November 2, 2009

WAYBACK MACHINE MEMO

Report of the

Discovery Practices and Procedures Subcommittee

of the

Enforcement Committee:

Brian O'Bleness
Dana C. Jewell
J.C. Sebastian Pinckaers
James Weinberger
Jami A. Gekas
Janusz Marceli Fiolka
Jorge Molet
Laura E. Goldbard
Megan Dredla
Michael Cryan
Peter Eduardo Siemsen
Philip J. Kerr
R. Charles Henn
Rick McMurtry
Scott R. Miller
Shane David Hardy
Stephanie C. Alvarez
Lynda E. Roesch

With special thanks to John Plumpe of Charles River Associates and Charles Henn for research

An Overview of The Wayback Machine

MEMO – November 2, 2009

While conducting research for its draft resolution regarding the admissibility of internet and electronic evidence, the Discovery Practices & Procedures Subcommittee of the Enforcement Committee recognized the potential importance that the Internet Archive's Wayback Machine may play in discovery during global trademark and related litigation matters. The following is a summary of some of the information that the Subcommittee gathered while researching the Wayback Machine.

1. Background of Internet Archive / The Wayback Machine

The Wayback machine is an internet-based service provided by Internet Archive, a 501(c)(3) non-profit organization.[1] Internet Archive was founded in 1996 with the purpose of building an Internet library that offered permanent access for researchers, historians, and scholars to historical collections that may exist only in digital formats.[2] Originally limited to only archiving web pages, in 1999 Internet Archive began incorporating texts, audio, moving images, and software in its collection.[3]

Internet Archive states that it works with organizations such as the Library of Congress and the Smithsonian to prevent the Internet and other digital media from disappearing without any record.[4] With a typical lifespan of 44-75 days,[5] web pages are not as permanent as printed media. The Internet Archive's activities capture and store web pages and other digital media for future

[1] http://www.archive.org/about/about.php#storage.
[2] *Ibid.*
[3] *Ibid.*
[4] *Ibid.*
[5] http://www.archive.org/web/web.php.

reference. Accessing the actual archive directly requires Unix programming ability and a user account.[6] However, the Wayback Machine functions as an interface to the Internet Archive for those without such skills and the general public. It is Internet Archive's service that allows public access to its archived digital media.[7] Currently, the Wayback Machine offers access to over 85 billion web pages, dating back to 1996 in some instances.[8]

2. Functionality of the Wayback Machine

a. Crawling the Internet

The Wayback Machine accesses archived web sites that are provided by a "web crawler" operated by Alexa Internet and donated to Internet Archive.[9] Web crawlers are programs or automated scripts which browse the World Wide Web in a methodical, automated manner.[10] Some uses of web crawlers include checking links on web pages, collecting email addresses, and downloading web page source codes for archiving.

A web crawler typically starts with a list of web sites and then identifies all the hyperlinks in those pages. It then adds those hyperlinks to the original list of web sites to visit.[11,12] The process is then repeated for the duration of the web crawl. Following this iterative process, it is possible for a web site to be archived only once during a web crawl or multiple times a day. If no links to a web site appear on any other web site, it will not be included in the web

---

6 http://www.archive.org/web/researcher/intended_users.php.
7 http://www.archive.org/about/faqs.php.
8 http://www.archive.org/about/faqs.php.
9 http://www.archive.org/about/faqs.php. Alexa Internet is a web information company that conducts web crawls, creates online directories of web sites, and monitors web traffic (http://www.alexa.com/site/company). Each web crawl takes approximately 2 months to complete and collects about 4.5 billion web pages from 16 million web sites. (http://www.alexa.com/site/company/technology).
10 http://en.wikipedia.org/wiki/Web crawler.
11 *Ibid.*
12 http://www.alexa.com/site/help/webmasters.

crawl. Many web sites, including search engines like Google, use some type of web crawler to index web sites.[13]

There are several ways to be included in Alexa's web crawl.[14] The first method is to access Alexa's "Webmasters" page and manually add a web site to be included in Alexa's next crawl of the web.[15] The second method is to visit a web site with the Alexa Toolbar installed in a web browser. This will automatically include the web site in Alexa's next web crawl. It is reported that web sites that are added to Alexa's web crawl are usually crawled within eight weeks of submission. In addition to the eight week delay between submission and crawling, there is typically a six month lag between when a site is crawled and when it is available through the Wayback Machine.[16]

b. Web Sites Collected

The web crawls donated to Internet Archive do not always capture entire web sites for every available date. Instead, when browsing an incompletely archived site, the Wayback Machine will grab a linked page with the closest available date to the page the user is currently viewing. This makes it possible for the user to view older or newer pages when surfing a web site archived on a specific date.[17] In addition, if the Wayback Machine does not have the requested link archived, it will attempt to find the link on the current web page and redirect the user there.[18] The URLs created by the Wayback Machine include a code for the date that each particular web page was visited by the web crawler. The code uses the format "yyyymmddhhmmss."[19] For example, October 1, 2009 12:00:00 PM would appear as 20091001120000.

[13] http://en.wikipedia.org/wiki/Web_crawler.
[14] http://www.archive.org/about/faqs.php.
[15] http://www.alexa.com/site/help/webmasters.
[16] http://www.archive.org/about/faqs.php.
[17] *Ibid.*
[18] *Ibid.*
[19] *Ibid.*

While the web crawls donated to Internet Archive are fairly exhaustive, the Wayback Machine does not have a complete collection of all web sites that existed since 1996. The web crawls only collect publicly available web sites. Pages that either require a password to access, are tagged for "robot exclusion" by their owners, are accessible only when a user types into and sends a form, or exist on a secure server are not archived.[20] According to Internet Archive, other difficulties in archiving web sites include the use of Javascript, server side image maps, orphan pages (web sites that are not linked to by any other web pages), and unknown sites.[21] Simple HTML web sites are the easiest to archive. Crawled web sites are stored in 100 megabyte .ARC files, which are composed of many individual files.[22] Archived web sites in the Wayback Machine do not always appear as they did on the live web for reasons such as the previously mentioned difficulties in archiving web sites.[23]

There are a number of ways for web sites to be manually excluded from the Wayback Machine. Web site owners can submit an online request for their sites to be excluded from the Wayback Machine; a "Blocked Site Error" message designates these sites in the Wayback Machine.[24] Alternatively, web site owners can include a "robots.txt" file in their web site header. This file adheres to the Standard for Robot Exclusion, a voluntary convention to prevent cooperating web crawlers from accessing all or part of a web site that is otherwise viewable by the public.[25] Once a web site includes a robots.txt file, the Alexa crawler will stop visiting the specified locations and will

[20] *Ibid.*
[21] *Ibid.*
[22] http://www.archive.org/about/about.php.
[23] http://www.archive.org/about/faqs.php.
[24] *Ibid.*
[25] The robots.txt file in a web site header acts as a request to all or specific web crawlers to ignore specified files or directories during the crawl. http://www.archive.org/about/faqs.php, http://en.wikipedia.org/wiki/Robots exclusion standard.

retroactively make unavailable all files previously gathered from the site. This creates a "Robots.txt Query Exclusion" message in the Wayback Machine.[26]

3. Legal Use of the Wayback Machine

Use of documents and information obtained from the Internet Archive Wayback Machine as evidence in legal proceedings implicates several different evidentiary issues, including authentication, the hearsay rule, and inherent untrustworthiness concerns. The application of the hearsay rule to Wayback Machine evidence generally depends on the manner in which the evidence is sought to be used. For example, proponents of Wayback Machine evidence have avoided problems under the hearsay rule by offering the evidence for a non-hearsay purpose[27]—such as to show declarant's state of mind—or by coming within a hearsay exception—such as where the declarant is a party opponent.[28] For authentication, however, at two different positions have emerged among U.S. federal courts regarding the requirements for satisfying Federal Rule of Evidence 901 for Wayback Machine Evidence.

[26] http://www.archive.org/about/faqs.php; http://www.archive.org/about/exclude.php. In at least once case, however, the failure of robots.txt files properly to exclude access through the Internet Archive, has led to litigation. *See Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey*, 497 F. Supp. 2d 627 (E.D. Pa. 2007). Due to a malfunction of Internet Archive's servers, when the Harding firm used the Wayback Machine to access archived versions of Healthcare Advocates's website, the firm was able to view pages to which its access that should have been restricted by operation of a robots.txt. file in the website's code. *Id*. at 630, 632. The Healthcare Advocates asserted that the Harding firm's actions constituted hacking and brought an action against the firm alleging copyright infringement, violation of the Digital Millennium Copyright Act (DMCA), violation of the Computer Fraud and Abuse Act (CFAA), and conversion and trespass to chattels under Pennsylvania law. *Id*. at 630, 633. The court found for defendant on summary judgment on all counts, holding that the state law claims were preempted by the federal Copyright Act, that Harding's use of the copyrighted material constituted a fair use, and that for the DMCA and CFAA claims, the Harding firm did not circumvent the robots.txt file and plaintiff had provided no evidence that the firm intentionally exceeded its authorized access. *Id*. at 634-50.

[27] *See, e.g.*, Telwizja Polska USA, Inc. v. Echostar Satellite Corp., No. 02-3293, 2004 WL 2367740, at *5 (N.D. Ill. Oct. 15, 2004) ("To the extent these images and text are 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.") (citation omitted).

[28] *See id*., at *5 (finding that even if the Wayback Machine evidence were offered for the truth of the matter asserted, "the contents of [plaintiff]'s website may be considered an admission of a party-opponent, and are not barred by the hearsay rule").

a. Authentication of Wayback Machine Evidence In U.S. Federal Courts

It is the stated policy of Internet Archive that the Wayback Machine was not created for legal use and that Internet Archive strives to be a disinterested third party in all disputes involving its archived material.[29] Due to limited resources, Internet Archive recommends seeking judicial notice or stipulation from the opposing party regarding the authenticity of archived web pages before enlisting Internet Archive for assistance in authenticating web pages.[30]

Internet Archive does, however, offer authentication services for web pages from the Wayback Machine along with a standard affidavit affirming the same for a fee.[31] The affidavit and authenticity services affirm that printed web pages from the Wayback Machine are true and accurate copies of its records. Internet Archive states that it remains the requesting party's burden to prove to the finder of fact that the archived web pages were available at the date and time shown in the URL.[32]

Several U.S. federal courts have found Rule 901 satisfied for the admission of Wayback Machine where the proponent offers such an affidavit of an Internet Archive employee to authenticate the evidence,[33] particularly where the objecting party has not actually denied the evidence's authenticity or presented any evidence that it is not genuine.[34] Authenticating affidavits by Internet

---

[29] http://www.archive.org/legal/faq.php.
[30] *Ibid.*
[31] http://www.archive.org/legal/, http://www.archive.org/legal/affidavit.php.
[32] http://www.archive.org/legal/faq.php.
[33] *See, e.g.*, , *e.g.*, SP Technologies, LLC v. Garmin International, LLC, No. 08 C 3248, 2009 WL 3188066, at *3 (N.D. Ill. Sep. 30, 2009) (accepting Internet Archive evidence authenticated by affidavit of manager at Internet Archive); *Telwizja*, No. 02-3293, 2004 WL 2367740, at *6 (denying motion in limine to bar Wayback Machine evidence on hearsay and authentication grounds where proponent attached "an affidavit of Ms. Molly Davis, verifying that the Internet Archive Company retrieved copies of the website as it appeared on the dates in question from its electronic archives"); *see also* Mortgage Market Guide, LLC v. Freedman Report, LLC, No. 06-CV-140, 2008 WL 2991570 (D.N.J. July 28, 2008) (although finding the issue not controlling, the court recognized that "other federal courts typically reject documents obtained from web archive services, unless they are accompanied by a 'statement or affidavit from [a representative with personal knowledge of the contents of the [archive] website.' ").
[34] *See* Masters v. UHS of Delaware, Inc., No. 4:06-CV-1850, 2008 WL 5600714, at *2 (E.D. Mo. Oct. 21, 2008) (admitting plaintiff's Wayback Machine evidence where plaintiff submitted an affidavit of a person

Archive employees that have been accepted by courts have included attestation to the affiant's personal knowledge of the Internet Archive, an explanation of how the Internet Archive works, and attestation to the authentication of the exhibit in dispute.[35] Conversely, several courts have denied admission of Wayback Machine evidence for lack of authentication where the proponent failed to submit an affidavit from an employee of the Internet Archive.[36]

Courts in the Second Circuit have adopted a different view of the requirements for authenticating Wayback Machine evidence under Federal Rule 901.[37] Second Circuit courts have required the testimony or affidavit of an employee of the company hosting the original website that the Wayback Machine's archived web pages purport to represent. For example, in *Novak v. Tucows, Inc.*, the Eastern District of New York held that the plaintiff's Wayback Machine evidence could not be authenticated as required under the Rules of evidence because the plaintiff offered "neither testimony nor sworn statements

---

who attested to having personal knowledge of how information is collected and stored at the Internet Archive and where defendant did not contend that the exhibit contained inaccurate or false representations of defendant's website); *Telwizja*, No. 02-3293, 2004 WL 2367740, at *6 ("Plaintiff has neither denied that the exhibit represents the contents of its website on the dates in question, nor come forward with its own evidence challenging the veracity of the exhibit. Under these circumstances, the Court is of the opinion that Ms. Davis' affidavit is sufficient to satisfy Rule 901's threshold requirement for admissibility.")

[35] *See SP Technologies*, No. 08 C 3248, 2009 WL 3188066, at *3 (denying motion to strike Wayback Machine evidence where proponent attached an affidavit from a manager at the Internet Archive, explaining how the website saves old web pages and that proponent's exhibit was created in 1999); *see also* St. Luke's Cataract & Laser Institute, P.A. v. Sanderson, No. 8:06-CV-22, 2006 WL 1320242, at *2 (M.D. Fla. May 12, 2006) (describing the contents of the authenticating affidavit accepted in the *Telwizja* case: "Ms. Davis' affidavit was submitted to verify that the copies of the web pages retrieved from Internet Archive were accurate representations of the web pages as they appeared in Internet Archive's records. Her affidavit also described in detail the process Internet Archive uses to allow visitors to search archived web pages through its 'Wayback Machine.' " Most importantly, the affidavit contained specific attestations of authentication as to the web page in dispute.") (internal citations omitted).

[36] Zinn v. Seruga, No. 05-3572, 2009 WL 3128353, at *27 n.8 (D. N.J. Sep. 28, 2009) (excluding proponent's Wayback Machine evidence, noting that the proponent had "not called a witness from that organization to authenticate its compilation and storage of such information, or provided any other valid means to authenticate" the evidence); Audi AG v. Shokan Coachworks, Inc., 592 F. Supp. 2d 246, 278 (N.D.N.Y. 2008) (refusing to consider Wayback Machine evidence in support of plaintiffs' Motion for Summary Judgment where plaintiffs submitted the print-outs without authentication from a representative from the Internet Archive); *see also St. Luke's Cataract & Laser Institute*, No. 8:06-CV-22, 2006 WL 1320242, at *2 (denying admission of Wayback Machine evidence, finding insufficient for authentication purposes the declarations of the individuals who conducted the Internet Archive search and a certified copy the affidavit of the Internet Archive representative that was submitted in the *Telwizja* case).

[37] *See* Novak v. Tucows, Inc., No. 06-CV-1909, 2007 WL 922306, at *5 (E.D.N.Y. Mar. 26, 2007), *aff'd* No. 07-2211-CV, 2009 U.S. App. LEXIS 9786, at *6 (2d Cir. May 6, 2009).

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."[38] The court focused on the fact that "the web pages archived within the Wayback Machine are based upon 'data from third parties who compile the data by using software programs known as crawlers,' who then 'donate' such data to the Internet Archive.' "[39] The court emphasized 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."[40] This ruling by the Eastern District of New York was recently affirmed by the Second Circuit Court of Appeals[41] and also has been cited with approval (albeit in dicta) by the Southern District of New York.[42]

b. Authentication of Wayback Machine Evidence Before the U.S. Trademark Trial and Appeal Board

The United States Trademark Trial and Appeal Board has similarly recognized the authentication issues involved in introducing Wayback Machine evidence in a legal proceeding.[43] In *Paris Glove of Canada, Ltd. v. SBC/Sporto Corp.*, for example, the Board, citing *St. Luke's Cataract and Laser Institute,* and *Novak v. Tucows, Inc.*, noted that "in recent cases that have discussed or dealt

[38] *Id*. *Contra SP Technologies*, No. 08 C 3248, 2009 WL 3188066, at *3 (rejecting opponent's suggestion that the Wayback Machine print-out would be admissible only if a person with direct knowledge of the website's existence at the time the site was archived testified that the print-out was a true and accurate copy of the contents of the website on that date; the court held that "[s]uch a high standard is not required for other types of evidence, and is beyond what Rule 901 requires").
[39] *Novak*, No. 06-CV-1909, 2007 WL 922306, at *5.
[40] *Id*.
[41] Novak v. Tucows, Inc., No. 07-2211-CV, 2009 U.S. App. LEXIS 9786, at *6 (2d Cir. May 6, 2009) ("[T]the District Court did not err, much less abuse its discretion, in admitting into evidence certain of defendants' affidavits, and in denying the admission of certain of Novak's exhibits.").
[42] Chamilia, LLC v. Pandora Jewelry, LLC, No. 04-CV-6017, 2007 WL 2781246, at *6 n.4 (S.D.N.Y. Sep. 24, 2007) (denying defendant's motion to strike as moot, but stating in dicta that plaintiff's Wayback Machine evidence "suffers from fatal problems of authentication under Fed.R.Evid. 901.") (citing *Novak*, No. 06-CV-1909, 2007 WL 922306, at *5).
[43] *See* Paris Glove of Canada, Ltd. v. SBC/Sporto Corp., 84 U.S.P.Q.2d 1856, 2007 WL 2422997, at *3 (T.T.A.B. Aug. 22, 2007) ("As to applicant's argument that the Internet Archive makes the holding in *Raccioppi* obsolete, the database itself is not self-authenticating, and there is no reason to treat its existence as authenticating the pages in its historical record.").

with evidence from the Internet Archive, supporting declarations accompanied the evidence."[44] And, although the Board has acknowledged "the general unacceptability" of evidence obtained from the Wayback Machine, the Board has accepted and considered such evidence.[45]

c. Authentication of Wayback Machine Evidence in Australia

An Australian court that recently considered the admissibility of Wayback Machine evidence identified the same evidentiary issues that have raised concerns for tribunals in the U.S. In *E & J Gallo Winery v. Lion Nathan Australia Pty Ltd*, the trial judge rejected as purported evidence of past trademark use two brochures obtained from an archived web page through the Internet Archive Wayback Machine. The judge found the evidence inadmissible both on hearsay grounds and because the it did not satisfy the requirements of the Evidence Act. Specifically, the court found that the evidence did not fall within the business records exception to the hearsay rule and did not meet the Evidence Act's requirements for admissibility of evidence produced by computers or machines. Similar to some American courts, the Australian judge also criticized the Wayback Machine evidence as inherently unreliable.[46]

1702889v2

[44] *Id*.
[45] *See* Hiraga v. Arena, 90 U.S.P.Q.2d 1102, 2009 WL 723334, at *4 (T.T.A.B. Mar. 18, 2009) (allowing introduction of Wayback Machine website print-outs over opposing party's objection where the opposing party had also relied on Wayback Machine print-outs).
[46] *E & J Gallo Winery v. Lion Nathan Australia Pty Ltd* (2008) 77 IPR 69 at [124]-[129].