JONATHAN BAND PLLC

TECHNOLOGY LAW & POLICY

21 Dupont Circle NW
Washington, DC 20036
policybandwidth.com

jband@policybandwidth.com
(202) 296-5675 (phone)
(202) 872-0884 (facsimile)

A NEW DAY FOR WEBSITE ARCHIVING: *FIELD v. GOOGLE* AND *PARKER v. GOOGLE*

SUMMARY

As the World Wide Web has evolved into a major means of distributing information, librarians have recognized that archiving websites is an important element of cultural and historic preservation. However, the threat of copyright liability has prevented the systematic archiving of these sites by libraries. Fortunately, two recent district court decisions, *Field v. Google*, 412 F. Supp. 2d 1106 (D. Nev. 2006), and *Parker v. Google*, 422 F. Supp. 2d 492 (E.D. Pa. 2006), provide several viable legal theories for library website preservation.

The district court in *Field* identifies two strong defenses against copyright liability for libraries seeking to engage in the systematic archiving of websites: implied license and fair use. These defenses will be strongest with websites that do not employ "no-archive" meta-tags. The *Field* decision also suggests that the display of an archived site might not constitute a direct infringement by the library because it does not involve a volitional act by the library. The *Parker* decision could be understood as applying this volitional act theory to both the reproduction and the display of the website. In the absence of authority to the contrary, these two district court opinions from different circuits provide a solid basis for arguing that copyright law currently permits libraries to archive websites.

INTRODUCTION

The Supreme Court has observed that the Internet is "a unique and wholly new medium of worldwide communication." *Reno v. ACLU*, 521 U.S. 844 (1997). It "constitutes a vast platform from which to address and hear from a worldwide audience

of millions of readers, viewers, researchers and buyers." *Id*. at 853. The Court marveled at the "vast democratic fora of the Internet," *id*. at 868, including thousands of newsgroups, "each serving to foster an exchange of information or opinion on a particular topic running the gamut from, say, the music of Wagner to Balkan politics to AIDS prevention to the Chicago Bulls." *Id*. at 851. While the Internet has made more information more accessible to more users than any previous medium, it has a significant drawback: impermanence. Unlike a published book, which can remain accessible to users for centuries through libraries, the information on the World Wide Web is available only as long as it is stored on a computer attached to the Internet.

As the Internet has evolved into an increasingly important means of distributing government information, news, financial data, and artistic expression, librarians have become aware of the need to find a way to preserve this information of significant historical and cultural value. The vast size and dynamic nature of the Internet make website preservation an overwhelming technical task. But perhaps even more daunting than the technical challenge of gathering, storing, and indexing millions of constantly changing websites is the legal barrier of copyright. Much of the content of these websites falls within copyright's protection, and libraries will not copy this content if doing so will subject them to legal liability. While many librarians believed that the fair use privilege, 17 U.S.C. 107, would permit website preservation, there was no case law sufficiently on point to provide librarians with enough certainty to justify proceeding with large-scale website archiving projects. However, the recent decisions in *Field v. Google*, 412 F. Supp. 2d 1106 (D. Nev. 2006), and *Parker v. Google*, 422 F. Supp. 2d 492 (E.D. Pa. 2006),[1] provide several viable legal theories for library website preservation.

FIELD v. GOOGLE

An attorney named Blake Field "decided to manufacture a claim for copyright infringement against Google in the hopes of making money from Google's standard practice" of caching the content it spiders and presenting users with links to the cached

[1] *Parker* currently is on appeal to the Third Circuit.

copies. 412 F. Supp. 2d at 1113. He wrote 51 short stories and posted them on his website. Field was aware that he could use a "bot exclusion" header in a robot.txt file, which would instruct Google's software that gathers material from the Web – the "Googlebot" -- not to copy the site into Google's search index. He also was aware that he could include a "no archive" meta-tag in the robot.txt file, in which case Google would copy the site into its search index, but it would not keep a cached copy of the site, nor provide users with links to the cache. Instead, he set the permissions within the robot.txt file to allow all search engine "bots," including Google's, to visit and index the site, and he intentionally did not employ a "no-archive" meta-tag. Detecting neither a bot exclusion header nor a "no archive" meta-tag, Google crawled Field's site, included his stories in its search index, and provided links to cached copies of the stories.

Field sued Google for copyright infringement. He argued that the Google Cache feature, by allowing Google users to link to archived copies of his stories indexed by Google, infringed his copyrights. Field did not argue that Google infringed by virtue of copying his stories into its index in the first place. He presumably was familiar with *Kelly v. Arriba Soft*, 336 F.3d 811 (9th Cir. 2003), where the Ninth Circuit found that a search engine's storage and display of thumbnail-sized images copied from websites, with links to the full-sized images on the original websites, constituted a fair use. Since *Kelly* is binding precedent on the federal court in Nevada, Field likely assumed that a court would treat Google's copying and storing of his stories as a fair use. By centering his complaint on Google's serving the cache copy of his stories to users, rather than on Google's act of making and storing the cache copy, Field probably hoped to distinguish his case from *Kelly*. The *Kelly* court's fair use ruling was based in large measure on its conclusion that a display of a thumbnail sized image did not supersede the display of a full-sized image, and that a user still had to go to Kelly's site to see the full-sized image. Here, by contrast, the stories served from the Google cache were identical to the stories displayed on Field's website.

Nonetheless, the district court granted summary judgment in favor of Google on five independent bases, several of which have direct applicability to website archiving by libraries.

A. No Volitional Act. The *Field* court found that Google did not directly infringe copyright by serving a webpage from the Google Cache, because Google did not engage in any volitional activity with respect to the serving of the cached webpage. Rather, a user initiated the serving of the cached copy by clicking on the cache link. Then, Google's servers displayed the copy to the user by an automatic process. In reaching this conclusion, the district court relied on *Religious Tech. Ctr. v. Netcom On-Line Commc'n Servs., Inc*., 907 F. Supp. 1361, 1369-70 (N.D. Cal. 1995), where the court held that direct infringement requires a volitional act by the defendant, and that the automated copying by machines owned by the defendant was insufficient to trigger liability. The *Field* court also relied on *CoStar Group, Inc. v. LoopNet, Inc*., 373 F.3d 544, 555 (4th Cir. 2004), where the Fourth Circuit stated that "[a]greeing with the analysis in *Netcom*, we hold that the automatic copying, storage, and transmission of copyrighted materials, when instigated by others, does not render an ISP strictly liable for copyright infringement…."

The *Field* court only found that Google's display of the cache copy did not involve a volitional act by Google. It did not consider whether Google's caching of the copy by sending out the Googlebot to crawl and copy websites involved volition because those acts were not the subject of this lawsuit. Thus, the *Field* court's holding concerning volitional acts protects libraries only with respect to their providing access to archived websites in response to user requests. Since the decision did not reach Google's caching of websites, there is no way to know whether the *Field* court would find volition in a library's automated copying and storing of websites. Still, at the very least, the volitional act theory could protect libraries in connection to an important part of the website archiving function.

B. Implied License. More helpful to libraries is the court's holding that Field's posting an "allow all" robot.txt header and then failing to set a "no archive" meta-tag

indicated that he impliedly licensed search engines to permit users to access the cached copies. The court found that "the 'no-archive' meta-tag is a highly publicized and well-known industry standard" of which Field was aware. 412 F. Supp. 2d at 1116. The court further found that "Google reasonably interpreted the absence of meta-tags as permission to present 'Cached' links to the pages of Field's site." *Id.* On this basis, the court concluded that Field granted Google an implied license for this use.

This ruling suggests that libraries can argue that website operators have authorized libraries both to archive and display their sites if the operators fail to employ a "no archive" meta-tag. This implied license defense would work best with respect to actions a library takes prior to receiving a cease-and-desist letter from the website operator. However, once the library received such a letter, it could not argue that it had an implied license to continue displaying the archived website. Nonetheless, the library might succeed in convincing a court that since it reasonably relied on the absence of the "no-archive" meta-tag when it initially cached the website, it should be permitted to continue to display the cache. While a court might permit the library to continue displaying the archived website, it probably would not permit the library to archive updates to the website.

It should be noted that this implied license defense has a significant practical limitation: many websites with valuable content, such as government websites, often employ "no-archive" meta-tags or bot exclusion headers. A library obviously could not claim an implied license to archive such websites.

C. Fair Use. The *Field* court extended *Kelly* from the display of thumbnail images to the display of the complete text of web sites. Considering the first fair use factor, the purpose and character of the use, the *Field* court found that the display of cached copies served at least three different purposes from the original work and therefore was transformative and did not supersede the original.

First, and most relevant to library archiving,

> Google's 'Cached' links allow users to view pages that the user cannot, for whatever reason, access directly. A Web page can become inaccessible to Internet users because of transmission problems, because nations or service providers seek to censor certain information, because too many users are trying to access the same page at the same time, or because the page has been removed from its original location. In each case, users who request access to the material from the inaccessible site are still able to access an archival copy of the page via the 'Cached' link in Google's search results. Google's users, including those in academia, describe this functionality as highly valuable.

Id. at 1111 (citations omitted). The court further explained that this archiving function also benefited the original website publishers because it allowed them "to recover copies of their own sites that might otherwise have been lost due to computer problems." *Id*. at 1112 (citation omitted). The court observed that the "Internet is replete with references from academics, researchers, journalists, and site owners praising Google's cache" for providing this archiving functionality. *Id*. at 1118. When Google provides access to an otherwise inaccessible website, "Google's archival copy of a work obviously does not substitute for the original." *Id*.

Second, the cache copy allows users to detect changes to a website, which can be significant for political, educational, and legal reasons. A user can identify changes only by comparing the current website with the original archived by Google. Third, the cache copy allows a user to understand why a page is responsive to a query, because the queried term is highlighted in the cache copy.

In sum, even though the cache copy is identical to the original, the court found the display of the cache copy to be transformative because it "serves different and socially important purposes" from the original. *Id*. at 1119. If Google's archiving is transformative, then a library's preservation of websites is *a fortiori* transformative.

Field argued that Google's commercial objectives cut against fair use. The court dismissed this argument by stressing that there was no evidence that Google profited from its use of Field's stories: "Field's works were among *billions* of works in Google's

database." *Id*. at 1120. A library could make an even stronger first factor argument than Google because its archiving obviously serves a non-commercial preservation purpose.

Turning to the second factor, the nature of the copyright works, the court noted that Field had made his works "available to the widest possible audience for free" through his website. *Id*. The court evidently concluded that this sort of use by its owner implied that the work was not that creative and not that deserving of protection.

When discussing the third factor, the amount of the work used, the district court observed that the "multiple and socially valuable purposes" of the cache could not be achieved by the display of only portions of web pages:

> Without allowing access to the whole of a web page, the Google Cached link cannot assist Web users (and content owners) by offering access to pages that are otherwise not available. Nor could use of less than the whole page assist in the archival or comparative purposes of Google's 'Cached' links. Finally, Google's offering of highlighted search terms in cached copies of Web pages would not allow users to understand why a Web page was deemed germane if less that the whole Web page were provided.

Id. at 1121. In sum, Google "displayed no more of the works than is necessary" to accomplish its purpose. *Id*.

With respect to the fourth factor, the market impact of the cache copy, the court noted that Field made his works available for free on the Internet, and never received any revenue for them from any source. Since there was no evidence of any market for Field's works, there could be no harm to that market. Further, the court found that "there is no evidence ... of any market for licensing search engines the right to allow access to Web pages though 'Cached' links, or evidence that one is likely to develop." *Id*. at 1122. To the contrary, the court observed that

> [t]here is compelling evidence that site owners would not demand payment for this use of their works. Notwithstanding Google's long-standing display of 'Cached' links and the well-known industry standard protocols for instructing search engines not to display them, the owners of literally billions of Web pages choose to permit such links to be displayed.

Id. In the absence of a market for the licensing of "caching rights," Google's caching of his

stories did not deprive him of any revenue. So, too, a library's preservation of a website typically would not deprive the website operator of any revenue. Website archiving could conceivably affect the market for a commercial website by diverting traffic from that site, but presumably such a site would use either a bot-exclusion header or a "no-archive" meta-tag. If such an operator failed to use a "no-archive" meta-tag, he would have difficulty convincing a court that the site's content had any commercial value that archiving could harm.

The *Field* court then considered a fifth, non-statutory, fair use factor: "whether an alleged infringer has acted in good faith." *Id.* The court found that Google honored industry standard protocols such as bot exclusion headers and "no-archive" meta-tags. Google's website provided website operators with explanations on how to deploy these protocols. Google also provided an automated mechanism for a website operator to prevent the further display of a cached website after Google cached it. Finally, in this instance, Google disabled the links to the cache as soon as Field filed his complaint. Hence, the court concluded that Google acted in good faith.

While libraries could easily emulate all of Google's caching policies, they might be reluctant to provide an automated mechanism for website operators to prevent access to archived material. Because a library understandably would oppose outside interference with its preservation decisions, a court would likely still conclude that a library acted in good faith even if it did not adopt this particular feature.

However, a library might have a more difficulty using *Field*'s fifth factor as a precedent if the library ignored "no-archive" meta-tags. And, as noted above, a library's position on market harm is stronger if it respects "no-archive" meta-tags. Although a library may still prevail in its fair use defense if it archived a website in contravention of a "no-archive" meta-tag, the library is more likely to succeed in cases involving websites that did not employ "no-archive" meta-tags.

4. Digital Millennium Copyright Act. The *Field* court held that the Section 512(b) caching "safe harbor" for online service providers in the Digital Millennium Copyright Act applied to Google's caching of website content. Section 512(b) limits a

service provider's liability for the intermediate and temporary storage of material on its system if (a) the material is made available online by a person other than the service provider; (b) the material is transmitted from the person who made the material available online to another person at the direction of that "other person;" and (c) the storage is carried out through an automatic technical process for the purpose of making the material available to users. The *Field* court concluded that the "other person" in (b) could be Google itself: "Field transmitted the material in question, the pages of his Web site, to Google's Googlebot at Google's request. Google is a person other than Field. Thus, Google's cache meets the requirements of Section 512(b)(1)(B)." *Id*. at 1124.

The system caching safe harbor has typically been understood to protect a service provider that cached material in the course of the material's transmission from a website to an end-user. It was assumed that for purposes of Section 512(b), the "other person" was the end-user, and not the service provider. At the same time, the *Field* court's interpretation is not inconsistent with the plain language of the statute; the service provider is a person other than the person who made the material available online. Thus, on its face, Section 512(b) could apply to website archiving.

Nonetheless, Section 512(b) might be of limited utility to libraries. Field argued that Section 512(b) did not apply because Google's storage was not "intermediate and temporary." The court noted that Google's cache was "intermediate" because it was "a repository of material that operates between the individual posting the information, and the end-user requesting it." *Id*. at 1124. With respect to "temporary," the court noted that the Ninth Circuit in *Ellison v. Robertson*, 357 F.3d 1072 (9th Cir. 2004), considered AOL's storage of a Usenet posting for 14 days to be "transient" for purposes of 17 U.S.C. § 512(a). Accordingly, the *Field* court found that Google's cache of material for 20 days to be "temporary" under Section 512(b). While a library might succeed in arguing that its archival copy is intermediate because its objective is providing end-users with access to the material, a library presumably will want to archive websites for far longer than 20 days. Courts probably will not treat a copy that resides in a computer server for

years as "temporary."

While Section 512(b) does not appear to provide libraries with adequate protection for website archiving, the *Field* court's holdings with respect to implied license and fair use provide libraries with two viable defenses. Additionally, although the *Field* court's ruling concerning volitional acts might not help libraries with respect to the archiving of websites, it could assist libraries with the display of archived sites.[2]

PARKER v. GOOGLE

Just two months after the *Field* decision, a federal district court in Pennsylvania issued a decision that built on some of *Field*'s holdings. Parker represented himself in the case, so the court had difficulty understanding exactly what he was claiming. As a result, the opinion is unclear on some points of fact and law. Parker, an author, posted his writings on his website. He also posted them to Usenet groups. Google automatically archived his Usenet postings. Additionally, Google cached and displayed the content of his website. Finally, when Google presented a list of hyperlinks in response to search queries, it excerpted portions of his website.

Parker claimed that these Google actions constituted direct copyright infringement. The court granted Google's motion to dismiss. Relying on *CoStar* and *Netcom*, the court found that Google had not engaged in a volitional act:

> When an ISP automatically and temporarily stores data without human intervention so that the system can operate and transmit data to its users, the necessary element of volition is missing. The automatic activity of Google's search engine is analogous. It is clear that Google's automatic archiving of USENET postings and excerpting of websites in its results to users' queries do not include the necessary volitional element to constitute direct copyright infringement.

---

[2] The *Field* court also ruled that Field was estopped from asserting a copyright claim because he induced Google to infringe by using software code that invited Google to scan and cache his website, but then intentionally failed to instruct Google not to serve the cached copies. This defense turns on the peculiar facts of this case, and would not assist libraries when they sought to archive most websites.

422 F. Supp. 2d at 497. Here, the court appears to suggest that the archiving of a website, if performed automatically and without human intervention, does not involve a volitional act. Thus, the *Parker* court seems to go beyond the *Field* court. While the *Field* court ruled that the display of an archived work did not involve a volitional act by Google, the *Parker* court indicated that the initial archiving did not involve a volitional act either.[3]

Even if a library does not engage in a volitional act when archiving a website or displaying it, could it nonetheless be secondarily liable for the infringement made by the user when he accesses the site? After all, the user makes a copy in the random access memory of his computer, and he might also print out a hard copy. *Parker* suggests that a library would not meet the knowledge requirement for contributory infringement. It found that Google never had the requisite knowledge of the copying facilitated by its automatic systems. So, too, a library would not have knowledge that a specific use made by a specific user was infringing, until after the website operator informed the library that it did not want the library to archive and display its website. If the website did not employ a bot exclusion header or a "no-archive" meta-tag, the library could reasonably assume that the website operator was permitting not only the library's reproduction and display of the site, but also the user's RAM and hard copies.

Parker also suggests that a library would not satisfy the criteria for vicarious liability. The *Parker* court did not find any evidence of financial benefit resulting from any specific act of infringement. The court quoted with approval Nimmer's language that "[l]arge commercial ISPs derive insufficient revenue from isolated infringing bits, in the context of the billions of bits that cross their servers, to characterize them as financially benefiting from the conduct of which complaint is made." *Id.* at 500. The court also noted that in *Ellison*, the "record lack[ed] evidence that AOL attracted or retained subscriptions because of the infringement or lost subscriptions because of AOL's eventual obstruction of the infringement." *Id.* If the *Parker* court found that Google did

[3] Citing *Field* as authority, the *Parker* court also found that caching safe harbor at Section 512(b) of the DMCA sheltered Google's caching of web pages as a means of indexing websites and producing results to search queries. The *Parker* court did not address the meaning of the term "temporary storage."

not benefit financially from the infringing conduct of its users, it is hard to imagine that any court would find that a library benefits financially from its users' infringing use of it archive.

Finally, and perhaps most importantly, a library can be secondarily liable for its users' activities only if those activities are themselves infringing. In the overwhelming majority of situations, the user of a library's website archive would have a strong fair use argument excusing her from copyright liability.

CONCLUSION

The *Field* court provides two powerful defenses against copyright liability for libraries considering the systematic preservation of websites: implied license and fair use. These defenses will be strongest with websites that do not employ "no-archive" meta-tags. *Field* also suggests that the display of an archived site might not constitute a direct infringement by the library because the display does not involve a volitional act by the library. The *Parker* decision could be viewed as applying this theory to both the reproduction and the display of the website. In the absence of authority to the contrary, these two district court opinions from different circuits provide a solid basis for arguing that copyright law currently permits libraries to archive websites.

Moreover, library archiving of websites will take place in an environment where all the leading search engines, including Google, Yahoo, MSN, and Ask, routinely cache millions of websites. A court will evaluate a library's implied license and fair use defenses the context of this industry practice.