-archive-wayback-machine-introduces-new-beta-version-with-calendar-view/; *Updated Wayback Machine in Beta Testing*, WEB ARCHIVING AT ARCHIVE.ORG BLOG (Jan. 24, 2011), http://iawebarchiving.wordpress.com/2011/01/.

[31] *What's the difference between the classic Wayback Machine and the new BETA test*

declared, "This first release of the Wayback Machine beta already provides a 100% improved experience with major improvements both in navigation and in the amount of data available for a specific page."[32]

Therefore, while there are many other means of caching or archiving websites, the WayBack Machine represents one of the most consistent and respected sources in the industry. The Bluebook even uses the Internet Archive's database as an example of the rule for citing to archival Internet sources.[33] Because of the Wayback Machine's impressive database and improved navigation function, it will, if it has not already, become the most relied upon resource for past versions of websites. This respected position across the Internet also presents new responsibilities, including its role in the development of the use of screen shots as evidence.

B. Screen Shots as Evidence

1. Federal Rules of Evidence

The main evidentiary issues that arise when submitting screen shots into evidence relate to hearsay and authentication. Hearsay is defined as a statement that "the declarant does not make while testifying at the current trial or hearing" and "a party offers in evidence to prove the truth of the matter asserted in the statement."[34] An archived screen shot from the Wayback Machine is a written assertion of what a web page looked like on the date provided. However, there are several exceptions to the rule against hearsay that may cover this situation.

First, most courts reject a hearsay objection to testimony generated by a machine, such as a screen read of a clock.[35] Here, the Wayback Machine is archiving screen shots taken by automated web crawlers from the date and time indicated. Therefore, it could be argued that they meet the machine exception to hearsay. Another possible exception may be found in FED. R. EVID. 801(d)(2), which provides that an opposing party's statement offered against the opposing party is not hearsay.[36] This exception could be applied because the contents of the party's website may be deemed an admission.

version?, INTERNET ARCHIVE WAYBACK MACHINE BETA (Dec. 21, 2010), http://faq.web.archive.org/whats-the-difference-between-the-classic-wayback-machine-and-the-new-beta-version/.

[32] Price, *supra* note 30.

[33] THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION R. 18.2.2(h), at 169 (Columbia Law Review Ass'n et al. eds., 19th ed. 2010).

[34] FED. R. EVID. 801(c).

[35] CHRISTOPHER B. MUELLER & LAIRD C. KIRKPATRICK, EVIDENCE UNDER THE RULES 122 (7th ed. 2011).

[36] FED. R. EVID. 801(d)(2).

Additionally, FED. R. EVID. 803(6) provides that a record of "an act, event, condition, opinion, or diagnosis" is admissible as nonhearsay if it is a record that was kept in the course of a regularly conducted activity of an organization whether or not for profit.[37] The making of the record must also be a regular practice of the organization's activity.[38] The Internet Archive is a non-profit entity whose regularly conducted activity is to build an Internet library with permanent access to archived web pages and its regular practice is to archive screen shots obtained for its library.[39] Therefore, screen shots from the Wayback Machine could also fall within this hearsay exception.

However, as FED. R. EVID. 803(6)(D) states, in order to be admissible, all evidence is subject to the requirement of authentication. In order to satisfy the requirement of authentication, FED. R. EVID. 901(a) provides that the proponent of any proffered evidence "must produce evidence sufficient to support a finding that the item is what the proponent claims it is."[40] FED. R. EVID. 901(b) provides examples of some types of evidence that will satisfy the authentication requirement,[41] including the testimony of a witness with knowledge[42] and evidence about a process or system.[43] Additionally, FED. R. EVID. 902 provides instances where evidence is self-authenticating and does not require additional evidence in order to meet the authentication requirement.[44] Under FED. R. EVID. 902(11), evidence is self-authenticating when the original or a copy of the record that meets the requirements of FED. R. EVID. 803(6)(A) to (C) for a regularly conducted business activity is submitted with a certification of the custodian or another qualified person.[45]

An alternative avenue to get screen shots into evidence is to argue that the screen shots fall under the Best Evidence Doctrine. The Best Evidence Doctrine, also known as the original writing or original document rule, requires that a writing's original be provided if it is offered to prove the contents of the writing.[46] FED. R. EVID. 1001(d) provides that "[f]or electronically stored information, 'original' means any printout—or other output readable by sight—if it accurately reflects

[37] *Id.* at 803(6).
[38] *Id.* at 803(6)(c).
[39] *See About the Internet Archive*, *supra* note 3.
[40] FED. R. EVID. 901(a).
[41] *Id.* at 901(b).
[42] *Id.* at 901(b)(1).
[43] *Id.* at 901(b)(9).
[44] *Id.* at 902.
[45] *Id.* at 902(11).
[46] MUELLER & KIRKPATRICK, *supra* note 35, at 881; FED. R. EVID. 1002.

the information."[47] However, because the archived pages may not produce the entire web page, a screen shot could fall under the definition of "duplicate" because it is "a counterpart produced by a[n] . . . . electronic, or other equivalent process or technique that accurately reproduces the original."[48] At least one author has suggested that "the best evidence rule and its allowance for the admission of duplicates provide a superior framework in which to address concerns that may arise with respect to the uniformity between the original Web page and the archived copy."[49] However, thus far, no courts have applied the Best Evidence Doctrine when assessing the authenticity of archived screen shots from the Wayback Machine.

2. Federal Courts' Treatment of Admissibility

Although some courts cite to the Internet Archive's web archive index without comment or hesitation,[50] evidentiary issues still arise.[51] In addressing hearsay objections, federal courts have taken a variety of approaches regarding the admissibility of archived screen shots. Some courts have held that the screen shots are not statements or that they are not offered to prove the truth of the matter asserted and thus are not within the definition of hearsay.[52] Other courts have held that the

[47] FED. R. EVID. 1001(d).

[48] *Id.* at 1001(e); *see* Deborah R. Eltgroth, Note, *Best Evidence and the Wayback Machine: Toward a Workable Authentication Standard for Archived Internet Evidence*, 78 FORDHAM L. REV. 181, 211-12 (2009).

[49] Eltgroth, *supra* note 48, at 212.

[50] *See e.g.*, GoPets Ltd. v. Hise, 657 F.3d 1024, 1028 (9th Cir. 2011) (citing the Internet Archive's records to describe content of the website at issue); Santos *ex. rel.* Beato v. United States, 559 F.3d 189, 201 n.7 (3d Cir. 2009) (quoting from a healthcare provider's archived website for what defendant could have known); Moorish Sci. Temple of Am. 4th & 5th Generation v. Super. Ct. of N.J., 2012 WL 123405, at *2 n.2 (D.N.J. 2012) (citing an article regarding the concept of "sovereign citizenship" in discussion of the Moorish movement); Pounds v. Katy Indep. Sch. Dist., 730 F. Supp. 2d 636, 640 n.2 (S.D. Tex. 2010) (citing the archived website of the service provider of holiday cards featuring children's artwork, which was at issue in the case); Am. Casino & Entm't Props, LLC v. Marchex Sales, Inc., No. 2:12-cv-01054-GMN-VCF, 2012 WL 2674611, at *4 (D. Nev. July 5, 2012) (considering a Wayback Machine screen shot to determine if the use of a domain name constitutes sufficient use of a trademark to rebut the presumption of ownership); Parsi v. Daioleslam, No. 08-705(JDB), 2012 WL 4017720, at *3 (D.D.C. Sept. 13, 2012) (citing to the Internet Archive's archived version of a website because the original URL now directs to an unrelated website).

[51] *See e.g.*, Market-Alerts Pty. Ltd. v. Bloomberg Fin. L.P., 2013 WL 443973, at *4 n. 12 (D. Del. Feb. 5, 2013) (noting that courts have reached varying conclusions regarding the reliability of documents generated by the Wayback Machine).

[52] *See e.g.*, Foreword Magazine, Inc. v. OverDrive, Inc., 2011 WL 5169384, at *4 (W.D. Mich. 2011) ("When a printout from a third party website is offered merely to show that certain images and text appeared on the website, they are not statements at all and thus fall outside the ambit of the hearsay rule"); Telewizja Polska USA, Inc. v. Echostar Satellite Corp., 2004 WL 2367740, at *5 (N.D. Ill. 2004) (citing Perfect 10, Inc. v. Cybernet Ventures, Inc., 213 F. Supp. 2d 1146, 115 (C.D. Cal. 2002) that "[t]o the extent these images and text are

archived versions of the websites are party admissions.[53] A few courts have held that screen shots are not admissible under any hearsay exception.[54] However, for the basic purpose of understanding when archived screen shots are likely to be admitted, most courts admit such evidence under a hearsay exception if properly authenticated.

Regarding the authentication requirement, most federal courts have concluded that an affidavit from the Internet Archive is sufficient to authenticate screen shots from the Wayback Machine.[55] The two circuits whose district courts have been the most skeptical of the admissibility of archived screen shots are the Fourth Circuit and the Second Circuit.

In the Fourth Circuit, the United States District Court for the District of Maryland discussed the authentication methods for electronic evidence at length in its opinion in *Lorraine v. Markel American Insurance Co.*[56] In *Lorraine*, the court ultimately held that counsel failed to establish the authenticity of their exhibits, resolve potential hearsay issues, and meet several other evidentiary obligations.[57] This established an incredibly strict standard to admit electronic evidence, including archived screen shots.

In 2010, the District Court of Maryland spoke again on the issue by granting the defendants' motion in limine to exclude plaintiff's exhibits downloaded from the Wayback Machine insofar as defendants may not

being introduced to show the images and text found on the websites, they are not statements at all—and thus fall outside the ambit of the hearsay rule").

[53] *See e.g., In re* Hydroxycut Mktg. & Sales Practices Litig., 810 F. Supp. 2d 1100, 1115 (S.D. Cal. 2011) (overruling hearsay objection because the archived web pages were deemed authentic representations of the website at issue in a related case and therefore are party admissions); Telewizja Polska USA, Inc. v. Echostar Satellite Corp., 2004 WL 2367740, at *5 (N.D. Ill. 2004) (holding that the contents of the website may be considered an admission of a party-opponent and are not barred by the hearsay rule).

[54] *See e.g.,* Novak v. Tucows, Inc., 2007 WL 922306, at *5 (E.D.N.Y. 2007) (holding that the managers of Internet Archive do not ensure that the material posted accurately represents the previous versions of the websites, and therefore, the screen shots could not be authenticated by such testimony), *aff'd*, 330 Fed. Appx. 204 (2d Cir. 2009).

[55] *See e.g.,* St. Luke's Cataract & Laser Inst., P.A. v. Sanderson, 2006 WL 1320242, at *2 (M.D.Fla. 2006) (holding that printouts from a website could be authenticated by "a statement or affidavit from an Internet Archive representative with personal knowledge of the contents of the Internet Archive website"), *aff'd*, 573 F.3d 1186 (11th Cir. 2009); Keystone Retaining Wall Sys., Inc. v. Basalite Concrete Prods., LLC, 2011 WL 6436210, at *9 n.9 (D. Minn. 2011) (citing cases that accepted affidavits as sufficient to authenticate for the proposition that "[t]he Internet Archive has existed since 1996, and federal courts have regularly accepted evidence from the Internet Archive"); Mahmood v. Research in Motion Ltd., 2012 WL 242836, at *4 n.2 (S.D.N.Y. 2012) (rejecting plaintiff's argument that the website snapshot is "'not authenticated'" by citing the affidavit of the Office Manager of the Internet Archive); Telewizja Polska USA, Inc. v. Echostar Satellite Corp., 2004 WL 2367740, at *6 (N.D. Ill. 2004) (admitting evidence from the Internet Archive as properly authenticated by an affidavit from the Internet Archive).

[56] Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 537-83 (D. Md. 2007).

[57] *Id.* at 585.

make any reference to the Wayback Machine or reference any dates provided by the Wayback Machine.[58] The court also denied the motion in part, holding that the defendants may show website documents that can be authenticated by witnesses who can testify that they in fact viewed the relevant website at a particular time.[59] Although the court's order does not reveal much about the plaintiff's authentication attempts in this case, it does illustrate that this court is requiring a much higher standard than is required in most federal courts.

In the Second Circuit, the United States District Court of the Eastern District of New York pointed out in *Novak v. Tucows, Inc.* that "the authorized owners and managers of the archived websites play no role in ensuring that the material posted in the Wayback Machine accurately represents what was posted on their official websites at the relevant time."[60] It held that because the proffering party offered "neither testimony nor sworn statements attesting to the authenticity of the contested web page exhibits by any employee of the companies hosting the sites from which plaintiff printed the pages," the exhibits could not be authenticated.[61]

However, in 2011, the Eastern District Court of New York adopted the Report & Recommendation of Magistrate Judge Reyes regarding patent claim construction.[62] A footnote in the Report & Recommendation mentions that although litigants seeking to submit archived images from the Wayback Machine as evidence run into authentication problems, "the federal courts are familiar with this internet resource."[63] This footnote arose after the court stated that sample websites from the years 1996 and 1998 supported Google's proposed construction of a patent.[64] Although the court does not discuss the potential authentication issues while considering the archived websites as extrinsic evidence during its claim construction analysis, this case does suggest that the court may be moving toward greater acceptance of this form of evidence.

The United States District Court for the Southern District of New York also expressed distrust of this form of evidence in *Chamilia, LLC v. Pandora Jewelry, LLC*.[65] In *Chamilia*, the court mentioned in a footnote

[58] Schwartz v. J.J.F. Mgmt. Servs., Inc., 2010 WL 1529241, at *1 (D. Md. 2010).

[59] *Id.*

[60] Novak v. Tucows, Inc., 2007 WL 922306, at *5 (E.D.N.Y. 2007), *aff'd*, 330 Fed. Appx. 204 (2d Cir. 2009).

[61] *Id.*

[62] Web Tracking Solutions, LLC v. Google, Inc., 2011 WL 3418311, at *4 (E.D.N.Y. 2011).

[63] Web Tracking Solutions, LLC v. Google, Inc., 2011 WL 3418323, at *16 n.25 (E.D.N.Y. 2011), *report and recommendation adopted*, 2011 WL 3418311 (E.D.N.Y. 2011).

[64] *Id.* at *16.

[65] *See* Chamilia, LLC v. Pandora Jewelry, LLC, 2007 WL 2781246, at *6 n.4 (S.D.N.Y. 2007).

that the plaintiff submitted a series of web pages from the Wayback Machine to support its claim; however, the court, citing *Novak*, held that "[t]his putative evidence suffers from fatal problems of authentication under FED. R. EVID. 901."[66] However, *Chamilia* was not the court's last word regarding the authentication of archived screen shots. In *Mahmood v. Research in Motion Ltd.*, the District Court for the Southern District of New York rejected the plaintiff's argument that the website snapshot used by the defendant was not authenticated by citing the affidavit of the Office Manager of the Internet Archive.[67] Therefore, as recently as last year, the court has aligned itself with the majority of federal courts by holding that an affidavit from the Internet Archive is sufficient to authenticate archived screen shots.

There may also be some movement to permit other means of authenticating archived screen shots. In *United States v. Bansal*, the Third Circuit affirmed the admission of archived screen shots from the Wayback Machine based upon the testimony of a witness regarding how the Wayback Machine website works and the reliability of its contents.[68] The court noted that the witness, based upon her personal knowledge, concluded from her review of previously authenticated and admitted images from the same website that the screen shots were authentic.[69] While the Third Circuit is in the minority with this less strict approach to authentication of screen shots from the Wayback Machine, along with the apparent acquiescence of the courts of the Second Circuit, it appears that when presented with a proper affidavit or testimony, the authentication of archived screen shots is not an issue for most federal courts.

Although this review does not sufficiently explain the approach of each court within each circuit,[70] it attempts to demonstrate the potential evidentiary issues that may arise and the general consensus regarding the admissibility of archived screen shots. When archived screen shots are supported by an affidavit from the Internet Archive, most federal courts will admit those images into evidence under some variation of a hearsay exception or because the evidence does not fall within the definition of hearsay. However, each federal court's approach may vary and therefore further considerations must be made before concluding that any screen shots will be admitted.

[66] *Id.*

[67] Mahmood v. Research in Motion Ltd., 2012 WL 242836, at *4 n.2 (S.D.N.Y. 2012).

[68] United States v. Bansal, 663 F.3d 634, 667-68 (3d Cir. 2011).

[69] *Id.*

[70] *See e.g.*, Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 537-83 (D. Md. 2007) (discussing the federal rules of evidence regarding electronic evidence); Eltgroth, *supra* note 48, at 202-11 (discussing the evolution of authentication standards).

3. Internet Archive's Authentication Policies

The Internet Archive specifically states that "[t]he Wayback Machine tool was not designed for legal use."[71] However, the Internet Archive provides a legal page, standard affidavit, and a frequently asked questions page specifically for lawyers.[72] As described above, an affidavit from someone at the Internet Archive describing how the Wayback Machine's process for compiling and storing previous websites is almost a prerequisite to admissibility.[73] This legal requirement has made the legal resources web pages on the Internet Archive's website necessary.

The Internet Archive's policy begins by stating that it is "not in the business of responding to requests for affidavits, or authenticating pages or other information from the Wayback Machine."[74] It also highlights that they are a nonprofit with limited resources. However, if an affidavit authenticating printouts is absolutely necessary, you must send an electronic list of the extended URLs for each page you would like to be printed out.[75] It is crucial that the pages be properly attached to the affidavit in order to be considered admissible.[76] The standard fee of $250 per request, plus an additional $10 for each URL listed, must be sent along with the list of URLs to be included with the affidavit.[77] The Internet Archive also charges an additional $100 fee for the affidavit to be notarized,[78] which is highly recommended if not required for the affidavit to be valid.[79] Finally, any URLs with printable files cost an

[71] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "How can I get pages authenticated from the Wayback Machine? How can [I] use the pages in court?")

[72] *Id.*

[73] *Compare, e.g.*, Specht v. Google Inc., 758 F. Supp. 2d 570, 580 (N.D. Ill. 2010) (holding that it was an improper authentication method when not authenticated by an officer or employee of the Internet Archive), *and* Sam's Riverside, Inc. v. Intercon Solutions, Inc., 790 F. Supp. 2d 965, 981 (S.D. Iowa 2011) (admitting the screen shots attached to the Internet archive's employee's affidavit but excluding those that were not attached), *with, e.g.*, Mortg. Mkt. Guide, LLC v. Freedman Report, LLC, 2008 WL 2991570, at *12 n.3 (D. N.J. 2008) (permitting the use of archived web pages from an archive company without an affidavit because the content of the archived web pages was not a controlling issue in the case).

[74] *Legal: Information Requests*, INTERNET ARCHIVE, http://www.archive.org/legal/ (last visited Nov. 19, 2012).

[75] *Id.*

[76] *See, e.g.*, *Sam's Riverside, Inc.*, 790 F. Supp. 2d at 981 (admitting the screen shots attached to the Internet archive's employee's affidavit but excluding those that were not attached).

[77] *Legal: Information Requests*, *supra* note 74.

[78] *Id.*

[79] *See Affidavit*, BLACK'S LAW DICTIONARY 66 (9th ed. 2009) ("A voluntary declaration of facts written down and sworn to by the declarant before an officer authorized to administer oaths.") *and Notary Public*, BLACK'S LAW DICTIONARY 1161 (9th ed. 2009) ("A person authorized by a state to administer oaths . . . ."); FED. R. EVID. 902(11).

additional $30 per URL.[80] Payment is required before the Internet Archive will begin working on any request.[81] Notably, there is no refund or credit for requests with mistakes or incorrect URLs.[82] The Internet Archive usually takes about five business days from receipt of payment to return a request, but it may take longer if there are a large number of URLs and the request must be limited.[83] Requests cannot be expedited.[84] Finally, they "reserve the right to refuse any request that [they] deem to be unreasonable, and may require you to reimburse other costs [they] incur as a result of [the] request."[85]

Once the request is processed, they will return the printed screen shots and a standard affidavit of authenticity.[86] The affidavit describes how the Wayback Machine operates and the format of its printouts.[87] The Internet Archive may be willing to change its standard affidavit on a case-by-case basis but requires compensation for any attorney's fees incurred due to reviewing the changes.[88] The Internet Archive's final admonition is that the affidavit "only affirms that the printed document is a true and correct copy of our records. It remains your burden to convince the finder of fact what pages were up when."[89]

All of the Internet Archive's legal resources highlight the fact that an affidavit may not be necessary in a legal proceeding.[90] Before seeking authentication from the Internet Archive, it requests that individuals requesting an affidavit "seek judicial notice or simply ask [the] opposing party to stipulate to the documents' authenticity."[91] Additionally, they recommend contacting "the party who posted the information on the URLs at issue" or "someone who actually accessed the historical versions of the URLs."[92] However, as noted above, the party that posted the information may be adverse to supplying it to a litigating party,[93] and

[80] *Legal: Information Requests*, *supra* note 74.

[81] *Id.*

[82] *Legal: Frequently Asked Questions*, INTERNET ARCHIVE, http://www.archive.org/legal/faq.php (last visited Feb. 10, 2013) [hereinafter *Legal: FAQ*] (scroll to "I submitted an incorrect request, can I have a refund or a credit?").

[83] *Legal: Information Requests*, *supra* note 74.

[84] *Legal: FAQ, supra* note 82 (scroll to "My request is urgent! Can the Internet Archive provide the documents and affidavit immediately?").

[85] *Legal: Information Requests*, *supra* note 74.

[86] *Id.*; *see Standard Affidavit*, *supra* note 15.

[87] *Standard Affidavit*, *supra* note 15.

[88] *Legal: FAQ*, *supra* note 82 (scroll to "Can the Internet Archive change its standard affidavit to fit my needs?").

[89] *Id.* (scroll to "Does the Internet Archive's affidavit mean that the printout was actually the page posted on the Web at the recorded time?")

[90] *See Legal: Information Requests*, *supra* note 74; *Standard Affidavit*, *supra* note 15.

[91] *Legal: Information Requests*, *supra* note 74.

[92] *Id.*

[93] *See, e.g.*, Netbula, LLC v. Chordiant Software, Inc., 2009 WL 3352588, *1 (N.D.

someone who accessed the historical versions of the URLs via the Wayback Machine has been inadequate to authenticate the screen shots in previous cases.[94] Nevertheless, the Internet Archive's initial recommendation to seek judicial notice or a stipulation of the parties is excellent advice that should be pursued before requesting an affidavit from the Internet Archive.

II. BENEFITS AND PROBLEMS WITH USE OF THE WAYBACK MACHINE IN LITIGATION

A. Benefits

The Wayback Machine can prove an invaluable resource in proving a case at trial. Through its vast database of archived web pages, an individual can find an exact replica of a specific site on certain listed dates. Perhaps the best way to illustrate the potential benefits of use of the Wayback Machine is through a hypothetical trademark infringement litigation. If a website owner used another's registered trademark on his or her website, the trademark owner could search the Wayback Machine for the website owner's web page and request authenticated copies from the Internet Archive to show not only actual use but also the time period of use to aid in calculation of damages. Additionally, the trademark owner may be able to demonstrate the potential for consumer confusion by comparison to the use of the trademark by the actual owner.

The Third Circuit case *Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey* provides an excellent real-world illustration of how the Wayback Machine might help assess and defend a potential trademark infringement suit. In *Healthcare Advocates*, the patient advocacy organization alleged that a law firm unlawfully obtained copyrighted material when the firm used the Wayback Machine to access screenshots of Healthcare Advocates' website.[95] The Third Circuit held that screen shots were not unlawfully obtained and granted summary judgment for the Harding firm on that issue. However, more importantly to the point at issue, the Third Circuit stated that "[v]iewing the content that Healthcare Advocates had included on its public website in the past was very useful to the Harding firm in assessing the merits of the trademark infringement and trade secret misappropriation claims brought

Cal. Oct. 15, 2009).

[94] *See, e.g.*, Specht v. Google Inc., 758 F. Supp. 2d 570, 580 (N.D. Ill. 2010) (holding that it was an improper authentication method when not authenticated by an officer or employee of the Internet Archive).

[95] Healthcare Advocates v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 630 (E.D. Pa. 2007).

against their clients."[96] Therefore, in a case brought against individuals for trademark infringement, the Wayback Machine proved useful in assessing the validity of the claims.

B. Problems

Although it can be advantageous to have the data collected by the Wayback Machine to bolster arguments in litigation, there are several complications that can make it difficult to use effectively. Perhaps most importantly, the database only includes screen shots that have been generated randomly by Alexa Internet's web crawler. Therefore, only screen shots from sporadic dates and times may be available. Unfortunately, those screen shots available may not be helpful to the case because they may not meet the time period in question.

For example, the United States District Court for the Southern District of Iowa concluded that Wayback Machine screen shots were inadmissible in litigation alleging trademark infringement because the shots were taken on or after December 4, 2004, and "the alleged infringement began no later than January 2004."[97] Therefore, because the screen shots were taken significantly after Sam's Riverside terminated its relationship with Intercon, the court determined they were "not relevant to the issue of whether [Sam's Riverside] established protectable rights in the phrase 'Sam's Riverside' prior to the commencement of the alleged infringement."[98]

Another difficulty that the Wayback Machine's format presents is that it is not keyword searchable. The Internet Archive's website explains:

> The Wayback Machine is not like a typical search engine in that it cannot search for specific terms or keywords. Therefore, the Internet Archive cannot respond to requests such as: "All records containing the term 'Prelinger Archives'" or "All records related to the Web site www.archive.org." Instead, provide a list of the extended URLs for each page on the Wayback Machine that you want us to authenticate.[99]

Therefore, using a trademark infringement issue as an example, in order to search for potentially infringers, a trademark owner must first know of the web pages where the allegedly infringing mark appears.

[96] *Id.* at 630.

[97] Sam's Riverside, Inc. v. Intercon Solutions, Inc., 790 F. Supp. 2d 965, 982 (S.D. Iowa 2011).

[98] *Id.*

[99] *Legal: FAQ*, *supra* note 82 (scroll to "Can the Internet Archive search for pages on the Wayback Machine using particular keywords or other search terms?").

However, the Internet Archive hopes to "implement a full text search engine at some point in the future."[100] Additionally, once the web page of a potential infringer is discovered, it may or may not be available in the Wayback Machine's database. If the automated web crawlers are not aware of the existence of the web page, they will not generate screen shots of the specific web page.

Finally, websites archived on the Wayback Machine may have limited functionality. "When a dynamic page contains forms, JavaScript, or other elements that require interaction with the originating host, the archive will not contain the original site's functionality."[101] Also, the archived web pages are intended to be a "snap shot" of the website, and therefore some of the "images or links might be missing."[102] Files over 10 MB are also not included in the archived version of a website.[103] Therefore, some relevant and important information may not be obtainable even if the web page has been archived. As an example, a search of http://www.yahoo.com will generate several screen shots of the home page on several dates.[104] After selecting one of the crawls from a specific date, it will display a plethora of news stories for the date in question; however, only some of the listed links will generate the actual story. By way of example, even if there appears to be a link to an article regarding the dispute over Dubai's "sinking" islands, when the link is clicked, it does not display the actual article because it was not archived.[105] In other words, only if the article itself was crawled by Alexa Internet will it also appear in the archive.

While the Wayback Machine may prove useful in generating evidence for litigation, it often cannot be relied upon because of its limitations. There is no guarantee that the web page in question will be included in the archived records and, even if it is, it may not cover the time period at issue. Also, the Wayback Machine will not help companies discover potential past infringers because it does not have a keyword search option. However, despite these limitations, the Wayback

[100] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "Can I search the Archive?").

[101] *Id.* (scroll to "How do you archive dynamic pages?"); *See e.g.*, YOU TUBE Jun. 15, 2011, 16:10:24), http://web.archive.org/web/20110615032834/http://www.youtube.com/ (accessed through the Internet's Archive's Wayback Machine index).

[102] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "Where is the rest of the archived site? Why am I getting broken or gray images on a site?").

[103] *Id.*

[104] YAHOO!, http://wayback.archive.org/web/*/http://www.yahoo.com (last accessed on Feb. 10, 2013 by searching for "www.yahoo.com" in the Internet Archive's Wayback Machine index).

[105] YAHOO! (Jan. 29, 2011, 01:57:20), http:// http://web.archive.org/web/20110129015720/http://www.yahoo.com/ (accessed through the Internet's Archive's Wayback Machine index and then clicking "Dispute over Dubai's 'sinking' islands").

Machine is an invaluable resource that can be immensely helpful in proving a case at trial, even if used by the other side.

III. WAYS TO PREVENT ACCESS AND RECOMMENDATIONS

A. *Internet Archive's Policies*

The Internet Archive declares that it "is not interested in offering access to Web sites or other Internet documents whose authors do not want their materials in the collection."[106] The Internet Archive also follows the Oakland Archive Policy for Managing Removal Requests and Preserving Archival Integrity.[107] This policy was created in 2002 to address the possibility that authors and publishers may request that their documents not be included in publicly available archives or web collections.[108] The policy details recommendations for the proper responses for particular removal requests, including requests based on intellectual property claims.[109] The Internet Archive provides that "[w]hen a URL has been excluded at direct owner request from being archived, that exclusion is retroactive and permanent."[110] Instructions on how to exclude a website from the archive are detailed on the Internet Archive's website[111] as well as in an addendum to the Oakland Archive Policy.[112]

When a website owner contacts the Internet Archive directly to request that they stop crawling or archiving a site, the Internet Archive "endeavor[s] to comply with these requests."[113] However, within the Internet Archive's exclusion instructions, it advises website owners to read the Oakland Archive Policy to determine if it applies before sending a request.[114] Therefore, requested removal is limited to the categories and corresponding policies detailed in the Oakland Archive Policy.[115]

However, individualized requests are not the Internet Archive's

[106] *Removing Documents From the Wayback Machine*, INTERNET ARCHIVE, http://www.archive.org/about/exclude.php (last visited Feb. 10, 2013) [hereinafter *Removing Documents*].

[107] *The Oakland Archive Policy*, UNIV. OF CAL., BERKELEY INFO. MGMT. AND SYS. (Dec. 13-14, 2002), http://www2.sims.berkeley.edu/research/conferences/aps/removal-policy html.

[108] *Id.*

[109] *Id.*

[110] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "Some sites are not available because of robots.txt or other exclusions. What does that mean?").

[111] *Id.*

[112] *The Oakland Archive Policy*, *supra* note 107.

[113] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "Some sites are not available because of robots.txt or other exclusions. What does that mean?").

[114] *Removing Documents*, *supra* note 106.

[115] *See The Oakland Archive Policy*, *supra* note 107.

primary vehicle for removing content; instead, it recommends that a website owner place a robots.txt file at the root of the website's domain.[116] The robots.txt file, which will be further explained below, will remove all the domain's documents from the Wayback Machine and it will tell web crawlers not to crawl the site in the future.[117] The Internet Archive also offers information on how to exclude the Internet Archive's crawler only.[118] However, it is crucial to understand how the robots.txt file works and if the content exclusion it provides is really "retroactive and permanent."

B. Robots.txt

"Web robots . . . are programs that traverse the Web automatically" to collect information.[119] Web owners can use a robots.txt file to give instructions to web robots that visit its site.[120] These instructions are called "The Robots Exclusion Protocol."[121] Therefore, by installing the robots.txt file, when web robots, such as those employed by Alexa Internet, want to crawl the website, they will first check for the robots.txt file instructions and may observe the Robots Exclusion Protocol, which instructs the web robot that it is disallowed from visiting any pages on the website.[122]

While a robots.txt file appears to provide a simple solution to the unwanted archiving of web content on the WayBack Machine, there are several important considerations. First, web robots can ignore the robots.txt file.[123] The robots.txt standard was originally created in 1994 with "A Standard for Robot Exclusion" document, representing a consensus of robot authors and others interested in robots.[124] This document expressly states:

> [Robots.txt] is not an official standard backed by a standards body, or

[116] *Removing Documents*, *supra* note 106.

[117] *Id.*

[118] *Id.*

[119] *The Web Robots Pages*, ROBOTSTXT.ORG (Aug. 23, 2010, 19:18:05), http://www robotstxt.org/.

[120] *About robotstxt.org*, ROBOTSTXT.ORG (Aug. 23, 2010, 19:18:05), http://www robotstxt.org/robotstxt html [hereinafter *About Robotstxt.org*].

[121] *Id.*

[122] *See id.*; *Webmaster Tools: Block or remove pages using a robots.txt File*, GOOGLE.COM, http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=156449&from=35237&rd=1 (last visited Feb. 10, 2013) [hereinafter *Block or Remove Pages Using a Robotos.txt File*].

[123] *About Robotstxt.org*, *supra* note 120.

[124] Martijn Koster, *A Standard for Robot Exclusion*, ROBOTSTXT.ORG (Aug. 23, 2010, 19:18:05), http://www robotstxt.org/orig html.

> owned by any commercial organization [sic]. It is not enforced by anybody, and there [is] no guarantee that all current and future robots will use it. Consider it a common facility the majority of robot authors offer the WWW community to protect WWW server against unwanted accesses by their robots.[125]

Therefore, there is no enforcement or incentive for robots to adhere to the file's instructions. However, Alexa Internet, the company that crawls the web for Internet Archive, does respect robots.txt instructions and even does so retroactively.[126]

Second, the robots.txt file is a publicly available file, which means that anyone can see what sections of the server the website owner does not want robots to access.[127] When "/robots.txt" is added to a URL, anyone can observe what that website owner has decided to exclude from web robots. For example, the University of Colorado Law School URL is "www.colorado.edu/law" when "/robots.txt" is added to the end of the URL in a browser's website field, it displays what is disallowed.[128] Unfortunately, it can seem complicated to translate this computer language but it can easily be done by someone with intermediate computer knowledge. Another example is from Colby College, a small liberal arts college in Waterville, Maine. Colby's URL is "www.colby.edu" but with "/robots.txt" attached, it reveals that Colby has requested that web robots do not crawl the Alumni Section of the Spring 2004 Magazine.[129] Therefore, by adding this file to a URL, almost anyone can see exactly what web robots have been requested not to crawl. Obviously, this might not necessarily be an issue if it is something as innocent as the Alumni Section of a college's magazine, but, from a litigation standpoint, it may encourage inquiry into why this particular web page has been requested to be inaccessible to web robots.

Third, the robots.txt file must be properly entered into the top-level directory of the web server or it will not work.[130] As the Robotstxt.org resource explains, "When a robot looks for the '/robots.txt' file for URL, it strips the path component from the URL (everything from the first single slash) and puts '/robots.txt' in its place."[131] Additionally, where

[125] *Id.*

[126] *The Wayback Machine: FAQ*, *supra* note 2 (scroll to "Some sites are not available because of robots.txt or other exclusions. What does that mean?").

[127] *About Robotstxt.org*, *supra* note 120.

[128] UNIV. OF COLO. LAW SCH., http://www.colorado.edu/law/robots.txt (last visited Feb. 10, 2013) (accessed by adding "robots.txt" to the University of Colorado Law School website address).

[129] COLBY COLL., http://www.colby.edu/robots.txt (last visited Feb, 10, 2013) (accessed by adding "robots.txt" to the Colby College website address).

[130] *About Robotstxt.org*, *supra* note 120.

[131] *Id.*

the file should be placed depends on the web server software used.[132]

Chief Judge Alex Kozinski of the United States Court of Appeals for the Ninth Circuit presents a cautionary tale for the importance of properly implementing a robots.txt file.[133] The Judicial Council of the Third Circuit concluded that Chief Judge Kozinski's retention of a sexually explicit email on the subdirectory of his personal computer, which was publically accessible, was in violation of the Code of Conduct for United States Judges.[134] On June 11, 2008, the *Los Angeles Times* published an article exposing the presence of sexually explicit matter on the Chief Judge's public website.[135] These sexually explicit materials were personal files that were accessible through the use of web server software.[136] The Chief Judge was advised to protect his web server from spiders or crawlers, and his adult son, a computer hobbyist, suggested and installed a robots.txt file.[137] However, due to a miscommunication or configuration error when his son installed the file, it was installed only in the subdirectory created for the group of judges and did not protect the rest of the server's directory.[138] Therefore, the Chief Judge's subdirectory was indexed by Yahoo! and was available through other Internet search engines for public access.[139] In conclusion, if the Chief Judge had instead relied on a computer expert to properly install the robots.txt file, he would have saved the federal judiciary and himself a great deal of embarrassment.[140]

Fourth, website owners can be compelled to disable the robots.txt file. Once the robots.txt file is disabled, previously crawled web content will again be accessible via the Wayback Machine and future crawls will be permitted.[141] In *Netbula, LLC v. Chordiant Software, Inc.*, the District Court for the Northern District of California granted Chordiant's motion to compel Netbula to allow access to Netbula's past web pages that had previously been archived by Internet Archive.[142]

At issue in the case was whether Chordiant had an express or implied license to use Netbula's software and if the damages claimed

[132] *Id.*

[133] *In re* Complaint of Judicial Misconduct, 575 F.3d 279, 288-89 (3d Cir. 2009).

[134] *Id.* at 284.

[135] *Id.* at 280.

[136] *Id.* at 286.

[137] *Id.* at 288-89.

[138] *Id.* at 289.

[139] *Id.*

[140] *See id.* at 289.

[141] *See* Healthcare Advocates v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 632 (E.D. Pa. 2007).

[142] Netbula, LLC v. Chordiant Software, Inc., No. C08–00019 JW (HRL), 2009 WL 3352588, at *2 (N.D. Cal. Oct. 15, 2009).

were accurately based on past pricing.[143] Because there was no dispute as to relevance and Chordiant was seeking discovery of Netbula's past web pages, the request to compel fell within relief authorized by FED. R. CIV. P. 37.[144] Additionally, the court held that Netbula's claim that it lacked "control" over the information because it does not control the Internet Archive's "archiving activities" was unfounded because the issue was over control of the access, which Netbula unilaterally blocked by employing a robots.txt file on its website.[145] Therefore, because Netbula had control and the request fell within relief authorized by FED. R. CIV. P. 37, the court ordered Netbula to disable its website's robots.txt file for a period of two weeks to allow Chordiant to inspect and copy any relevant documents from past versions of Netbula's website that were available through the Internet Archive.[146]

Interestingly, Netbula also argued that Chordiant could access the information from Internet Archive directly through the use of a FED. R. CIV. P. 45 subpoena, but the court concluded that such retrieval would be extremely burdensome, expensive, and disruptive to Internet Archive's operations, whereas Netbula could allow access in a matter of minutes.[147] Therefore, even if a party could claim it was unable to disable its robots.txt file, it appears that that a FED. R. CIV. P. 45 subpoena could be an alternative option to gain access. Additionally, in *Verigy US, Inc. v. Mayder*, despite the fact that the defendant's account with its former Internet service provider, Network Solutions, had been suspended for nonpayment, the District Court of the Northern District of California ordered the defendant to contact Network Solutions to obtain access to documents that were previously available through the Wayback Machine.[148]

Fifth, the Internet Archive's servers that block the excluded content can malfunction. While this might not be a common occurrence, it was relevant in at least one recent case.[149] In *Healthcare Advocates*, the first count of Healthcare Advocates's Second Amended Complaint alleged that Harding violated the Digital Millennium Copyright Act (DMCA) by circumventing Healthcare's security measures employed to prevent

[143] *Id.* at *1.

[144] *Id.*

[145] *Id.*; Defendant's Notice of Motion and Motion to Compel Plaintiffs to Allow Access to Archived Web Pages at 13, Netbula, LLC v. Chordiant Software, Inc., No. C08–00019 JW, 2009 WL 3352588, at *2 (N.D. Cal. Oct. 15, 2009), 2009 WL 3462369.

[146] *Netbula*, 2009 WL 3352588, at *1-2.

[147] *Id.* at *2.

[148] Order Granting in Part and Denying in Part Plaintiff's Motion to Compel Documents at 2-3, Verigy US, Inc. v. Mayder, No. C07-04330 RMW (HRL), 2008 WL 4786621 (N.D. Cal. Oct. 30, 2008).

[149] *See* Healthcare Advocates v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 632 (E.D. Pa. 2007).

access to Healthcare's copyrighted material.[150] The Harding firm was looking for information about Healthcare Advocates and used the Wayback Machine to access screenshots of what Healthcare Advocates' public website looked like prior to the filing of the complaint.[151] The screenshots were very useful in assessing the merits of the claims brought against the firm's clients.[152] Healthcare Advocates alleged that the Harding firm's use of the Wayback Machine to obtain archive screenshots constituted "hacking," because it had a robots.txt file in place to prevent access to the archived screenshots of www.healthcareadvocates.com.[153] However, on July 9, 2003 and July 14, 2003, when the Harding firm accessed the Wayback Machine, Internet Archive's servers that checked for the robots.txt file and blocked the images was malfunctioning and provided the archived images to those who requested them.[154] The court granted summary judgment for the Harding firm on the DMCA claim, because the Harding firm only received the archived screenshots because of a malfunction in the Internet Archive's servers and "[m]aking requests for archived images via the Wayback Machine, even after some requests were denied, is not avoiding or bypassing the [robots.txt] measure.[155] Therefore, although the Internet Archive provides that exclusion will be "retroactive and permanent," there can be occasions when its servers do not work properly and provide an unexpected exception.

Finally, only a server administrator can include the robots.txt file.[156] Therefore, the individual website maintainers using the server cannot make this change.[157] However, this problem can usually be easily remedied by contacting the server administrator, who should be able to include the file.

Although the robots.txt standard is an imperfect solution to web crawling, if a website owner concludes that it should be considered, there are several resources to help with the implementation process. The Web Robots Pages is an information resource that provides information about web robots and how to obtain and use a robot on a website.[158] Google also offers information about how to install a robots.txt file and provides a tool to test the robots.txt file to ensure the installed web robot is

[150] *Id.* at 632-33.

[151] *Id.* at 630.

[152] *Id.*

[153] *Id.* at 630-31.

[154] *Id.* at 632.

[155] *Id.* at 646.

[156] *A Standard for Robot Exclusion, supra* note 124.

[157] *Id.*

[158] *About Robotstxt.org, supra* note 120.

working properly.[159] Perhaps most importantly, a website owner should ensure that the file is properly installed to avoid the embarrassing fate of Chief Judge Kozinski.[160]

C. Recommendations

Understandably, this recitation of the admissibility of screen shots from the Wayback Machine as evidence and the difficulties of adequately preventing recordation of a website may seem somewhat discouraging. However, this innovative and developing service could also be advantageous not only to consumers but also to businesses looking to protect their intellectual property.

As described above, installing a robots.txt file is probably not a very effective solution to prevent the crawling of a website. Perhaps most notably, it appears at least from some cases that the information may be discoverable if the court orders the file to be removed.[161] However, in the unlikely event that there are no screen shots already available in the Wayback Machine's database, including if the website is brand new, a website owner could install the robots.txt file immediately to prevent any record. This way, even if ordered to disable the robots.txt file, the Wayback Machine will have no prior records stored and any web crawling that could be completed while the file is disabled will only be from present screen shots, which could be obtained without disabling the robots.txt file.

Therefore, before beginning a website, businesses and individuals should consider this option. If it seems undesirable for whatever reason to have a website documented throughout time, a website owner should consult online resources to familiarize himself or herself with the procedure.[162] The website owner should also contact a computer network professional who can be relied upon to install the file properly. However, if this seems like the appropriate option for a website owner or its business, it is important to keep in mind that the use of the robots.txt file is visible to anyone on the Internet. Therefore, consider not being too selective about which web pages are excluded and then be consistent for subsequently developed web pages.

However, if an already developed website exists and screen shots

[159] *Block or Remove Pages Using a Robots.txt File*, *supra* note 122.

[160] *See In re* Complaint of Judicial Misconduct, 575 F.3d 279, 289 (3d Cir. 2009).

[161] *See* Netbula, LLC v. Chordiant Software, Inc., No. Co8-00019 JW (HRL), 2009 WL 3352588, at *2 (N.D. Cal. 2009).

[162] *See, e.g.*, *The Web Robots Pages*, *supra* note 119; *Block or Remove Pages Using a Robots.txt File*, *supra* note 122; *and* Christopher Heng, *How to Set Up a robots.txt to Control Search Engine Spiders*, THE SITE WIZARD (Oct. 27, 2010), http://www.thesitewizard.com/archive/robotstxt.shtml.

are already available on the Wayback Machine's database, the options are somewhat limited. Obviously, no responsible individual or business would intentionally put content on its website that might generate liability; however, misunderstanding the law or innocent mistakes make liability a distinct possibility. The Wayback Machine also teaches the difficult lesson that once something is on the Internet, it may be there forever.[163] While the Wayback Machine does not record every change to a website, consistent use of certain content, such as descriptions or trademarks, are likely to appear if the site is crawled regularly. Therefore, it is essential that website owners have policies in place for review of all materials posted on its web pages. Ideally, legal counsel with special attention to product liability and intellectual property considerations should review the content to minimize potential risk.

While it may seem disadvantageous to have portions of a website's or business's history available on the Internet, there are several positive aspects to be considered as well. In addition to providing a way to prove first use of a trademark or prior use (before a patent) or independent creation (for copyright), the Wayback Machine also provides an opportunity to protect a website owner's intellectual property. For example, just as another company could use the Wayback Machine to prove unlicensed use of their trademark, any website owner could do the same. The Wayback Machine could become an excellent resource for monitoring current and potential infringers.

There are several options to help ensure that infringers' websites are documented in the Wayback Machine's database. First, the Internet Archive describes how a website owner can get its site listed in major Internet directories on the Wayback Machine Beta Frequently Asked Questions page.[164] By including a website on major Internet directories, such as dmoz.org, it ensures that web crawlers are able to find the website.[165] Instructions to add a website site to dmoz.org's Open Directory are available on its website.[166]

However, there are several important limitations on submission to the Open Directory Project. The Open Directory Project, or dmoz.org, does not accept all submitted websites because its goal is to make the

[163] *See* Bob Spankle, *The Internet is Forever . . . Sort of*, BIT BY BIT BLOG (Nov. 12, 2009), http://bobsprankle.com/bitbybit_wordpress/?p=1714; Patrik Fredriksson, *What happens online stays online. Forever*, TOOLBOX.COM (Dec. 11, 2007), http://it.toolbox.com/blogs/the-security-outlet/what-happens-online-stays-online-forever-21145.

[164] *My site's not archived! How can I add it?*, WAYBACK MACHINE BETA: FREQUENTLY ASKED QUESTIONS (Dec. 21, 2010), http://faq.web.archive.org/my-sites-not-archived-how-can-i-add-it/.

[165] *Id.*

[166] *Submitting Your Site*, DMOZ (Aug. 25, 2012 08:38), http://www.dmoz.org/docs/en/help/submit html.

directory as useful as possible for its users.[167] The Open Directory Project also has a policy against the inclusion of sites with illegal content, including material that infringes any intellectual property right."[168] It may take several weeks for any submission to be reviewed.[169] Additionally, the Open Directory Project provides that a site's placement in the directory is subject to the editor's sole discretion.[170] Finally and most importantly, the Terms of Use implicitly require that the individual submitting the request have ownership of the submitted website, e.g., the submitter assigns copyright in the material submitted for inclusion.[171] Therefore, a website owner could not submit a current or potential infringer's website to the Open Directory Project to ensure that it would be crawled. However, this option is still available if a website owner would like to potentially use the Wayback Machine's database to prove the past content of its own websites.

The second option described by the Internet Archive is currently not available. An update dated February 7, 2011 provides:

> At some point in the past couple of months, Alexa removed the ability to submit your site for crawling (option 2 above). At the moment, we do not have an alternative method of submission. As always, the best way to make sure we find your site is to make sure lots of people link to it. We'll notify users on this page if another solution becomes available. Thanks![172]

Fortunately, there appears to be a third option to ensure that Wayback Machine is aware of websites to be included in the archive. When searching for a website that is not yet archived by the Wayback Machine, a message will appear that says, "Hrm. Wayback Machine doesn't have that page archived."[173] However, when one clicks on the "Latest" button, which is available underneath the search box, the website is redirected to the website originally searched with a Wayback Machine message at the top that reads: "The Wayback Machine does not

[167] *Id.* (scroll to "How long does it take for my site to be listed in the ODP?").

[168] *See e.g., Open Directory Project Terms of Use*, DMOZ (Oct. 9, 2011 23:06:19), http://www.dmoz.org/docs/en/termsofuse html (scroll to "Claims of Copyright Infringement").

[169] *Submitting Your Site*, *supra* note 166 (scroll to "How long does it take for my site to be listed in the ODP?").

[170] *Open Directory Project Terms of Use*, *supra* note 168 (scroll to "Netscape Discretion Over Content, Use, And Operation of the ODP").

[171] Id. (scroll to "Copyright Assignment").

[172] *My site's not archived! How can I add it?*, *supra* note 164.

[173] *See, e.g.,* Ferreira Research Group, COLORADO STATE UNIVERSITY, http://web.archive.org/web/*/http://franklin.chem.colostate.edu/emferr/Ferreira_Research_Group/Home html (accessed through the Internet's Archive's Wayback Machine index) (last visited Feb. 10, 2013).

have this URL. Here is the page from the Live Web."[174] Months after searching for another not-yet-archived website on the Wayback Machine, archived images from at least eight captures of the website have been posted.[175] Therefore, months after searching for a little known website, an archival record could be available in the Wayback Machine's index.

Upon further review, the Internet Archive provides that "there is a 6-14 month lag time between the date a site is crawled and the date it appears in the Wayback Machine."[176] It further explains that "[it] generally takes six months or more for pages to appear in the Wayback Machine after they are collected, because of delays in transferring material to long-term storage and indexing . . . ."[177] However, in some cases, content "can appear in a much shorter timeframe – as little as a few weeks from when it was crawled."[178] Therefore, although this process may not appear to work after a few months, the archived image just may be delayed due to processing and indexing.

Another option is to install the Alexa Internet toolbar.[179] Those that choose to download the Alexa toolbar are considered members of the Alexa Toolbar community.[180] These members contribute to the information that Alexa Internet stores "about the web, how it's used, what's important and what isn't."[181] While Alexa Internet highlights the benefits to users of its toolbar, including providing key statistics about each site and related links that may be of interest to the user, it also "donates a copy of each crawl of the web to the Internet Archive."[182] Therefore, by using the toolbar, individuals are encouraging Alexa Internet to crawl the websites that they visit. While the possible implications of this service are beyond the scope of this note, the Alexa Internet Toolbar is an option that website owners could explore in order to ensure particular websites are crawled.

Therefore, when a website owner discovers an infringer or a possible infringer, the owner can take steps to ensure that the website is crawled by Alexa Internet to serve as a monitoring device. However, even though Alexa Internet may be aware of the website, it does not

[174] *See e.g., id.* (click on "Latest").

[175] *See e.g.,* MY CONSUMER TIPS, http://liveweb.archive.org/http://myconsumertips.info (accessed through the Internet's Archive's Wayback Machine index) (last visited Feb. 10, 2013).

[176] *The Wayback Machine: FAQ, supra* note 2 (scroll to "Where is the rest of the archived site? Why am I getting broken or gray images on a site?").

[177] *Id.* (scroll to "Why are there no recent archives in the Wayback Machine?").

[178] *Id.*

[179] *See The Alexa Toolbar for Firefox*, ALEXA INTERNET, INC., http://www.alexa.com/toolbar (last visited Feb, 10, 2013).

[180] *Alexa Technology, supra* note 14.

[181] *The Alexa Toolbar for Firefox, supra* note 179.

[182] *Alexa Technology, supra* note 14.

ensure that it will be crawled frequently enough to prove useful. Also, website owners must keep in mind that these third party website owners may have also educated themselves on the robots.txt file and ways to exclude their website content on the Wayback Machine. Therefore, these efforts to have their website included may be in vain. Therefore, some additional personal monitoring may be necessary.

Another option that is unrelated to the Wayback Machine is to take an independent screen shot or screen capture of the web page and make sure that a time and date stamp is included. On a Macintosh computer, this is easily done by opening the "Grab" tool under "Utilities" or use keyboard shortcuts to obtain the same results.[183] On Windows XP or Windows Vista, the Print Screen Key usually labeled "PrtSc" or "Print Screen" will create a screen capture.[184] Windows Vista also has a "Snipping Tool."[185] However, with this archiving option, it is important to keep in mind the federal rules of evidence, including the personal knowledge requirement of FED. R. EVID. 602.[186]

CONCLUSION

In conclusion, the most important lesson the Wayback Machine provides is that Internet content is easily copied and stored for future use. Such use may be legally adverse to a website owner but, by requiring legal review of website content, this risk may be mitigated. However, the use of such archived images may also be used positively to the benefit of the website owner by using the Internet Archive's archive service to monitor its own intellectual property. As the majority of federal courts admit such evidence with reasonable authentication requirements, a website owner can never be too aware of its options.

[183] *Find out how: Capture Screen Shots*, APPLE, http://www.apple.com/findouthow/mac/#capturescreen (last visited Feb. 10, 2013).

[184] *Take a screen shot*, WINDOWS, http://windows microsoft.com/en-US/windows-xp/help/setup/take-a-screen-shot (last visited Feb. 10, 2013).

[185] *Use Snipping Tool to capture screen shots*, WINDOWS, http://windows microsoft.com/en-US/windows-vista/Use-Snipping-Tool-to-capture-screen-shots (last visited Feb. 10, 2013).

[186] *See* FED. R. EVID. 602.