NOTE

"CAN YOU DO A WAYBACK ON THAT?" THE LEGAL COMMUNITY'S USE OF CACHED WEB PAGES IN AND OUT OF TRIAL

Matthew Fagan[*]

TABLE OF CONTENTS

[*] J.D. Candidate, Boston University School of Law, 2007; B.S. Computer Science, Carnegie Mellon University, 2004; B.A. Philosophy, Carnegie Mellon University, 2004.

I. INTRODUCTION

It is the year 2054, and there is no more murder.

Any time someone contemplates killing a fellow human being, the police know immediately. They see the homicide as it will happen on their computer screens, and they arrive at the scene of the contemplated crime long before it can occur. Upon the killer's arrival they arrest him, stopping the murder before it happens.

Welcome to the world of Twentieth Century Fox's *Minority Report.*[1] While precognitive crime fighters may not be in our immediate future, what if the legal community could do something similar in reverse? Instead of seeing a crime before it happens, we wait for the crime to happen, then go to our computers. A few keystrokes later, we see exactly what occurred: perfect recall. Alternatively, suspects could go into the system and prove their innocence by demonstrating the truth of their alibis.

In the Internet-crime context, we *can* do something like that now. The Internet has a history, and for the most part, it is open for anyone to view. This history can be an incredibly useful tool for attorneys, but they should heed the warning of the character Gideon from *Minority Report*:

"Careful, Chief . . . . You dig up the past, all you get is dirty."[2]

Cached web pages allow users to look into a web site's past. Instead of viewing the page as it exists now, cached copies of the page might reflect how it looked a week, a month, or even ten years ago. These pages are available to the public through a variety of sources called "webcaching services."

[1] MINORITY REPORT (Twentieth Century Fox 2002).

[2] *Id.*

Many attorneys already use cached pages in their daily practice.[3] This use is widespread[4] and, for the most part, unquestioned. This is primarily because the attorneys are utilizing the pages to, for instance, direct investigations or learn what information might be available for discovery.[5] These uses can lead to undeniably useful and credible evidence. Some attorneys have begun to take the next step by seeking to admit the cached web pages themselves as evidence in trials.[6] Because of this, we are beginning to see the first legal challenges to cached pages.[7] As we are in the early stages of these challenges, it is important for attorneys to know the strengths and weaknesses of caching so that they can limit themselves to certain safe (or saf*er*) uses of cached material.

Can cached web sites be introduced as evidence at trial in civil and criminal cases, and, more pointedly, should they be? Ken Strutin, director of legal information for the New York State Defenders Association, recently addressed this question:

> the issue of whether archived Web sites are admissible as evidence hasn't been widely tested in criminal cases, where the burden of proof is higher than in civil court . . . . Lawyers seeking to submit evidence from a cache system or the Wayback Machine would face a much more rigorous standard.[8]

The Federal Rules of Evidence allow a wide breadth of evidence and then take a "let the jury decide what information is credible" approach.[9] Does that give an unfair advantage to attorneys seeking to admit cached web pages? How valuable are these pages to trial lawyers, or lawyers in general? How reliable? What obstacles exist to using them, and can attorneys overcome these obstacles?

[3] David Kesmodel, *Lawyers' Delight: Old Web Material Doesn't Disappear*, WALL ST. J., July 27, 2005, at A1.

[4] For example, over 100 journals use a resource called "WebCite" (http://www.webcitation.org), which allows them to substitute cached links for regular links in their articles, ensuring that the links that users follow will accurately reflect the web site as it was when the author wrote their article. Webcitation.org, WebCite Members List, http://www.webcitation.org/members (last visited January 30, 2006).

[5] *See* Kesmodel, *supra* note 3 .

[6] *Id.*

[7] *See*, e.g., Telewizja Polska USA, Inc. v. Echostar Satellite Corp., No. 02 C 3293, slip op. (N.E Dist. Ill. Oct. 14, 2005), *available at* http://cyberlaw.stanford.edu/packets/echostar.pdf.

[8] Kesmodel, *supra* note 3.

[9] *See* Telewizja Polska USA, Inc., slip op. at 14 (quoting United States v. Harvey, 117 F.3d 1044, 1049 (7th Cir. 1997)).

In Part II, I describe webcaching technology and the different kinds of webcaching services to provide a better understanding of the strengths and weaknesses of this technology. Following this is an overview of how caching is currently used in the legal community and a discussion of representative cases where the issues prevalent in caching arise. Parts III and IV discuss the strengths and weaknesses, respectively, of caching services. Next, in Part V, I review the admissibility of cached web pages under the Federal Rules of Evidence. Finally, Part VI presents recommendations on how the legal community can effectively use cached web pages while avoiding some of the major pitfalls of archived web sites.

II. BACKGROUND

A. *What is Caching?*

1. Caching in general

In computer science, "caching" refers to the temporary storage of information where it can be easily accessed for future use. For example, imagine that all the important documents your company has are stored in a central filing cabinet. When you need a document, you typically go to the filing cabinet, wait in line to retrieve the document, use the document, and then return it to the cabinet. However, if you plan to use the document more than once, it would be more efficient to make a copy to keep at your desk. That way, you can get the file once and view your own personal copy thereafter. This also saves you from having to wait in line for the document. The more in demand the document, the more time you save by avoiding the line.

Information is cached on computers in a similar way. Computers use caching in many different circumstances. For instance, data present on a hard drive can be copied into the computer's RAM, which has a much faster access speed, but has far less capacity for storage than the hard drive. The next time that the computer needs to access this data, it can find it more quickly in the RAM than it could on the hard drive.[10]

2. Caching on the Internet

To illustrate how caching occurs in the context of the Internet, consider the following analogy. Imagine that your company is divided into departments. If some documents are particularly popular, especially among certain

[10] *See* I. Trotter Hardy, *Symposium: Copyright Owners' Rights and Users' Privileges on the Internet: Computer RAM "Copies": a Hit or a Myth? Historical Perspectives on Caching as a Microcosm of Current Copyright Concerns*, 22 DAYTON L. REV. 423, 463 n.5 (1997).

departments, it might be advantageous for the head of each department to keep a department-wide copy. Then, when a member of the department needed a certain document, that person could first check to see if the department has a copy, and if none is available, ask the department to get the document from the central file and maintain a copy for future reference by anyone in the department. This allows for greater efficiency, and can be combined with individual copying to further reduce wait times.

In the Internet context, caching similarly means "the storing of copies of content [that subscribers wish to see most often] at locations in the network closer to subscribers than their original sources . . . in order to provide more rapid retrieval of information."[11] Web browsers like Firefox and Internet Explorer store cached web pages at the location closest to the user (the local computer itself),[12] while Internet service providers ("ISPs") cache web pages on proxy servers in order to provide streamlined access to the most popular pages among a large group of users.[13] In this way, users get their copies from the proxy server instead of the site owner's server, which is likely to be slower and more congested than the proxy server. Browser caching is akin to making a personal copy of an important document, as in the first example. while ISP caching is like department-wide copying. This ability to cache web pages allows the Internet to run quicker and more smoothly by reducing bandwidth constraints.[14] The main difference between documents and web pages is that web pages can change frequently. While some may remain stagnant for months, others (like news sites) can change every few minutes. Therefore, the cached copies need to be updated regularly to keep up with the pace of change.

3. Search engine caching

Some search engines also cache web pages, but for different reasons. A

[11] *In re* Inquiry Concerning High-Speed Access to the Internet Over Cable & Other Facilities*,* 17 F.C.C.R. 4798, ¶ 17 n.76 (2002).

[12] *See* Microsoft, Internet Explorer: How and Why to Clear Your Cache, http://www.microsoft.com/windows/ie/using/howto/customizing/clearcache.mspx (last visited Jan. 30, 2006); Mozilla.org, Firefox Help: Firefox FAQ, http://www.mozilla.org/support/firefox/faq.html#profiles (last visited Jan. 30, 2006).

[13] *See* RON WHITE, HOW COMPTUERS WORK SEVENTH EDITION 339 (Que 2004).

[14] *See* Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967, 999(2005) (noting that "caching obviates the need for the end user to download anew information from third-party Web sites each time the consumer attempts to access them, thereby increasing the speed of information retrieval."); (citing *In re* Inquiry Concerning High-Speed Access to the Internet Over Cable & Other Facilities, 17 FCC Rcd. 4798, at ¶ 17 n.76 (2002)); *see also* Richard S. Vermut, *File Caching on the Internet: Technical Infringement or Safeguard for Efficient Network Operation?,* 4 J. INTELL. PROP. L. 273 (Spring 1997).

search engine works by employing a program called a "web crawler" or "web spider."[15] This program "crawls" the web, visiting every web page it can find, and stores information about those web pages in an index. Indexes might be very small or very large; for instance, a very simple index might hold a list of specific keywords present on each web page the web crawler found. When a user types a query into the search engine, the engine checks the query against the list of keywords in the index, and returns a list of web pages containing the query. The order in which the results are presented depends on an algorithm employed by the search engine; for example, a simple algorithm might place web pages where the query terms occur close together at the top of the list, while placing web pages where the query terms are further apart closer to the bottom. In this way, pages with more relevant information are more likely to appear higher on the list.

Some indexes, including Google's, store a complete copy of the web page instead of just a list of keywords.[16] Google's success is due, in part, to the algorithm it uses to determine which pages are most relevant to the searcher.[17] Storing a cached copy of the entire web page simplifies the calculations that Google makes to determine this "pagerank."[18]

Temporary storage of popular sites for ease of access and storage of sites by search engines in order to determine page relevance are the two primary ways caching is used on the World Wide Web. They can, however, serve other functions, as attorneys have discovered to their benefit.

B. Caching Services

Several companies that use cached web pages for various purposes make these pages available to end-users. ISPs necessarily make these pages available. Typically, when a user accesses a popular page, they are actually viewing their ISP's (or some other ISP's, or their own browser's) cached copy.[19] These cached copies do not last very long: web pages generally tell the

[15] eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1061 n.2 (N.D. Cal. 2000) ("Programs that recursively query other computers over the Internet in order to obtain a significant amount of information are referred to in the pleadings by various names, including software robots, robots, spiders and web crawlers.").

[16] GoogleGuide, How Google Works and Google's Indexer, http://www.googleguide.com/google_works.html (last visited Jan. 30, 2006).

[17] Bart Eisenberg, Sidebar: An Interview with Danny Sullivan, Creator of Search Engine Watch, http://www.gihyo.co.jp/magazine/SD/pacific/SD_0408.html (last visited April 3, 2006) (trans. by Hiroshi Iwatani for "Pacific Connection" Series in the Japanese Magazine *Software Design*, Aug. 2004).

[18] *See* Phil Craven, *Google's Pagerank Explained*, WEBWORKSHOP.NET, http://www.webworkshop.net/pagerank.html (last visited Jan. 30, 2006).

[19] *See* WHITE, *supra* note 13, at 339-341.

proxy server how long to keep a cached version before that version becomes "stale" and must be reloaded.[20] This allows for some individualization. For instance, a personal homepage that is updated once a month would need to be re-cached only about once a month, but a highly dynamic page like CNN's homepage might need to be updated every minute. This increases the chance that when a user requests a page, the cached page that they actually see is the correct "fresh" version of the page.[21]

On the other hand, search engines only update their cached copies when they crawl the web, which happens far less often than ISPs update their cached copy. Many search engines measure their updates in terms of weeks or months, while ISPs measure their updates in minutes or hours.[22] This means that a cached web page in a search engine looks further back into a web page's "past" than a cached version by an ISP. Users looking at a search engine's cached copies[23] will most likely see the web page as it existed some time in the past, rather than the way it looks in the present. This could be useful in case a web page is changed before someone gets a chance to come back to a site in which they were interested, or if a page is taken down but a user still needs to access it (a situation known as "linkrot"). Realizing the potential utility of these cached copies, some Internet search services have made them available to the searching public.

There are three primary sources for viewing cached web pages online: Google, Yahoo!, and the Internet Archive.

1. Google[24] Caching

In 1997, Google began to offer users access to its cached versions of web pages.[25] This was the first time that searchers could gain access to a page through a search engine even after it had been removed from the Internet.[26]

Google's webcrawler is called "Googlebot."[27] Googlebot crawls the web

[20] Caching Tutorial for Web Authors and Webmasters, http://www.mnot.net/cache_docs/ (last visited Jan. 30, 2006).

[21] *Id.*

[22] *See* GoogleGuide, *supra* note 16.

[23] By "search engine's cached copies" I mean the internal copies that a search engine uses to determine how relevant a page is to a users search, and not the page that appears when a user clicks on a link after performing a search. *See* Field v. Google, Inc., 412 F. Supp. 2d 1106, 1110-11 (D. Nev. 2006). The former copies are the primary topic of this Note, while the latter copies are the ISP-or-browser-cached sites referred to previously.

[24] Google Home Page, http://www.google.com (last visited Jan. 25 2007).

[25] Stephanie Olsen, *Google's Cache Causes Copyright Concerns*, CNET NEWS, July 10, 2003, http://news.zdnet.co.uk/internet/ecommerce/0,39020372,2137329,00.htm.

[26] *Id.*

[27] AMY N. LANGVILLE & CARL D. MEYER, GOOGLE'S PAGERANK AND BEYOND: THE

based on an algorithm that determines how often it visits each page;[28] thus, sites that Google deems more important will be visited more often than those Google deems less important.[29] A copy of each page that Googlebot crawls is saved in Google's massive server farm.[30] These copies, or "snapshots," usually contain all of the text on a site, but may lack other data, like pictures and sounds.[31]

Any time a user performs a Google search and receives the results of her query, a "cached" link is displayed below the page's description. Clicking on that link takes the user to the cached snapshot. Though this caching service is extremely popular, several other mainstream search engines, including Yahoo, provide similar services.

2. Yahoo! My Web (Beta)[32]

Yahoo!'s My Web service also allows users to view snapshots of web pages as they were when they were cached.[33] A "cached" link, similar to Google's, is placed below the site's description, and takes the user to the page saved by Yahoo!'s webcrawler the last time it visited the site.[34]

However, Yahoo! provides an additional feature to Yahoo! users. After performing a search, a user is presented with the appropriate results for the user's search criteria. Under the description and beside the "cached" link is a "save" link. Yahoo! members can click the save link to take their own snapshot of the page. When they view this cached snapshot, the page will

SCIENCE OF SEARCH ENGINE RANKINGS 15-16 (Princeton University Press 2006); *see also id. at* 27-28.

[28] *Id.*

[29] *Id.* For instance (as an informal test only), when the author viewed the cached version of CNN's homepage at 11:41 PM EST on November 12, 2005, Google reported that the page had last been crawled at 10:13 PM EST (3:13 GMT) on November 11, 2005, for a difference of about twenty-five and a half hours. Meanwhile, the author's personal homepage, when accessed at the same time, was last cached at 4:05 AM EST (9:05 GMT) on November 9, 2005, for a difference of about ninety-one and a half hours.

[30] *See* GoogleGuide, *supra* note 16.

[31] Internet Archive Frequently Asked Questions, http://www.archive.org/about/faqs.php (follow "Why am I getting broken or gray images on a site?" hyperlink) (last visited Jan. 30, 2006) [hereinafter Internet Archive FAQ #18].

[32] My Web Beta, http://myweb.search.yahoo.com/ (last visited Jan. 25, 2007) A second beta version of My Web is available at http://myweb2.search.yahoo.com/ (last visited Jan 25, 2007).

[33] My Web 2.0 Beta FAQ, http://myweb2.search.yahoo.com/myresults/faq#saveacopy, (follow "11. I read that . . ." hyperlink) (last visited Jan. 30, 2006).

[34] *Id.*

appear as it did when the user clicked the save link.[35] Google's service overwrites cached snapshots whenever Googlebot visits the page, so a page that is visited often will appear nearly current in a cached snapshot.[36] MyWeb users can save the page to prevent it from being lost during the next web crawl.[37]

However, users must manually save a page to prevent its loss.[38] There is no way for MyWeb users to view a page that they have not saved, if the cached snapshot has already been overwritten.[39]

3. The Internet Archive Wayback Machine[40]

The Internet Archive provides an interesting solution to the problem of finding cached web pages overwritten by other search engines. The Internet Archive's service, called the "Wayback Machine" (named for the time machine in *Rocky and Friends*) ,[41] stores every snapshot taken from web sites that the Alexa webcrawler has visited since 1996.[42] For instance, the Wayback Machine contains 2701 versions of "www.cnn.com" from 2001, when it took snapshots as often as several times a day.[43] Those snapshots are available for anyone to view.

However, as of January 30, 2006, the Wayback Machine had only eighteen versions of "www.cnn.com" from 2005.[44] No pages are archived in the most recent six months, and sometimes versions of a page are not added to Wayback's database for up to twelve months.[45]

[35] *Id.*

[36] Access Deleted Pages with the Google Cache and Internet Archive, The Information Bank, http://www.bankblog.info/2005-11/access_deleted_.html (last visited Jan. 30, 2006).

[37] MyWeb 2.0 Beta FAQ, *supra* note 33.

[38] *Id.*

[39] This is changing slightly with MyWeb 2.0 (beta). MyWeb 2.0 has introduced a "Community" feature which allows users to search the cached web pages of other users in their community. So, it might be possible to find a cached web page stored by another user. However, *some* user must manually save the web page for it to be viewed by anyone.

[40] Internet Archive, http://www.archive.org/ (last visited Jan 30, 2006).

[41] Internet Archive: Information from Answers.com, http://www.answers.com/topic/internet -archive (last updated Jan. 30, 2006) [hereinafter Answers.com].

[42] Internet Archive Frequently Asked Questions, http://www.archive.org/about/faqs.php (follow "How can I get my site included in the Wayback Machine?" hyperlink) (last visited Jan. 30, 2006) [hereinafter Internet Archive FAQ #1].

[43] Internet Archive Wayback Machine, Results of a Search for "http://www.cnn.com", http://web.archive.org/web/*/http://www.cnn.com (last visited Jan. 30, 2006).

[44] *Id.*

[45] Internet Archive FAQ #1, *supra* note 42.

The Wayback Machine does not save a copy of a page every time the page is updated. Instead, the Wayback Machine saves a copy every time it actually visits the page by crawling the web.[46] Users can also request to have their sites removed from the Wayback archive[47] although the Wayback may not honor these requests.[48] Even with these limitations, the Wayback Machine provides a wealth of cached snapshots and is the most comprehensive archive of web history available.[49] Over 40 billion snapshots are available from among more than a petabyte of data.[50]

4. Other Resources

In addition to the three primary caching services, a number of other sources offer cached content to the public. These alternative sources usually contain fewer pages than the primary three and may contain older information. As I discuss below, this older content may be a useful feature of these secondary sources.

Gigablast[51] - Gigablast is a search engine released in 2000.[52] It indexes and archives web pages in much the same way as Google does. Archived copies can be several months to several years old.

Fagan Finder[53] - Fagan Finder is not itself a caching service, but it provides links to a number of other services. Think of it as a search engine for webcachers.

Spurl[54] - Spurl, like MyWeb, allows users to cache web pages manually. Pages are cached on the Spurl server.

Furl[55] - Furl is very similar to Spurl.

Feedster[56] - Feedster primarily caches popular weblogs.

DayPop[57] - DayPop archives news sites, weblogs, and RSS feeds.[58]

[46] Internet Archive Frequently Asked Questions, http://www.archive.org/about/faqs.php (follow "What Does it Mean when a Site's Archive Data has been 'Updated'?" hyperlink) (last visited Jan. 30, 2006).

[47] The Internet Archive's Policies on Archival Integrity and Removal, http://www.sims.berkeley.edu/research/conferences/aps/removal-policy.html (last visited Jan. 30, 2006).

[48] *Id.*

[49] *See* Answers.com, *supra* note 41.

[50] *Id.*

[51] Giga Blast Home Page, http://www.gigablast.com/ (last visited Mar. 3, 2007).

[52] Gigablast, About Us, http://www.gigablast.com/about.html (last visited Jan. 30, 2006).

[53] Fagan Finder Home Page, http://www.faganfinder.com/ (last visited Mar. 3, 2007). No affiliation with the author.

[54] Spurl.net Home Page, http://www.spurl.net/ (last visited Mar. 3, 2007).

[55] LookSmart's Furl Home Page, http://www.furl.net/ (last visited Mar. 3, 2007).

[56] Feedster Home Page, http://feedster.com/ (last visited Mar. 3, 2007).

Incy Wincy[59] - Incy Wincy is a search engine whose caching features function similarly to Google's.

C. The Issue

1. Web wrongs

A number of crimes and civil wrongs occur on the Internet, and some are particular to it. Copyright and trademark infringement are prevalent online, where users can post anything they want and "hide in the crowd" of the millions of other Internet users to keep from being prosecuted. Other crimes and civil wrongs, like libel, regularly occur online.

There are a number of reasons why it may be difficult to catch an infringer in the act. If an infringer suspects that he is being targeted for enforcement, he is likely to remove the offending material from the Internet. He might change his site, or remove the site from the Internet entirely. Web caching services aid enforcement by showing that the offending web site in fact contained infringing content. A potential plaintiff can use webcaching services as a fairly reliable way to determine whether their work has been infringed, or whether the potential defendant actually posted criminal material online. Web caching services can also verify a potential defendant's defenses (for instance, that the infringer had been using the author's purportedly copyrighted material before the author published it).[60]

2. A solution

Lawyers routinely use web caching services in their research. The practice has become so common that many attorneys, when presented with a web site, ask their assistants to "do a Wayback on that."[61] In cybersquatting cases,[62]

[57] Daypop Home Page, http://www.daypop.com (last visited Mar. 3, 2007).

[58] RSS is a format used to deliver short pieces of online content, including headlines, that may contain a link to a more detailed story. RSS is widely used by news services and weblogs.

[59] Incywincy: The Invisible Web Search Engine Home Page, http://www.incywincy.com/ (last visited Mar. 3, 2007).

[60] Kesmodel, *supra* note 3. (stating "In 2003, [Playboy's senior intellectual property attorney] says, the company cited the Wayback Machine during a court hearing to prove that a defendant used the term 'sex court' on his Web site only after Playboy aired a TV show with the same name. In his defense, the site operator asserted he had been using the name months before. The case was settled midtrial.").

[61] *Id.*

[62] Cybersquatting cases usually involve one party buying a domain name with a well-known trademark in it before the trademark owner purchases it. The cybersquatter then

attorneys are beginning to use the Wayback Machine "as a matter of course."[63]

Recently, lawyers have tried to introduce these cached pages as evidence during trials.[64] Opposing attorneys have raised a number of objections including hearsay (both single and double hearsay) and inherent unreliability.[65] Regardless of the cached pages' admissibility under the Federal Rules of Evidence ("FRE"), attorneys must consider other questions regarding cached pages' reliability and utility. An attorney will not want to rely on evidence if it is easy for opposing council to undermine it. Likewise, she will not want to rely on evidence if explaining what that evidence is, how she got it, and why it is useful confuses a jury. These problems are addressed below.

D. Recent Cases

A number of recent cases are relevant to this discussion. A growing number directly tackle the issue of cached web sites.

1. Telewizja Polska USA, Inc. v. Echostar Satellite Corp[66]

Echostar sought to introduce Wayback snapshots as evidence that Polska had maintained Echostar's trademarks on Polska's web site after their right to use the trademark had expired.[67] On October 15, 2004, the Court denied a motion *in limine* seeking to bar the snapshots, ruling that the cached pages are admissible under the Federal Rules of Evidence.[68] The plaintiff argued that the snapshots were hearsay and unauthenticated, unreliable sources.[69]

The Court replied, "to the extent these images and text are being introduced to show the images and text found on the web sites, they are not statements at all - and thus fall outside the ambit of the hearsay rule."[70] In response to the second contention, the Court said that Federal Rule of Evidence 901 leaves the question of authenticity and probative value up to the jury, and that "plaintiff is free to raise its concerns regarding reliability with the jury."[71] The Court also noted that Polska should be able to show that Echostar's snapshots are

offers to sell the domain name for an inflated price. This activity is actionable under the Lanham Act, 15 U.S.C.§ 1125(d)(1) (2000).

[63] Kesmodel, *supra* note 3.

[64] *Id.*

[65] Telewizja Polska USA, Inc. v. Echostar Satellite Corp., No. 02 C 3293, slip op. at 12-13 (N.D. Ill. Oct. 14, 2005).

[66] *Id.*

[67] *See* Kesmodel, *supra* note 3.

[68] *Id.*

[69] *See Telewizja Polska USA, Inc.*, slip op. at 13.

[70] *Id.*.

[71] *Id.* at 14.

inaccurate by introducing copies of its own archived web site updates.[72]

2. The "Jonathan Murder Trial"[73]

In February 2005, a Canadian criminal court declared a mistrial in the murder trial of a 12-year-old Toronto boy, because the prosecution's star witness had posted comments on a web site which the judge said damaged the witness's credibility.[74] A reporter for Canada's National Post discovered the comments using Google Cache and the Wayback Machine.[75] Without these tools, the trial would have progressed, as the defense attorneys were unaware of the existence of the archives and had only researched the witness with Yahoo!'s primary search engine.[76]

III. THE BENEFITS OF CACHING SERVICES

Cached web pages are more reliable than certain other kinds of evidence because they are not under the control of the plaintiff or defendant, but are instead usually controlled by a third, disinterested party. If a plaintiff notices their trademark on defendant's web site and takes a screenshot, a jury may be wary of that evidence. If the defendant denies that the plaintiff's screenshot is accurate, then the jury must decide who is more trustworthy in a "he-said-she-said" situation. On the other hand, if the plaintiff can present a cached version of defendant's web page from the Wayback Machine, along with verification from one of Internet Archive's employees, the value of the evidence is greater. This evidence is useful for a number of purposes, including investigation, discovery, settlement, and introduction at trial.

A. *Investigation*

Caching services can provide access to information that otherwise would be lost or destroyed. This is especially important in an investigation. If the police are searching for forensic evidence at an outdoor crime scene and it begins to rain, much of value will be washed away. Thanks to webcaching services, this is no longer the case in the Internet context. Even if a potential defendant tries to cover their tracks by removing content from the web, the information trail

[72] *Id.*

[73] Siri Agrell, *Teenage witness feared jeopardizing Johnathan trial, Web log reveals: Electronic evidence a first: Girl would have been 'torn apart' on stand over Post revelations, says professor*, NAT'L POST, Feb. 17, 2005, at A7, *available at* http://osgoode.yorku.ca/media2.nsf/0/9a4636df2c34903c85256fab006c8cb7?OpenDocument.

[74] *Id.*

[75] Kesmodel, *supra* note 3.

[76] *Id.*

can be restored with cached web pages. For example, Playboy Enterprises routinely does Wayback searches to find possible infringers of their Playboy Bunny trademark.[77]

B. Discovery

Cached pages can help to direct discovery.[78] Caching helps even the playing field in terms of access to information by giving investigators tools for viewing documents that might otherwise only be available to the other side. The presence of a cached copy of a web page could lead to a request for a webmaster to provide backup copies of their web pages on the date in question. Statements made on weblogs, forums, and on other web sites could lead to other useful evidence.

C. Avoiding Trial

Cached pages can encourage settlements. If a plaintiff can point to a caching service to show that she already knows that the defendant has done something wrong, the defendant will be less likely to want to stand trial. The stronger the evidence a potential plaintiff has before trial, the better her chances of extracting favorable terms in a settlement offer. An innocent defendant, on the other hand, can use cached pages in order to prove her innocence in cases where she does not have a backup copy of her own web site

D. At Trial

Cached pages are especially valuable for rebuttal. If a party claims to have had certain content on her web site (or to not have had certain content), another party can point to a cached web page to refute that assertion. For instance, computer maker Dell was able to use cached web sites recently in arbitration to show that Innerversion Web Solutions had registered the domain name "dellcomputerssuck.com" in bad faith.[79]

[77] *Id.*

[78] *See Discovery in a World of Electronic Documents and Data,* 5 SEDONA CONF. J. 151 (2004); David K. Isom, *Electronic Discovery Primer for Judges*, 2005 FED. CTS. L. REV. 1 (2005); Shira A. Scheindlin & Jeffrey Rabkin, *Electronic Discovery in Federal Civil Litigation: Is Rule 34 Up to the Task?*, 41 B.C. L. REV. 327 (2000); Amy K. Thompson-Smith, Lisa A. Bail & Lennes N. Omuro, *Coming to Terms with Electronic Discovery*, 9-FEB HAW. B.J. 4 (2005).

[79] Dell Inc. v. Innervision Web Solutions, National Arbitration Forum (May 23, 2005), available at http://www.arb-forum.com/domains/decisions/445601.htm.

Innerversion asserted that they had made a fair use of the Dell trademark. The court noted:

> Prior to notice of the dispute, Respondent was not making a bona fide offering of goods or services. The <dellcomputerssuck.com> domain name diverted Internet users to

Cached pages are useful for rebuttal because it puts the burden of undermining the cached evidence on the other party. For instance, if a defendant says that she did not display the plaintiff's copyrighted photographs, but the plaintiff can point to a cached snapshot of the defendant's web page with the photographs present, then the defendant appears less credible to the jury. Moreover, the defendant will have to show that the cached page is *not* accurate.

Attorneys can also use cached pages to directly prove a point.[80] This is most useful in a situation where a party denies that their web site had contained the content in question, but it puts the burden on the party introducing the evidence to show that it is accurate. Whereas a party introducing a cached page for rebuttal can rely on the page until the party's opponent attempts to undermine it, a party introducing a cached page for direct evidence must first defend it against the opposition's attacks by justifying its relevance and reliability. Relying on caching services for direct evidence also poses several other problems, which I will discuss next.

IV. PROBLEMS WITH CACHING SERVICES

A. *Limitations*

1. Limitations of web pages in general

Cached web pages generally have the same evidentiary problems as normal web pages, along with other issues particular to cached sites. If a web page provides purported facts (e.g. "98% of U.S. patents contain mistakes"[81]), then the source and accuracy of the information must be considered. Generally, cached web pages are not used to present outside facts like these. Primary sources are better in a number of respects, and web sites are usually cited only when information cannot be found in traditional print formats, or when

> Respondent's web site at the <innervisionpc.com> domain name where Respondent markets computer systems and related products and services that are in direct competition with Complainant. Respondent's use of the disputed domain name to redirect Internet users searching for Complainant to Respondent's competing web site is not a use in connection with a bona fide offering of goods or services or a legitimate non-commercial or fair use .

[80] *See* David T. Cox, *Litigating Child Pornography and Obscenity Cases in the Internet Age*, 4 J. TECH. L. & POL'Y 1 (Summer 1999). *See also* Ty E. Howard, *Don't Cache Out Your Case: Prosecuting Child Pornography Possession Laws Based on Images Located in Temporary Internet Files*, 19 BERKELEY TECH. L.J. 1227 (Fall 2004).

[81] Press Release, CPA, Mistakes Found in 98% of Patents Sampled from the U.S. Patent & Trademark Office (Jan. 13, 2006), *available at* http://www.cpaglobal.com/media_centre/press_releases/press_release_34.

providing a web site will make it much easier to access the information.[82] Cached web pages are only necessary as evidence of an outside fact for information which is available solely online and then subsequently removed. If researchers need to dig this deep to learn information, the accuracy and value of that information becomes questionable. This is a problem with all web page content.[83]

2. Indexing of cached pages

Cached web pages are very useful in other respects. In order to show that something was present on a web site on a certain day (e.g. "During the month of November, defendant's web site included a picture of plaintiff's trademarked logo."), cached web sites may be the only independent proof that exists.[84] However, cached web pages have their own problems that attorneys should consider.

Not every page on the Internet is archived. Pages can be removed from some archives at the request of the site owner.[85] Additionally, not every page is crawled to begin with. Because of the way that the Internet is indexed by webcrawling robots, even the largest search engines are constantly adding pages to their database.[86] Because some web pages are not crawled, not every page will be cached. If a search for a cached web site yields no results, this does not mean that the web site did not exist on the date in question. While that is a possibility, it may also be the case that the site was on the Internet, but at the time was "invisible" to webcrawlers, and therefore was not indexed. Even if the web page was visible, a crawler may simply not have accessed the

[82] *See, e.g.*, THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION R. 18.2.2-18.2.3 at 155-6 (Columbia Law Review Ass'n et al. Eds., 18th ed. 2005).

[83] *See* St. Clair v. Johnny's Oyster & Shrimp, Inc., 76 F. Supp. 2d 773, 775 (S.D. Tex. 1999) (noting"[a]nyone can put anything on the Internet. No web-site is monitored for accuracy and nothing contained therein is under oath or even subject to independent verification absent underlying documentation. Moreover, the Court holds no illusions that hackers can adulterate the content on any web-site from any location at any time. For these reasons, any evidence procured off the Internet is adequate for almost nothing, even under the most liberal interpretation of the hearsay exception rules found in FED. R. CIV. P. 807.").

[84] *See* Van Westrienen v. Americontinental Collection Corp., 94 F. Supp. 2d 1087, 1109 (D. Or. 2000) (admitting allegedly "false representations on [defendant's] Internet web site" into evidence for purposes of assessing punitive damages because the statements made on the web site were "admissions of [the defendant].").

[85] *See* The Internet Archive's Policies on Archival Integrity and Removal (Dec. 13-14, 2002), http://www.sims.berkeley.edu/research/conferences/aps/removal-policy.html.

[86] In 2004, Google added a billion web pages to its index, increasing the number of searchable pages by about one-third. *Google Adds 1 Billion Pages to Search*, CNN.COM, Feb. 18, 2004, http://www.cnn.com/2004/TECH/internet/02/18/google.expands.ap/.

site on the day requested.

However, there are ways around this limitation. If a page is not cached automatically by a webcrawler, it can still be cached manually by services like MyWeb.[87] In addition, it is possible for users to instruct certain services to add web pages to their index for future archiving.[88]

3. Self-help measures by web site administrators

Even if a webcrawler knows about a site, it still may not be able to archive it. Password-protected sites are not archived because when a webcrawler attempts to visit the site it will be unable to gain access to the content.[89] In addition, webmasters can instruct certain webcrawlers not to visit their site. They do this by creating a "robots.txt" file and placing it in a certain folder on their web site.[90] A robots.txt file is a file that webmasters can place in the main folder on their web server. Depending on what commands are written in this text file, some or all webcrawling robots will be prevented from accessing the site.[91]

Unfortunately, webmasters often create robots.txt files incorrectly. Even with a number of existing validators,[92] many robots.txt files contain mistakes.[93]

[87] *See infra* at II.B.2.

[88] The Wayback Machine provides three ways to add web pages to the crawler's list of sites to index: fill out a simple web form with the URL of the site; install a toolbar and visit a site; or while visiting a site with Internet Explorer, click the "show related links" option. *See* Internet Archive FAQ #1, *supra* note 42.

[89] Internet Archive Frequently Asked Questions, http://www.archive.org/about/faqs.php (follow "How do you Protect my Privacy if you Archive my Site?" hyperlink) (last visited Feb. 18, 2007).

[90] The Web Robots Pages, http://www.robotstxt.org/wc/robots.html (last visited March 8, 2007).

[91] The specification for the robots.txt file is maintained at http://www.robotstxt.org/wc/norobots.html.

[92] A validator checks a robots.txt file and determines if its syntax is proper. One example is at Google: Webmaster Help Center, http://www.google.com/support/webmasters/bin/answers.py?answer=35237&topic=8475 (last visited Feb. 18, 2007).

Validators check only for syntax errors. They will not check all aspects of a robots.txt file. For instance, if a webmaster intends to exclude only Googlebot while allowing other robots onto their site, a validator will tell the webmaster whether she formed her robots.txt file correctly, but it will not check to see that the robots.txt file does exactly what is intended. Thus, a validator is comparable to a spell-checker in word processing software, which checks whether words are spelled correctly but not necessarily whether they convey the author's intended meaning.

[93] Search Engine World inspected nearly 75,000 robots.txt files and learned that "more than 5% of the robots.txt used bad style and up to 2% were so badly formed that they would

These mistakes prevent webcrawlers from determining whether they are allowed on the site in question, and by default they will access it. These mistakes can allow persons interested in accessing a cached web page to do so even if the site author clearly intends to exclude caching services from accessing the site.

Other quirks of robots.txt files may afford unwanted visitors access as well. For instance, some webcrawlers ignore robots.txt files completely.[94] Additionally, webcrawlers generally have no "memory" for robots.txt files. This lack of memory may be an issue if a webcrawler accesses the site before the site's webmaster adds a robots.txt file. The webmaster may later decide to add a robots.txt that will instruct some caching services to remove old versions of the site from their indices. If the webmaster later removes the robots.txt file, *all* previously cached versions of the page will become available again.[95]

not be recognized by any spider." *See* Robots.txt Survey, GADGETOPIA.COM, http://www.gadgetopia.com/post/4137 (last visited Feb. 18, 2007).

[94] A Standard for Robot Exclusion, http://www.robotstxt.org/wc/norobots.html (last visited Feb. 18, 2007) (noting the robots.txt file standard "is not an official standard backed by a standards body, or owned by any commercial organisation. 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 [sic] by their robots.").

[95] For example, assume that it is April and that I have maintained a web page since January. Further, assume that the Wayback Machine caches my page on the first of every month. There will be cached copies of my page from January, February, March, and April. I am going to post some sensitive information on my page in May, and while I want it to be viewable while it is up, I will remove it at the end of May and I do not wish to allow caching services to serve old copies of my page from May.

If, at the end of April, I add a robots.txt file to my web site in order to exclude the Wayback Machine's webcrawler, the Wayback Machine will not make a copy of my site in May. The robots.txt file will also stop providing users with access to my cached sites from January, February, March, and April. Essentially, my web page will not be available on the Wayback Machine. However, if at the end of May I remove the robots.txt file from my site, the January, February, March, and April versions of my page will again be accessible to Wayback users. This is because, although I had a robots.txt file, Wayback does not delete the older versions of cached web sites that it had before the robots.txt file existed; it simply filters them out of its search results.

The May version of my site will never show up, because when the Wayback Machine's webcrawler attempted to make a snapshot of the May version of my page, it encountered my robots.txt file and moved on without looking at my site. No May snapshot exists to be served. *See Suit Claims that Accessing Data Protected by Web Site's 'Robots.txt' File Violates DMCA,* 10 BNA: ELECTRONIC COM. & L. REP. 882 (2005) (Plaintiff's previously cached web sites visited even though plaintiff used a robots.txt file).

This could be important to a webmaster who discovered that sensitive information was

4. Other limitations

Images are often left out of cache archives. This may occur for a number of reasons. Sometimes robots.txt files block access to the particular directory of a web site where the site's images are stored, while many times the caching service simply does not copy the image to begin with.[96] This could be problematic if the web page is being offered as proof that copyrighted or trademarked graphics were present on a web site. A similar problem exists for music and video files.

Each individual caching service has unique limitations. Google is limited to the last incarnation of the web site. The Wayback Machine does not record every update of a web page. MyWeb sites need to be stored manually by the user. Consequently, potential plaintiffs need to anticipate which pages will be important in future litigation. While computer-savvy individuals may have the foresight to manually cache a site that they think may be useful in a lawsuit, most will not consider doing so until it becomes apparent that litigation is imminent. By the time they contact an attorney for advice, the offending material may have been removed.

B. E-Evidence Tampering[97]

1. General web site attacks

It is tempting to believe that once a user retrieves the cached version of a web site through a caching service, the page presented is an accurate representation of what that page looked like on the day in question. But there are a number of ways to replace accurate information with inaccurate information online. A malicious attacker could place inaccurate information into the system at any number of points; caching services may copy the fake web pages, or have their originals replaced with fakes directly. Attorneys should heed George Orwell's warning: "If all others accepted the lie which the Party imposed—if all records told the same tale—then the lie passed into

being served by the Wayback Machine and decided to include a robots.txt file to protect that information. If the file is later taken down, all the older sensitive information will be accessible again. Of particular concern is the situation in which a domain name is sold and the new owners do not include a robots.txt file, which would allow old versions of the site (under the old owner) to be accessed.

[96] *See* Internet Archive FAQ #18, *supra* note 31.

[97] A number of methods for exploiting and compromising web sites are explained in this section. These examples are provided for educational purposes only and should not be attempted.

history and became truth."[98]

Web pages themselves can be hacked.[99] Judges note that the hacking of web sites has become common,[100] with even Microsoft's sites having been compromised.[101] If a web page is the subject of an attack before it is cached, then the webcrawler will cache the hacked version. In this situation, even though the original page author had no responsibility for the content on the hacked web page, all that appears in the cached archive is the defaced version, leading users to think that the defaced version is the author's intended original version.

2. Webcrawler exploits

A hacker needn't actually replace the target's page with his own, because other, subtler attacks can accomplish the same result. As a consequence of some of these attacks, a site owner likely wouldn't know that her site had been compromised until she actually saw the cached version of their page. Before a page is cached, webcrawling robots can be susceptible to a number of exploits involving temporary redirects. Most of us have encountered "redirects" when we try to access a web site that has moved: a message appears stating "web site has moved, you will be redirected to the new URL in five seconds." Then, without any user input, your browser takes you to the new site.

Web pages can redirect robots, like Googlebot, to other pages as well. This

[98] GEORGE ORWELL, 1984 29 (Plume Books 1983) (1948).

[99] A great deal of debate surrounds the use of the terms "hack," "hackers," and "hacking." While most lay people use "hacker" to refer to a person that gains unauthorized access to a system through technological wizardry, real "hackers" come in a number of flavors. At their most basic, a hacker is simply a person who creates or modifies computer software; the word has a neutral connotation. It can also mean a programmer that creates inelegant, "quick fix" solutions to programming problems. "White hat hackers" are "ethical hackers," often employed by companies and paid to exploit computer systems, in order to expose and correct vulnerabilities. What most people mean when they say "hackers" are actually "crackers," or "black hat hackers:" programmers that pursue illegal or illicit activities. Because this is the common usage, I simply use "hackers" in this Note to refer to black hat hackers.

See generally WHITE, *supra* note 13, at 347-349.

[100] A common kind of hack, and the one considered in this section, is a "web defacement attack," in which the legitimate web site is modified to display information that the hacker chooses, instead of the page created by the owner. For instance, in 2001 the Code Red Virus spread around the Internet; it replaced a site's content with the message "HELLO! Welcome to http://www.worm.com! Hacked by Chinese!" The virus is thought to have infected more than 250,000 hosts. *See* CERT, Advisory CA-2001-19 (2001), http://www.cert.org/advisories/CA-2001-19.html (last visited January 25, 2006).

[101] *Microsoft's Korean Web Site Hacked*, FOX NEWS.COM, June 2, 2005, *available at* http://www.foxnews.com/story/0,2933,158463,00.html

is potentially dangerous because a temporary redirect can cause a webcrawling robot to think that one page is actually another.[102] In this way, someone who was illegally serving copyrighted material on his web site could redirect webcaching robots to a non-infringing site, making it appear that his site was legitimate. Likewise, a malicious attacker could redirect robots from a non-infringing web site to an infringing one, making it appear that the innocent party had copied the material.

3. Attacks on the locations where cached pages are stored

Hackers can also attack the server farms of webcaching companies where cached web pages are stored. On November 13, 2005, Spurl was the victim of a web defacement attack. A hacker defaced one of the pages within Spurl's domain, SpurlTalk,[103] which contains recent news articles about Spurl.[104]

[102] This is called a "302 exploit," named for the HTTP status code returned to your web browser when you request a page that has a redirect instruction. In very general terms, the exploit looks like this:

1. The robot visits http://www.badguy.xyz
2. The web server at badguy.xyz responds with an HTTP 302 redirect that informs the robot that the content has been temporarily moved to http://www.victim.xyz/
3. The robot dutifully follows the redirect to http://www.victim.xyz
4. The robot receives content from the web server at www.victim.xyz and indexes it. However, because it believes that the content has been moved only *temporarily* it indexes it under the www.badguy.xyz domain instead of the www.victim.xyz domain.
5. Some time later, a user hits the robot's search service (Google in most examples) and types in some keywords that appear at http://www.victim.xyz. The search engine finds the keywords which it has indexed under www.badguy.xyz, so it returns a link to http://www.badguy.xyz.
6. The user selects the link and is taken to the http://www.badguy.xyz site where "bad guy" has complete control over the content.

Posting Of "Accidental Geek," http://www.webmasterworld.com/forum30/28329-24-30.htm (registration required) (Mar. 24, 2005 19:07 GMT). The example above would fool a search engine (so if the user searched for your web site, he might get mine instead). In the caching context, the robot would be redirected from a legitimate site to a fake "virtual site." The virtual site could be indexed under the legitimate site's name, allowing hackers to substitute their own content for the legitimate content in the caching service.

This exploit has been known since 2005, and many sites that employ robots are attempting to fix the problem. But exploits of this form still exist, and likely will continue to.

[103] The defaced page was http://stream.spurl.net/Spurltalk/, but is no longer available.

[104] Cached copies of the hacked page are available at http://www.furl.net/search?search=cache&id=5621329&url=http%3A%2F%2Fstream.spurl.net%2FSpurltalk%2F (free registration required), and through Yahoo! MyWeb (on file with author; available on request)

Using the same methods, a hacker can gain control of a page hosted on Spurl's servers and change the cached pages stored on those servers. The hacker could then falsify any web page on the site, thus potentially making a guilty party look innocent or an innocent party look guilty.

Hackers can attack the server farms of caching services remotely. However, server farms can also be the victim of direct human tampering. For example, unscrupulous employees of a caching service could replace the server farms' copies of web pages with substitute pages by accessing the servers directly from their place of employment. This scenario is unlikely, though, because a tampering party would need an agent "on the inside" in order replace offending copies.

C Jury Confusion

Assume that an attorney wants to introduce as evidence an accurate cached web page. Before the attorney can introduce it, she must be sure that the jury understands what it is. While most people can easily gain an understanding of webcaching basics, the opposing party is likely to challenge the admission of cached pages, especially now, as admissibility of this form of evidence is still being tested in courts. The explanation of how a document can be falsified can be technical and complicated. The battle over how trustworthy web pages are is likely to sow confusion in the minds of jurors who frequently won't understand the mechanics of the Internet.

D. An Unfair Advantage

1. Web site maintenance

Web sites are usually dynamic entities. Sites with news coverage change many times a day, while webmasters will update static sites to keep them factually accurate. A site may go through a complete change very often, or the webmaster may add content regularly without old content being taken off at any set time. In those cases, it may be difficult to decide when to back up a site. If a regular backup schedule *is* maintained, it is likely to miss an update every once in a while. If a web page is backed up once a week, but is occasionally updated twice in one week, then one of those updates will not be reflected in the back up copies.

On the other hand, site owners may not back up their web content at all. Usually, a web page is contained in a single file with one file name. When a site administrator updates that page, they will often simply overwrite it with the new copy. If content has been removed, there would be no record of it once the file is overwritten.

Even if backup copies do exist, they are susceptible to a wide variety of problems. Though web site source files are not usually large, they can grow in size, especially if they contain a number of graphic elements that are updated regularly and stored with the backup copies. Backup files of other forms of multimedia like sound and video can grow to even larger sizes. If backups are maintained, it can be hard to know how long to keep them around; there are not many uses for old copies of a web site, other than for novelty purposes, and many copies of old versions of the site can clutter a hard drive very quickly. Backup copies stored on tangible media can degrade, and backup copies stored on other computer systems may be corrupted. No system is foolproof.

2. Rebutting the cached version

A site administrator's credibility with a jury might be damaged if the administrator cannot produce a necessary backup copy. Many people may suffer from the "hindsight is 20/20" effect: obviously this backup file is very important now, and maintaining a backup seems so easy. Therefore, if a backup file was not maintained, it must have been because the site owner did not want anyone to see what was there on the date in question. Many people do not care how the Internet works. They know "the basics" of how the Internet operates and simply don't question the rest.[105]

This gives a potentially unfair advantage to the side attempting to present cached web pages. Stating "this is what was present on the web site in question on the date in question" is easy. While explaining how that information was acquired can be more complicated, it can be done in a non-technical way. Explaining why a webmaster does not maintain backup copies of every source file in a web site can be very complex and can leave jurors with the impression that one party is using technical jargon to cover their misdeeds.

[105] Chuck Klosterman aptly describes this phenomenon in his "low culture manifesto" *Sex, Drugs, and Cocoa Puffs*:

> In less than a decade, millions of Americans went from (1) not knowing what the Internet was, to (2) knowing what it was but not using it, to (3) having an email address, to (4) using email pretty much every day, to (5) being unable to exist professionally or *socially* without it. For 98 percent of the world, the speed and sweep of that evolution was too great to fathom. Consequently, we learned how to use tools most of us don't understand. This has always been the case with technology, but not quite to this extent. I mean, I drive a car that I can't fix and that I could certainly never build, but I still understand how it works in a way that goes (slightly) beyond the theoretical. I could explain how a car works to a ten-year old. Conversely, I don't understand *anything* about the construction of the Internet, beyond those conventional *Newsweek* factoids that everyone knows (and which still seem borderline impossible). I have no practical knowledge of the "information superhighway." And I'm not interested in how it works; I just want to feel like I vaguely grasp its potential and vaguely understand how to use that potential to my advantage.

CHUCK KLOSTERMAN, SEX, DRUGS, AND COCOA PUFFS 114 (Scribner 2003).

V. CACHED WEB PAGES UNDER THE FEDERAL RULES OF EVIDENCE

There is a great deal of literature analyzing the treatment of electronic evidence in court proceedings.[106] Treatment of cached web pages specifically is sparse, but the general analysis can be applied in the specific context of caching. Because the analyses are so similar, I sketch here only the broad outlines, highlighting the relevant differences between cached web pages and other forms of electronic evidence.[107]

A. *Objections*

1. Hearsay

Objectors call cached web pages hearsay for the same reasons as regular web pages: they are usually introduced to show what the author of the web site did or said without direct testimony that the author was in fact responsible for the content. Cached pages are different from web pages for these purposes, though. If an attorney objects to a web page as hearsay (as contrasted with “double hearsay” below), this use is likely to be excused by Rule 807 (see “Hearsay Exceptions” below). Cached web pages are under the control of a (theoretically) neutral third-party. Webcaching services generally cache web sites automatically, without regard to whose web site it is, which strengthens the implication that cached web sites are more helpful than harmful to the truth-seeking process.

2. Double Hearsay

In addition, cached web pages are susceptible to attack on the grounds that they constitute “double hearsay.”[108] If a regular web page by itself is hearsay, there is an additional problem with cached pages because there is an additional level of removal (e.g. “the Internet Archive says that my web page said that I

[106] *See* Gregory S. Johnson, *A Practitioner’s Overview of Digital Discovery*, 33 GONZ. L. REV. 347 (1997-98); Shira A. Sheindlin & Jeffrey Rabkin, *Electronic Discovery in Federal Civil Litigation: Is Rule 34 up to the Task?*, 41 B.C. L. REV. 327 (2000); Tracey L. Boyd, *The Information Black Hole: Managing the Issues Arising from the Increase in Electronic Data Discovery in Litigation*, 7 VAND. J. ENT. L. & PRAC. 323 (2005); Shannon M. Curreri, *Document in the Digital Landscape of Electronic Discovery*, 38 LOY. L.A. L. REV. 1541 (2005); Leah Voight Romano, *Electronic Evidence and the Federal Rules*, 38 LOY. L.A. L. REV. 1745 (2005); Adam Wolfson, Note, *“Electronic Fingerprints”: Doing Away with the Conception of Computer-Generated Records as Hearsay*, 104 MICH. L. REV. 151 (2005).

[107] For a more thorough look at the admissibility of electronic evidence, see Romano, *supra* note 106.

[108] *See* Telewizja Polska USA, Inc. v. Echostar Satellite Corp., No. 02 C 3293, slip op. at 12 (N.D. Ill. Oct. 14, 2005), *available at* http://cyberlaw.stanford.edu/packets/echostar.pdf.

said this.”). Cached web pages are stored on third party servers, so we have to take a third step to arrive at the original statement.

Cached web pages are problematic for double hearsay purposes because we must rely on the word of the archiver who cached the page originally. In turn, this cached page may be based on a piece of hearsay itself: the original web page. Because the caching is done automatically, there is no way for an archiver to say with absolute certainty that the archived version they produce is the same as the original web page on the date in question.

3. Hearsay exceptions

Although the *Echostar* magistrate said that cached web pages were not hearsay at all,[109] others might disagree. However, if a judge were inclined to view these documents as hearsay, an attorney might still be able to fit the cached copies of web pages into one of the hearsay exceptions. Two possibilities are the Rule 803(6) “business records exception”[110] and the Rule 807 “residual exception.”[111] Because digital sources like web sites do not normally degrade in transmission (and if they do, that can be readily determined), and because cached copies are created automatically in the normal operation of the Internet, a Rule 807 argument may be a valid option.

4. Inherent Unreliability

Attorneys have also objected that cached web sites are inherently unreliable. However, the magistrate in *Echostar* left the question of reliability for a jury, admitting the evidence over the objections of Telewizja.[112] The magistrate’s analysis is consistent with the treatment of the Federal Rules of Evidence,

[109] *Telewizja Polska USA, Inc.*, slip op. at 13 (“‘to the extent these images and text are being introduced to show the images and text found on the web sites, they are not statements at all—and thus fall outside the ambit of the hearsay rule.’” (quoting Perfect 10, Inc. v. Cybernet Ventures, Inc., 213 F. Supp. 2d 1146, 1155 (C.D. Cal. 2002)).

[110] Rule 803(6) allows “[a] record or data compilation, in any form” to be entered into evidence “if kept in the course of a regularly conducted business activity, and if it was the regular practice of that business activity” to make it. FED. R. EVID. 803(6).

[111] Rule 807 provides:

> A statement not specifically covered by Rule 803 or 804 but having equivalent circumstantial guarantees of trustworthiness, is not excluded by the hearsay rule, if the court determines that (A) the statement is offered as evidence of a material fact; (B) the statement is more probative on the point for which it is offered than any other evidence which the proponent can cure through reasonable efforts; and (C) the general purpose of these rules and the interests of justice will best be served by admission of the statement into evidence.

FED. R. EVID. 807.

[112] *Telewizja Polska USA, Inc.*, slip op. at 14 (quoting United States v. Harvey, 117 F.3d 1044, 1049 (7th Cir. 1997)).

which take a liberal view toward admissibility:

> Federal Rule of Evidence 901 'requires only a prima facie showing of genuineness and leaves it to the jury to decide the true authenticity and probative value of the evidence.' Admittedly, the Internet Archive does not fit neatly into any of the non-exhaustive examples listed in Rule 901; the Internet Archive is a relatively new source for archiving web sites. Nevertheless, Plaintiff has presented no evidence that the Internet Archive is unreliable or biased. And Plaintiff has neither denied that the exhibit represents the contents of its web site 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 [the Internet Archive representative's] affidavit is sufficient to satisfy Rule 901's threshold requirement for admissibility. Plaintiff is free to raise its concerns regarding reliability with the jury.[113]

B. Authentication

A party proffering a cached web site as evidence will need to authenticate the cached copy. In *Echostar*, a representative of the Internet Archive submitted an affidavit "verifying that the Internet Archive Company retrieved copies of the web site as it appeared on the dates in question from its electronic archives."[114] The court found this sufficient. So long as the authentication condition is met, it appears that cached web pages will satisfy Rule 901's requirement.[115]

VI. RECOMMENDATIONS

A. Consider Caching for Certain Causes of Action

Cached web pages are of great value to attorneys in certain fields. Cybersquatting, copyright infringement, trademark infringement, and dilution cases are prime examples of situations where cached web pages may be of help to a civil attorney.

Cybersquatting cases, governed by the Anticybersquatting Consumer

[113] *Id.*

[114] *Id.* at 13.

[115] Rule 901(a) states:

> General provision. The requirement of authentication or identification as a condition precedent to admissibility is satisfied by evidence sufficient to support a finding that the matter in question is what its proponent claims.
>
> FED. R. EVID. 901(a).

Protection Act ("ACPA"),[116] involve several time-sensitive inquiries. For instance, one factor indicating "bad faith intent to profit" under the ACPA is "the person's prior use, if any, of the domain name in connection with the bona fide offering of any goods or services."[117] Attorneys can use caching services to check for this factor.

In copyright infringement and trademark infringement cases, cached pages can help to show copying and use, respectively, of the plaintiff's intellectual property. Dilution is a particular kind of trademark infringement that provides a good opportunity for the use of caching services. Dilution requires that a party use a famous mark, but is only actionable "if such use begins after the mark has become famous."[118] An attorney could use caching services to help establish a prima facie case by showing that the opposing party had used the mark after the mark became famous.

In the criminal context, cached web pages can be used to investigate witnesses (especially their comments on bulletin boards), to verify wrongdoing, and to encourage plea bargains. Yahoo!'s MyWeb Community is particularly valuable. It separates cached pages by keyword, including tags like "blogs" and "blogging," "forums," and even illegal activities like "hacking."[119]

B. *Limit the Use of Cached Web Pages to Certain Contexts*

Attorneys can use cached web pages for investigation, research, and verification. Cached pages can be used to determine whether a crime or civil wrong was committed in the first place. Alternately, they can verify a potential defendant's story.

Furthermore, attorneys can use cached web pages in arbitration and settlement, where jury confusion will not be a problem (although the arbitrator still needs to understand caching). Cached web pages may encourage early settlement on more favorable terms.

Likewise, attorneys can use cached materials before trial (e.g. during discovery) to gain reliable evidence. In addition to the other extraneous evidence that a cached web site may point an attorney towards, such as company records, a cached web site could itself lead to an actual, authentic backup copy of the web site. If an attorney has a dated cached copy of a web site, it may be possible to request the site owner's backup copy from that date,

[116] 15 U.S.C. § 1125(d)(1) (2006).

[117] 15 U.S.C. § 1125(d)(1)(B)(i)(III).

[118] 15 U.S.C. § 1125(c)(1).

[119] *See* Yahoo! Search MyWeb (Beta), Everyone's Tags, http://myweb.yahoo.com/myweb?ei=UTF-8&dg=6&dmode=vtags&sortby=count (last visited Mar. 9, 2007).

if one exists. If an exact date is not available, a range of dates can be identified by examining the different cached versions available online. For instance, if Yahoo's cached version is one month old and does not contain the information you are looking for, but Google's version is one week old and does contain this information, it is clear that the page was posted at some point during the intervening three weeks. The secondary caching services can be useful for this purpose because they often have a wider date range than the primary three. Also, an attorney could look for clues leading to a date range in the cached web page itself (links to news articles, references to current events, etc.). If no backup copy is available, it may be possible to have the site owner authenticate the web pages in question herself. Even if the site owner refuses to authenticate a page, this places the burden on her to show that the cached copy is inaccurate.

Cached web pages are especially useful for rebuttal. If a site owner makes a statement that contradicts what is on a cached web page, the cached web page may cast doubt on the site owner's statement. If the owner outright denies that the copy is accurate, the burden is on her to produce a backup copy showing the accurate version. If she claims that she is not sure if a cached copy is accurate, the jury may doubt the reliability of the site owner's memory.

C. Use Multiple Caching Services

It is important for attorneys to remember that if a page is missing from caching services' indices, it does not mean that that page did not exist at some point. The site author may have removed the page.[120] It may be possible to determine this by looking in the web site's root directory for a "robots.txt" file. Also, attorneys should remember that removing the robots.txt file will restore all previous cached copies on the Wayback Machine. If the file is taken down, the pages will reappear on Wayback the next time that the Internet Archive's webcrawler visits the site. If a robots.txt file is removed, a user can force a crawl by Wayback by installing the Alexa toolbar[121] and simply visiting the site Sites will be crawled 24-48 hours later, and a missing robots.txt file will restore previous archived copies.[122]

If the site is not cached in one caching service, it may be in another, especially one of the lesser-known ones. Yahoo! and Spurl's communities may be especially useful where a site may not have been automatically cached by Google or Wayback, but may have been manually cached by some other user. This possibility overcomes some of the limitations on cached web sites.

[120] *See* IV.A.2 – IV.A.3 *supra*.

[121] *See* Internet Archive FAQ #1, *supra* note 42; Alexa Toolbar, http://www.alexa.com/site/download/ (last visited Mar. 9, 2007).

[122] *See* Internet Archive FAQ #1, *supra* note 42.

A potential defendant might be able to defend herself against an inaccurate cached web page by showing that a different version of the web page exists on a different caching service. "A contradiction cannot exist . . . To arrive at a contradiction is to confess an error in one's thinking; to maintain a contradiction is to abdicate one's mind and to evict oneself from the realm of reality."[123] For example, if a plaintiff points to an erroneous cached page from Spurl that shows copyright infringement, the defendant could present a non-infringing version on the Wayback Machine. Bearing this in mind, if it is necessary to rely on cached copies, attorneys should use cached copies from several corroborating sources. For instance, they should check Google and the Wayback Machine to see if they contain identical copies of a web site for identical dates. Or, if using Yahoo! My Web, attorneys could store alternate copies with services like Spurl or Furl. Attorneys should also download the source code of the web page[124] as well, though this will only prove what was on the site, not where the page originated. Webcaching use should be limited at trial to avoid jury confusion. Finally, attorneys should avoid getting into a battle over technical details.

VII. CONCLUSION

Caching services can be a useful tool, and will probably stand up to scrutiny under the Federal Rules of Evidence. However, they have several limitations, and thus should not be relied upon heavily in court. The Federal Rules allow parties to admit a wide array of evidence, and then leaves it to a jury to decide which pieces are relevant and how much weight to give them. Attorneys need to take into consideration some important factors before introducing cached web pages as evidence, including the fact that the opposition can attack the procedures utilized by web caching services. In an attempt to explain these procedures, it will be very easy for an attorney to become bogged down in technical details, which could cause a jury to significantly downplay the importance of the cached page.

Bearing this in mind, cached web pages can be particularly useful in a number of situations. They can save time and resources during an investigation, they can help to direct discovery requests, and they are especially valuable in rebutting testimony.

The legal system's use of cached web pages may have other positive side effects, as well. Returning to the *Minority Report* illustration I described in my introduction, catching a criminal after the fact is not as effective as stopping them before they act. However, if online actors knew with near certainty that

[123] AYN RAND, FOR THE NEW INTELLECTUAL 126 (Penguin Books 1961).

[124] This can be done in most web browsers by selecting the View menu, the Page Source or Source menu item, and then selecting Save from the File menu.

their actions on the Internet could be traced by law enforcement after the fact, the likelihood of being caught would deter many of them. And in that sense, we *would* be stopping bad acts before they occurred.