3 B.U. J. Sci. & Tech. L. 12 April 15, 1997

Boston University
Journal of Science & Technology Law

Note

Protecting the Sweat of the Spider's Brow:
Current Vulnerabilities of Internet Search Engines

Michael J. Schmelzer

Table of Contents

Protecting the Sweat of the Spider's Brow: Current Vulnerabilities of Internet Search Engines[†]

Michael J. Schmelzer[*]

I. INTRODUCTION

1. Like the Internet itself, the World Wide Web[1] has experienced extraordinary growth in a rapid, organic, and disorganized fashion.[2] Web search engines represent an attempt to bring organization and manageability to the Web by indexing its contents. Although web search engines, also known as "robots" or "spiders," began as academic exercises, most of these engines are now commercial enterprises.[3] As commercial enterprises, most spiders currently rely on advertising for revenue. Any attempt to bypass the spider operators' advertising therefore represents a threat to their viability.

2. This paper examines the protections offered to spiders under current law. Although bypassing a spider operator's advertising is easy and prevents the advertising from being seen by its intended audience, there are valid reasons for a third party to do so.

3. Part II provides the history and technical information necessary to understand the issues involved. Part III examines the applicability of copyright law

[†] © 1997 Trustees of Boston University. Cite to this note as 3 B.U. J. SCI. & TECH. L. 12 (1997). Pin cite using the appropriate paragraph number. For example, cite the first paragraph of this note as 3 B.U. J. SCI. & TECH. L. 12 para. 1 (1997).

[*] A.B., 1988, Princeton University; M.S., 1990, Harvard School of Public Health; M.S., 1994, University of Wisconsin; J.D. (anticipated), 1997, Boston University School of Law.

[1] The World Wide Web ("the Web") is a "graphics-intensive environment running on top of the Internet." WIRED STYLE: PRINCIPLES OF ENGLISH USAGE IN THE DIGITAL AGE 111 (Constance Hale ed., 1996) [hereinafter WIRED STYLE].

[2] *See* ACLU v. Reno, 929 F. Supp. 824, 831 (E.D. Pa. 1996), *cert. granted*, 117 S. Ct. 554 (1996); *see* John W. Verity, *What Hath Yahoo Wrought?*, BUS. WK., Feb. 12, 1996, at 88, 88.

[3] *See* Verity, *supra* note 2, at 89; *see, e.g., History* (last modified Dec. 1, 1996) <http://www.yahoo.com/docs/pr/history.html> (noting that Yahoo began while its founders were at Stanford University).

to the query results returned by a spider. Part IV examines the feasibility of licensing in the absence of sufficient copyright protection. Part V examines the applicability of passing-off and unfair competition law on both the state and federal level. Part VI predicts how this pivotal field will develop.

II. BACKGROUND, HISTORY, AND TECHNICAL INFORMATION

A. *Growth of the Internet*

4. The Internet is a global network of networks that has experienced extraordinary growth in recent years. As of mid-1996, the Internet linked more than 9,400,000 computers.[4] The Internet can exchange files between computers.[5] The standard file exchange method uses the file transfer protocol ("ftp"). With ftp, a user on one computer can share files with any other computer on the Internet. The ability to transfer files led to the development large file archives accessible to anyone connected to the Internet.

5. As publicly available archives proliferated, it became more difficult to track them. In 1989, students at the McGill School of Computer Science[6] developed a system called “Archie"[7] that performed this task automatically. Archie "harvested" its data by periodically querying all known ftp archives, retrieving a list of each one's files, and then making the list accessible to queries.[8]

6. Although the rapid proliferation of archives made vast amounts of software available, archive access still required some technical sophistication.[9] To simplify this process and provide access to many Internet resources, a group from

[4] *See ACLU*, 929 F. Supp. at 831.

[5] The moving of files from one computer to another was one of the reasons behind the development of the Internet's precursor, the ARPAnet. *See id.*

[6] For more information on the McGill School of Computer Science, see *School of Computer Science’s Home Page* (last modified Jan. 12, 1997) <http://www.cs.mcgill.ca/>.

[7] Archie is an archive without the “v.” *See* DANIEL P. DERN, THE INTERNET GUIDE FOR NEW USERS 315 (1994).

[8] *See* DERN, *supra* note 7, at 331-32. This distinction between "harvesting" and "querying" is important and will be discussed *infra*. By late 1992, Archie catalogued over 1,500 sites containing a total of more than 2.5 million files occupying over 230 gigabytes of storage.

[9] For example, due to technical limitations, Mac and PC programs could not be stored on archives in their native "binary" format, but had to be converted to text files in order to insure proper storage and transfer. If a user wanted to download a PC-compatible executable, she would have to download the uuencoded file, and then "uudecode" the file back into its executable format. A user without a uudecoder program already on her PC would not be able to utilize Internet archives.

the University of Minnesota developed Gopher.[10] Gopher software provided a simple and unified interface to a variety of Internet resources such as Archie, ftp, Usenet news,[11] and Wide Area Information Server ("Wais").[12] "Veronica," an indexing system, helps users navigate the expanding "gopherspace."[13]

7. These Internet tools and resources have functional analogs in the Web: web sites fill the function of ftp archives; the hypertext transfer protocol ("HTTP") replaces the ftp protocol; browsers such as Netscape Navigator[14] and Microsoft Internet Explorer[15] replace the all-encompassing interface of the gopher client;[16] and spiders perform the indexing previously accomplished by Archie and Veronica.

B. *Growth of the World Wide Web*

8. The Web began at CERN, the European Particle Physics Laboratory in Geneva, Switzerland,[17] and gives universal access to large numbers of documents.[18] The Web consists of documents and hyperlinks ("links").[19] Documents have content, and links[20] point to other documents.[21]

[10] *See* DERN, *supra* note 7, at 310.

[11] Usenet is the informal network of machines that have agreed to store and forward each other's news. *See* WIRED STYLE, *supra* note 1, at 111.

[12] Wais is a sophisticated universal search tool. *See* DERN, *supra* note 7, at 347-48.

[13] *Id.* at 315. Veronica is an "archie companion" to Gopher. WIRED STYLE, *supra* note 1, at 152. The name may also be an acronym for Very Easy Rodent-Oriented Net-wide Index to Computerized Archives. *See* DERN, *supra* note 7, at 315.

[14] For more information on Netscape and its Navigator, see *Welcome to Netscape* (visited Mar. 29, 1997) <http://www.netscape.com> [hereinafter *Welcome to Netscape*].

[15] For more information on Microsoft and its Internet Explorer software, see *Microsoft Home Page* (visited Mar. 29, 1997) <http://www.microsoft.com>.

[16] Although tools such as standalone ftp and archie clients remain the fastest and most direct methods when employed by those with the technical knowledge, their functionality has been streamlined and integrated into web browsers such as Netscape Navigator and Microsoft Explorer. *See* Shea v. Reno, 930 F. Supp. 916, 929 (S.D.N.Y. 1996).

[17] *See* DERN, *supra* note 7, at 323.

[18] *See id.* (quoting World Wide Web developer Tim Berners-Lee).

[19] *See Shea*, 930 F. Supp. at 929.

[20] For the purposes of this paper, the terms "hyperlink," "link," and "clickable link" will be used interchangeably.

[21] *See Shea*, 930 F. Supp. at 929. For example, a page on a machine in Germany contains a list of international legal materials with links to other relevant web sites such as those of the World Trade Organization and UNICEF. *See International Legal Materials* (visited Mar. 22, 1997)

9. Like the Gopher project, web browsers provide a unified interface to all Internet services.[22] A click of a link now accomplishes the previous multistep searching, downloading, and converting process.

10. The Web's growth pattern paralleled that of its precursor Internet services. Just as the rapid proliferation of ftp archives required the development of Archie, the proliferation of web sites resulted in analogous harvesting and indexing tools called spiders.[23]

11. There were also some key differences between the growth of the Web and that of its precursor Internet services. The organization of the Web is more recursive. In contrast to the self-contained nature of an ftp archive, web pages contain links to other web pages. Like previous Internet indexing efforts such as Archie, a spider consists of two components, one for "harvesting" and one for "querying." To build their databases, spiders must "harvest" web sites. Spiders search recursively, putting pages into their databases, then adding all the pages to which the first page points, and then endlessly adding all the subsequent pages.[24]

12. The query component is the interface between the user and the database. When the user visits the spider's web site, the query component presents the user with a form into which the user can enter her search request.[25] The query component then takes this request and searches the harvested database for matches to its criteria. For each web page matching the user's criteria, the query component of a typical spider returns a pointer to that web page, known as its Uniform Resource Locator ("URL"),[26] a brief summary of that page's contents, and a clickable link to that page. Along with the list of matches, the typical spider also returns

<http://radbruch.jura.uni-mainz.de/~baab/materials.html>; Garry Ray, *Mosaic: The Killer App*, COMPUTERWORLD, Feb. 1, 1994, at 72, *available in* 1994 WL 13683414.

[22] For example, Netscape Navigator can download Adobe Acrobat Reader for the Macintosh Version 3.0 seamlessly, replacing the suite of standalone ftp programs and file converters required previously. *See Adobe Acrobat Free Reader for the Macintosh* (visited Mar. 22, 1997) <http://www.adobe.com/prodindex/acrobat/macdnld.html>; s*ee also* Ray, *supra* note 21, at 72.

[23] *See* Amy Cortese, *Cyberspace: Crafting Software that Will Let You Build a Business Out There*, BUS. WK., Feb. 27, 1995, at 78, 82.

[24] *See The Web Robots FAQ* (visited Mar. 31, 1997) <http://info.webcrawler.com/mak/projects/robots/faq.html> [hereinafter *Web Robots FAQ*].

[25] *See, e.g., Yahoo!* (visited Mar. 31, 1997) <http://www.yahoo.com>; *AltaVista Search: Main Page* (visited Mar. 31, 1997) <http://www.altavista.digital.com>; *Webcrawler Searching* (visited Mar. 31, 1997) <http://www.webcrawler.com>.

[26] For example, the URL for the author's home page is http://acs5.bu.edu:8001/~mschmelz/index.html. This URL translates to "use the hypertext transfer protocol ('http:') at port 8001 to transfer the file 'index.html', which can be found in the home directory ('~') of user mschmelz on the machine acs5.bu.edu." *See* DAVID FLANAGAN, JAVASCRIPT: THE DEFINITIVE GUIDE 546-50 (2d ed. 1997) (describing components of JavaScript's URL object.)

advertising.

C. *Commercialization of the Web: Revenue Models for Web Enterprises*

13. Although the World Wide Web began as an academic project, it is now a significant factor in the business world.[27] By making the Web searchable and accessible, spiders have made commercialization possible.[28] If the spider operators[29] are to be successful commercial enterprises themselves, they need revenue sources. The Web currently supports many commercial ventures with several revenue models. These include subscriptions, advertising, and enabling the purchase of goods and services.

14. Under the subscription-based revenue model, only authorized users who have paid a subscription fee have access to a web site's services.[30] Some subscription-based sites first offer a small amount of free information to entice the user to subscribe to get more information.[31]

15. A refinement of the subscription model is the pay-per-view ("PPV") model, in which the user pays to see a page at the time she wants to see it.[32] Unlike the subscription model, the user does not give any of her personal information to the web site, nor does the web site set up an account for her. Although the PPV model is not yet well-developed, it will be viable, if not dominant, once transaction costs are lowered.[33]

[27] *See*, *e.g.*, *On-line Capitalism*, ECONOMIST, Nov. 23, 1996, at 92 (noting that small businesses soon may make their initial public offerings on the Web); *Watching the Web*, WALL ST. J., Aug. 29, 1996, at B10 (describing new and interesting web ventures).

[28] *See* Philip E. Ross & Nikhil Hutheesing, *Along Came the Spiders*, FORBES, Oct. 23, 1995, at 210.

[29] Spider operators include Webcrawler, *Webcrawler Search Page* (visited Aug. 28, 1996) <http://webcrawler.com/>, Lycos, *Lycos Search Page* (visited Aug. 28, 1996) <http://www.lycos.com/>, AltaVista, *AltaVista Search: Main Page* (visited Aug. 28, 1996) <http://altavista.digital.com/>, InfoSeek, *Infoseek Search Page* (visited Aug. 28, 1996) <http://www.infoseek.com/>, Hotbot, *Hotbot Search* (visited Aug. 28, 1996) <http://www.hotbot.com>, and Excite, *Excite* (visited Aug. 28, 1996) <http://www.excite.com/>.

[30] The *Wall Street Journal's* web site is available by subscription and allows users with a credit card to pay for a one year subscription online. *The Wall Street Journal Interactive Edition* (visited Feb. 9, 1997) <http://wsj.com>.

[31] *See, e.g., ESPNET SportsZone* (visited Jan. 14, 1997) <http://espnet.sportszone.com/> (offering the day's results for major league sports, but giving subscribers in-depth information) [hereinafter *ESPNET SportsZone*].

[32] For more information on PPV systems, see *Pay Per View* (last modified Oct. 19, 1995) <http://gost.isi.edu/info/ppv/>; *see also USC ISI's Information Marketplace* (last modified May 25, 1995) <http://nii.isi.edu/market/> (using a PPV system to see recipes).

[33] *See* Tom Steinert-Threlkeld, *The Buck Starts Here*, WIRED, Aug. 1996, at 133, 135. One of

16. Other sites are interfaces for mail-order houses that sell goods and services such as books,[34] compact disks,[35] and airline tickets.[36] Buyers can send credit card information to vendors over secure protocols[37] without fear of eavesdropping. These businesses can maintain a huge, centralized inventory without high overhead costs.

17. Sites that attempt to replace traditional newspaper features, such as help-wanted ads[38] and movie listings[39] use the advertising-based revenue model. A computerized query-response system better serves these functions than the printed classified advertisement section.[40]

18. Under a broader advertising-based revenue model, users do not pay for use of a web site, but receive advertising with other web site content. This model is closely analogous to commercial television.[41] Some sites require the user to complete a questionnaire before giving free access.[42] Usually this allows the company to build a mailing list or better target its advertising.

19. The web site itself may function as an advertisement[43] or provide

the most popular sites, ESPNET SportsZone, has just introduced one-day subscriptions payable through an electronic transaction. *See* Rebecca Quick, *SportsZone Readies Daily-Fee Plan, Challenging Other On-Line Services*, WALL ST. J., Mar. 17, 1997, at B9; *ESPNET SportsZone*, *supra* note 31.

[34] *See Amazon Bookstore* (visited Jan. 5, 1997) <http://www.amazon.com>.

[35] *See CDNow!* (visited Feb. 7, 1996) <http://www.cdnow.com>.

[36] *See Travelocity air travel hotel online reservations* (visited Mar. 15, 1997) <http://www.travelocity.com>.

[37] For a full discussion of how secure protocols enable commerce to take place over the Internet, see *Netscape Security Solutions* (visited Feb. 9, 1997) <http://home.netscape.com/info/security-doc.html>.

[38] *See Jobfind* (visited Feb. 9, 1997) <http://www.jobfind.com/>.

[39] *See Philly Online at the Movies* (visited Feb. 9, 1997) <http://www.phillynews.com/online/movies/>; *Boston Phoenix - Movies* (visited Feb. 9, 1997) <http://bostonphoenix.com/alt1/standard/movies/index.html>.

[40] *See* Claudia Dreifus, *The Cyber-Maxims of Esther Dyson*, N.Y. TIMES, July 7, 1996, § 6 (Magazine), at 16 (noting that stock prices, and movie and apartment listings are conducive to an electronic searchable environment, and predicting their disappearance from traditional newspaper listings).

[41] Examples include most spider operators, as well as magazine-like sites such as Hotwired, *HotWired* (visited Aug. 28, 1996) <http://www.hotwired.com/> [hereinafter *HotWired*], Addicted to Noise, *Addicted to Noise* (visited Jan. 15, 1997) <http://www.addict.com> [hereinafter *Addicted to Noise*], and Slate, *Slate* (visited Oct. 15, 1996) <http://www.slate.com>.

[42] *See HotWired*, *supra* note 41.

[43] *See*, *e.g.*, *Miller Time!* (visited Feb. 9, 1997) <http://www.millerlite.com/> (showcasing the

customer information.[44] In these cases, the Web is another advertising outlet for traditional wares, analogous to an infomercial.

20. Although these various revenue models are not mutually exclusive,[45] most Web enterprises charge nothing for use and rely on advertising for their revenue. Advertising is quickly gaining respect as a good source of revenue on the Web.[46]

D. *The Threat to Commercial Spiders*

21. Advertising-based enterprises, including spiders, want users to see and read their advertisements. Current advertising is usually a small banner at the top of the page that imposes few burdens on the user or her connection.[47] End-users who do not want to see any advertising could create “filters” that show only the hard data, or the URLs that match her query, returned by the spider.[48]

22. The practice of a third party modifying the output of a spider for a user[49] could lead to two possible threats, "spoofers" and "metaspiders." A spoofer is an intermediator that substitutes its own advertising for that of the original spider, thus fooling the user into thinking that the spoofer, and not the original spider, did the harvesting and querying.[50] Spoofing does not result in any added functionality for the end-user who would have no reason to prefer the spoofing intermediator’s advertisements to those of the original spider.[51] The spoofer could benefit, however, by placing itself where end-users expect a legitimate spider to be, as in the recent

"Dick" series of Miller Lite advertisements).

[44] *United Parcel Service Home Page* (visited Jan. 15, 1997) <http://www.ups.com> (providing rate information and allowing customers to track their packages).

[45] *See, e.g.*, *Addicted to Noise*, *supra* note 41 (carrying advertisements and allowing users to directly purchase compact disks).

[46] Catharine P. Taylor, *Banner Year*, WIRED, Mar. 1997, at 120, 120.

[47] *Cf.* Taylor, *supra* note 46, at 122 (noting that advertisements are becoming more burdensome as they incorporate elements such as sound and animation).

[48] *See* FLANAGAN, *supra* note 26, at 227 (demonstrating JavaScript code for listing all links in a document.)

[49] Third party modification of spider output for a user will be called "intermediating" for the purposes of this paper.

[50] *See* DERN, *supra* note 7, at 378. Spoofing closely maps the concept of passing-off.

[51] The only evidence that the "legitimate" spiders are not engaged in this practice themselves is their huge disparity in search results. A search for the author’s name returned disparate results. *Compare Lycos* (visited Mar. 31, 1997) <http://www.lycos.com/> (returning two pages), *with HotBot -- Results* (visited Mar. 31, 1997) <http://www.hotbot.com/> (returning 547 pages).

case involving AltaVista.[52]

23. Metaspiders are the more insidious variety of intermediator. Metaspiders act on the principle that if querying one spider is good, querying many spiders is better. A metaspider will pass a user's query to several different spiders that search their databases simultaneously. No two spiders' databases are alike and any spider may have matches that the others do not.

24. The metaspider receives the spiders' reports and returns the results to the end-user. Metaspiders will not threaten individual spider operators if the metaspider returns results intact because the user sees the same results as if she had asked each spider individually. This approach might be beneficial to the spider operators as it would arguably increase exposure to end-users. The end-users would enjoy the convenience and thoroughness of one-stop shopping, resulting in an expanded audience for the spider's sponsors.

25. Given the likelihood of extensive duplication of results, however, the metaspider's first step would be to eliminate duplicate hits. The metaspider could compile the unique matches and return them on one page instead of forcing the user to click between individual spiders' results. Metacrawler implements this system noncommercially.[53]

III. PROTECTIONS OFFERED BY COPYRIGHT LAW

26. Given a credible threat to spider operators' main source of revenue, the operator may seek protections under copyright law. The first section of this Part will examine the copyrightability of the results returned by a spider in response to a user's query. The second section will examine theories of infringement that a spider could pursue against a metaspider.

A. *Copyrightability of Search Results*

27. To qualify for copyright protection, the search result must be an original work of authorship fixed in a tangible medium of expression.[54] Copyright vests upon

[52] David D. Kirkpatrick, *Tale of Two AltaVista Web Sites Teaches Useful Marketing Lesson*, WALL ST. J., Oct. 18, 1996, at B18 (reporting confusion for users looking for Digital's AltaVista search engine and finding Alta Vista Software's home page by mistake). Spoofing in general, and the AltaVista case in particular, will be covered *infra* in Part V.

[53] *Metacrawler Search* (visited Aug. 28, 1996) <http://metacrawler.cs.washington.edu/> (returning a list of sites, which search engines returned the sites, and each search engine's thumbnail sketch) [hereinafter *Metacrawler Search*].

[54] *See* 17 U.S.C. § 102(a) (1994) (defining when copyright vests in the author); § 101 (defining fixation).

fixation.[55] The standard for fixation is "embodiment in a copy which is sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration."[56] Although they appear ephemeral, a spider's search results meet this requirement.[57]

28. A spider's results may not meet the authorship requirement for a valid copyright. Inherent in "authorship" is minimal creativity.[58] If the metaspider is only passing along "facts," then the spider has no recourse to copyright law.[59]

29. While some metaspiders merely collect and collate URLs,[60] others return the complete pages returned by the spiders.[61] Ironically, this second type of metaspider is less of a threat to the original spider although the data returned demonstrates authorship and therefore copyrightability. This Part will discuss the various components of a typical spider's results, the copyrightability of each component, and the applicability of compilation copyright to the results as a whole.

1. Individual Components

30. The format of search results returned by a spider varies from spider to spider. All spiders include a page's title, while others also return URLs and summaries of page content.[62]

a. Titles and Uniform Resource Locators

31. Although the titles and URLs are the most important information conveyed by the spider, copyright law is unlikely to protect them because they most closely resemble facts.

[55] *See* §§ 101, 102(a).

[56] § 101.

[57] *Cf.* MAI Sys. Corp. v. Peak Computer Corp., 991 F.2d 511, 518 (1992) (holding that a program stored in a computer's RAM is a copy for purposes of the Copyright Act); *see also* Triad Sys. Corp. v. Southeastern Express Co., 64 F.3d 1330, 1335 (agreeing with the MAI court's use of RAM stored programs as copies).

[58] *See* Feist Publications, Inc. v. Rural Tel. Serv., 499 U.S. 340, 345-52 (1991).

[59] *See* 17 U.S.C. § 102(b) (copyright protection does not extend to facts or ideas); *Feist*, 499 U.S. at 358.

[60] *See, e.g.*, *Metacrawler Search*, *supra* note 53.

[61] *See, e.g.*, *All-In-One Search Page* (visited Aug. 30, 1996) <http://www.albany.net/allinone/>.

[62] *Compare Webcrawler Search* (visited Feb. 9, 1997) <http://www.webcrawler.com> (returning only a title for each match), *with AltaVista Search* (visited Feb. 9, 1997) <http://altavista.digital.com> (returning a title, URL, and a brief summary for each match), *and Excite Search* (visited Feb. 9, 1997) <http://www.excite.com> (returning a title, URL, summary, and a link to other similar sites).

32. When an end-user submits a query to a spider, the spider looks through its database for matches to that query. The spider built its database through “harvesting” information by automatically and continuously searching the Web, gathering titles and URLs.[63] The harvested information is unlikely to be considered a work of authorship because the database is a set of information gathered by the “sweat” of the harvesting program’s “brow.”[64] The Supreme Court unanimously rejected conferring copyright solely based on the sweat of the brow.[65]

33. It is highly unlikely that page titles and URLs would meet the authorship requirement. Even if they did, the creators of the web pages are the authors, not the gathering spiders. The spiders would not have standing to sue because they only catalogue the titles and URLs, but they do not create them.[66]

b. Thumbnail Sketches

34. Titles and URLs retrieved from a spider's database illustrate the idea/expression dichotomy.[67] The summaries prepared by some spiders raise the more interesting problem of authorship.[68]

35. Authorship requires the independent creation of a work with a “minimal degree” of creativity.[69] For each web page harvested, the spider submits the entire contents of the page to a "summarizer" program whose job it is to reduce the contents to a thumbnail sketch.[70] When a user’s query later “hits” the page, the

[63] User submissions of URLs supplement the automatic harvesting done by robots.

[64] “[T]he 1976 revisions to the Copyright Act leave no doubt that originality, not ‘sweat of the brow’ is the touchstone of copyright protection in directories and other fact-based works.” *Feist*, 499 U.S. at 359-60.

[65] *See id.* (rejecting the proposition that the labor put into an alphabetical ordering of white pages telephone listings suffices to confer authorship, and hence copyright protection, on the compilation). The copyrightability of the spider's database is not an important issue here, however, because the metaspider does not copy or misappropriate the database as a whole. In addition, at least one spider's operating company considers trade secret law to provide adequate protections for its database. Interview with Gary Levine, General Counsel of Lycos, in Newton Center, Mass. (Aug. 21, 1996) [hereinafter Levine Interview]. Lycos has licensed its database query engine and the database itself to third parties. *Id.*

[66] *See* 17 U.S.C. § 501(b) (1994) (giving only the legal or beneficial owner of a section 106 exclusive right standing to sue infringers).

[67] *See* § 102.

[68] For the purposes of this paper, a “thumbnail sketch” is a summary designed to give the end-user a quick impression of a web page’s contents. Some spiders do not prepare a summary. *See supra* note 62 and accompanying text.

[69] *Feist*, 499 U.S. at 345-52.

[70] *Compare*, *e.g.*, Charles Ulrich, *Bullwinkle's Corner* (visited Feb. 9, 1997) <http://mindlink.bc.ca/Charles_Ulrich/bc.html> (listing all Rocky and Bullwinkle cartoon titles), *with Lycos search: bullwinkle* (visited Feb. 9, 1997) <http://www.lycos.com/cgi-

spider returns the page's thumbnail sketch with the page's URL, allowing the user to judge for herself how well the page matches her query.[71] At least one metaspider passes these thumbnail sketches to the end-user.[72]

36. For the moment, let us defer the issue of nonhuman computer authorship and examine the nature of the thumbnail sketch itself. For this discussion, assume that a human produced the thumbnail sketch according to strict guidelines set out by the spider's operating company.[73]

37. The thumbnail sketch could be considered a joint work between the web page's original author and the summarizer.[74] Each author would own an undivided interest in the thumbnail sketch.[75] If the thumbnail sketch is a work of joint authorship, then the spider could pursue any infringer of that work independent of any of its coauthors, while enjoying immunity from being sued by the other coauthors.[76] Any of the coauthors can use the thumbnail sketch, however, without requiring the spider's permission.[77] For example, issues of enforceability aside, the original author of a web page could put a notice on the page stipulating that any thumbnail sketches be freely distributable.[78]

38. Similar results would follow if the thumbnail sketch was considered a derivative work of the original page.[79] The summarizer would be the sole author of

bin/pursuit?cat=lycos&query=bullwinkle> (illustrating Lycos' reduction of the Bullwinkle's Corner web page to a thumbnail sketch).

[71] For example, if a user wanted information about Rocky and Bullwinkle cartoons, a brief look at the returned thumbnail sketches would allow her to eliminate pages belonging to people who used Rocky as their login names as well as those pages dedicated to the Sylvester Stallone boxing movies.

[72] *See Metacrawler Search*, *supra* note 53.

[73] For simplicity, this human will also be called a "summarizer."

[74] A joint work is a "work prepared by two or more authors with the intention that their contributions be merged into inseparable or interdependent parts of a unitary whole." 17 U.S.C. § 101 (1994).

[75] *See* § 201(a); *see also* Community for Creative Non-violence v. Reid, 846 F.2d 1485, 1498 (D.C. Cir. 1988) (joint authors are treated as tenants in common).

[76] *See Community for Creative Non-violence*, 846 F.2d at 1498.

[77] *See id.*

[78] This notice would not have to be visible to a human reading the page on a browser. One way this provision could be implemented is through the definition and acceptance of a standard HTML "tag" incorporated into the document. An alternate approach might involve the definition and acceptance of a standard entry into the site's "robots.txt" file. These concepts will be discussed in more detail in Part IV.B.1, *infra*.

[79] *See* 17 U.S.C. § 101 (defining a derivative work as a "work based on pre-existing works").

its contribution,[80] and would not need to confer with the web page's author before pursuing an infringement remedy for the thumbnail sketch.[81] The spider might not possess a valid copyright in the thumbnail in the first place. If the web page author shows that the spider's thumbnail was an unauthorized derivative work, then the spider cannot pursue a copyright remedy for any subsequent infringement of the thumbnail.[82] The web page author could put the summarizer on notice by including on the original page a disclaimer prohibiting thumbnail sketches.

39. In the unlikely event that a web page author sued a spider, however, the spider could raise the affirmative defense of fair use.[83] Fair use allows otherwise infringing copying when it furthers the progress of science and the useful arts.[84] The fair use analysis includes four non-exclusive factors.[85]

40. The first factor of fair use analysis is "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes."[86] This factor would weigh against the commercial spider operator, but is not dispositive.[87]

41. The second factor is "the nature of the copyrighted work."[88] This varies greatly with each web page. Predicting a typical case is impossible because the summarizer treats all pages equally, whatever their nature. Consequently, the spider's operating company could argue that this factor is insignificant in this case.

42. The third factor is "the amount and substantiality of the portion used in relation to the copyrighted work as a whole."[89] The purpose of the thumbnail sketch

[80] *See* § 103(b) (granting the author of a derivative work sole copyright in the author's contribution only, and not extending or modifying the copyright in the underlying work).

[81] However, if the spider operator's suit included verbatim copying from the original web page, then the web page author would have to join the suit. *See id.*

[82] *See* § 106 (giving copyright owners exclusive rights to create derivative works); § 103(a) (witholding copyright protection in a derivative work when use of the underlying work was unlawful); § 501(b) (giving copyright owners rights to sue infringers). *Cf.* Wainwright Secs., Inc. v. Wall St. Transcript Corp., 558 F.2d 91, 96 (2d Cir. 1977) (holding almost verbatim abstracts of industry reports to be infringing and denying a fair use exception).

[83] *See* § 107.

[84] *See id.*

[85] *See id.*

[86] § 107(1).

[87] *See*, *e.g.*, Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 583-84 (1994)(commercial nature of use is not dispositive, but merely one of the factors that must be balanced).

[88] 17 U.S.C. § 107(2).

[89] § 107(3).

is to produce a summary of the page's contents, therefore, the "amount" part of this factor should weigh in the spider's favor.[90] Since the summarizer wants to produce an accurate thumbnail sketch that properly conveys the sense of the page, the summarizer may be copying "the heart" of the original work, causing the "substantiality" part of this factor to weigh against the spider.[91]

43. The fourth and final aspect of fair use analysis, "the effect of the use upon the potential market for or value of the copyrighted work,"[92] is arguably the most important,[93] and the only factor that weighs clearly in the spider's favor.[94] The spider can argue that the thumbnail sketch does not harm the market for the original work, but in fact provides free publicity, thus resulting in a positive effect on its market value.

44. Such suits would only be useful as nuisance suits, or perhaps as part of a concerted effort to shut down the spiders. It is difficult to take the threat seriously because spiders are seen as the keystones to the Web’s commercial success.[95] The possibility is important to recognize, however, because the threat of such a suit might be useful as a strategic move by the metaspiders. The metaspiders would admit they summarize the spiders if the spiders admit they summarize original web pages.

45. Even if the spiders may create thumbnail sketches as a fair use, the creators may not be authors because they are not human. Humans cannot summarize the millions of existing of web pages. These spiders do not employ

[90] *See* Salinger v. Random House, Inc. 811 F.2d 90, 98 (1987) (close paraphrasing of parts of author’s letters infringing); Norse v. Henry Holt & Co., 847 F. Supp. 142, 146-47 (N.D. Cal. 1994) (copying of 50 of 12,000 words did not copy the heart of the work and was not infringing). *Cf. Acuff-Rose*, 510 U.S. at 580-81 (finding parodies can copy just enough for audience to identify original song).

[91] *See* Harper & Row Publishers v. Nation Enters., 471 U.S. 539, 564-66 (1985) (holding that approximately 300 words from President Ford’s unpublished memoir was the heart of the work and therefore substantial copying); H.C. Wainwright & Co. v. Wall St. Transcript Corp., 418 F. Supp. 620, 625, *aff’d on other grounds*, 558 F.2d 91 (2d Cir. 1977) (“The takings have been substantial in quality, and absolutely, if not relatively substantial in quantity.”).

[92] 17 U.S.C. § 107(4).

[93] *See* Stewart v. Abend, 495 U.S. 207, 238 (1990); *Harper & Row*, 471 U.S. at 566. *But see Acuff-Rose*, 510 U.S. at 581-85 (all factors must be weighed in light of the purposes of copyright law).

[94] *See* MELVILLE B. NIMMER & DAVID NIMMER, 3 NIMMER ON COPYRIGHT, § 1305[A][4], at 13-183 to 13-184 (Release 41 1996) [hereinafter NIMMER ON COPYRIGHT] (noting that the fourth factor should cover not just the specific use in the case, but all similar uses by other individuals). *Cf.* American Geophysical Union v. Texaco, Inc., 37 F.3d 881, 895 (2d. Cir. 1994) (noting that determining market value is difficult because there is no traditional market for individual journal articles); Twin Peaks Prods., Inc. v. Publications Int’l, Ltd., 996 F.2d 1366, 1377 (2d Cir. 1993) (noting commercial use fair when it fills a market plaintiff would not enter otherwise).

[95] *See* Verity, *supra* note 2, at 88.

human summarizers but rely on automated programs to generate millions of web page summaries.[96] Although the thumbnail sketches may be works of authorship, the author is a computer. The law is unclear as to whether computer generated works satisfy the authorship requirement of the Copyright Act.[97]

46. Although commentators have generally accepted the possibility of computer generated authorship, the articles have been speculative, concentrating on computer authorship of de novo, rather than derivative, works.[98] There has not been any discussion of a program such as a summarizer, that consistently applies a specific, sophisticated set of rules to every work fed into it.

47. The Copyright Act does not explicitly require an author to be human.[99] The issue is undecided,[100] but the Final Report of the National Commission on New Technological Uses of Copyrighted Works suggests that the Copyright Act should recognize such nonhuman authorship.[101]

48. Lycos considers its summarizer to be highly sophisticated, and maintains the technology behind it as a closely guarded secret.[102] Lycos views their summaries

[96] AltaVista has summarized more than 30 million pages. *AltaVista Search: Main Page* (visited Mar. 31, 1997) <http://altavista.digital.com/>. Excite has summarized more than 50 million pages. *The New Excite Search* (visited Mar. 31, 1997) <http://www.excite.com/ice/new.html?1#size>. Lycos has summarized more than 51 million pages. *Lycos Catalogs 51 Million URLs* (visited Mar. 31, 1997) <http://www.lycos.com/press/51million.html>.

[97] 17 U.S.C. §§ 101-1101 (1994) (humanity never mentioned as requirement for authorship); *see also* 1 NIMMER ON COPYRIGHT, *supra* note 94, § 5.01[A], at 5-5 to 5-6 (stating that "the time may not be far off when that question demands an answer").

[98] For example, commentators have discussed the nonsensical yet spookily human-seeming book originally composed by the computer program Racter. WILLIAM CHAMBERLIN & THOMAS ETTER, THE POLICEMAN'S BEARD IS HALF CONSTRUCTED, COMPUTER PROSE AND POETRY BY RACTER (1984). *But see* Evan H. Farr, *Copyrightability of Computer-Created Works*, 15 RUTGERS COMPUTER & TECH. L.J. 63, 79 (giving computer authorship status "is absurd"). For a scientific explanation of the concepts behind Racter and similar efforts, see DOUGLAS R. HOFSTADTER, FLUID CONCEPTS AND CREATIVE ANALOGIES 158, 471, 480-81 (1995).

[99] *See* 17 U.S.C. §§ 101-1101 (1994). *But see* § 104 (conferring protection to unpublished works regardless of the author's citizenship, but proper citizenship of the author must be taken into account when determining ownership of the copyright of a published work).

[100] *See* 1 NIMMER ON COPYRIGHT, *supra* note 94, § 5.01[A], at 5-5; Arthur R. Miller, *Copyright Protection for Computer Programs, Databases, and Computer-Generated Works: Is There Anything New Since CONTU?*, 106 HARV. L. REV. 977, 980 (1993).

[101] NATIONAL COMMISSION ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS, FINAL REPORT ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS 43-46 (1979); s*ee also* OFFICE OF TECHNOLOGY ASSESSMENT, INTELLECTUAL PROPERTY RIGHTS IN AN AGE OF ELECTRONICS AND INFORMATION 70-73 (1986) (suggesting that computer programs are not merely "inert tools of creation," and could be "co-creators" of works with human authors).

[102] Levine Interview, *supra* note 65.

as copyrightable works of authorship that provide added value to the end-user.[103] This belief may be more important in determining the issue than any philosophical polemics: if the spider operators believe their summaries are copyrightable and a valuable asset, they will protect them. If intermediaries perceive a threat of viable legal action, they may choose not to risk an infringement suit to decide the issue.

49. An analogy can be drawn between spider operators and the performing rights societies that seek to regulate the playing of sound recordings and underlying musical works on the Web. The process of playing the music clip on the end-user's computer involves copying the clip onto the end-user's computer and playing it through the end-user's computer's speaker.[104] This process is closer to the copyright holder's "reproduction right" than the "performance right,"[105] because the computer makes a perfect copy of the song, the song is played for the end-user's benefit, and the end-user is probably not in public.[106] Nonetheless, the two major performance rights clearinghouses, the American Society of Composers and Performers ("ASCAP") and Broadcast Music Incorporated ("BMI"), believe that this practice implicates the performance right, and have drawn up agreements to this effect.[107] They must argue that the copying implicates the performance right and not the reproduction right because antitrust decrees restrict them from entering the field of blanket licensing for reproduction rights.[108]

50. Just as ASCAP and BMI can argue for performance rights despite the underlying physical and legal situation, so can the spiders argue for the copyrightability of their summaries.

[103] *Id.*

[104] This is usually done through a “plug-in” to the computer’s web browser. *See, e.g., Real Audio Home Page* (visited Mar. 31, 1997) <http://www.realaudio.com> (allowing users to download free software to hear sound over the Internet).

[105] 17 U.S.C. § 106 (1994). The reproduction right involves the making of physical copies such as phonorecords and sheet music. *See id.* The performance right involves the public performance of the copyrighted work, including live performance and radio broadcast. *See id.*

[106] *See* § 101 (defining public performance).

[107] *See, e.g.,* iRock--Internet Music Radio (visited Mar. 31, 1997) <http://www.irock.com/main.html> (billing itself as the first Internet site playing music with ASCAP/BMI permission); *see also* Mark F. Radcliffe, *Multimedia in the Digital World*, *in* MULTIMEDIA 1997: PROTECTING YOUR CLIENT’S LEGAL AND BUSINESS INTERESTS, at 9, 65-68 (PLI Patents, Copyrights, Trademarks, and Literary Property Course Handbook Series No. G4-4000) (noting that the Copyright Clearance Center is creating agreements to license digital rights and recommending that ASCAP and BMI create standard licenses for digital media).

[108] *See* Broadcast Music, Inc. v. CBS, Inc., 441 U.S. 1, 10-12 (1979) (describing ASCAP and BMI consent decrees).

c. Advertising

51. The final element that a spider returns in response to a user's query--the advertising accompanying the search results--is the element most clearly protected by copyright. Yet, ironically, it is also the one element that the spiders would least object to an intermediator copying. The spider could conceivably use a metaspider's copying of its advertising as the basis for an infringement suit, but that would not make sense strategically: the metaspider's next move would be to strip the spider's advertising, leaving the spider in a worse commercial position.[109]

52. In sum, the titles and URLs of web pages that match a user's query are probably uncopyrightable facts. Courts will probably regard the summaries produced by individual spiders as copyrightable works of authorship, despite issues of nonhuman authorship and the possibility that the summaries might be infringing, derivative works. The advertising accompanying a spider's search results are unquestionably copyrightable works of authorship, and a spider suing for copyright infringement would be on solid ground.

53. Thus, the safest way for any potential intermediator to avoid copyright liability may be to forward only the titles and URLs while discarding any advertising or summaries. The metaspider is free, of course, to visit all the matching web pages itself and generate its own thumbnail sketch with its own summarizer.[110]

2. Compilation Copyright

54. Compilation copyright can attach to the set of matches produced by the spider, but it would not prove useful against intermediators. Although a subjective selection of facts is sufficient authorship to confer copyright on a compilation of those facts,[111] the compilation copyright does not carry over to the underlying facts.

55. For example, assume that titles and URLs are considered uncopyrightable facts, and a spider did not attempt to commingle them with any copyrightable material such as thumbnail sketches, but simply returned a list of URLs. Note first that the order of this list is important. If the items were

[109] Such a move by the metaspider could, however, strengthen the spider's case for misappropriation or unfair competition, as will be discussed *infra* section IV.

[110] iFind! returns clickable URLs, "clustered" by their similarity to each other. It is not clear if iFind! relies on spiders' summaries for making the clustering calculations or if iFind! makes them from the original web pages. *See Inference Find!* (visited Mar. 31, 1997) <http://m5.inference.com/ifind/>.

[111] *See* 17 U.S.C. § 101 (defining compilation copyright); *see also* Key Publications, Inc. v. Chinatown Today Publ'g Enters., Inc., 945 F.2d 509, 514 (2d Cir. 1991) (holding that nomenclature of business classifications in yellow pages contains sufficient authorship); Eckes v. Card Prices Update, 736 F.2d 859, 862-63 (2d Cir. 1984) (finding sufficient authorship in selecting "premium" and "common" baseball cards).

presented in alphabetical order, there would most likely not be enough authorship for a compilation copyright to attach.[112] Assume instead that the spider orders the list by a subjective criterion, such as match score,[113] and that this is sufficient authorship to confer compilation copyrightability.

56. In this situation, the only protection offered by compilation copyright is against the intermediator presenting exactly the same list in the same order, with the spider-plaintiff bearing the burden of proving that this was the result of the intermediator's copying. In other words, compilation copyright offers only a thin protection that is difficult to prove.

57. Also, the authorship of the compilation must be considered in the copyright analysis. The end-user is sole author of the query. The subjective selection of the compilation stems directly from that query. The end-user, therefore, is at the very least a coauthor of any compilation copyright. As coauthor, the end-user would not have to confer with any other coauthors to exercise her exclusive rights as she sees fit.[114] Thus, the metaspider, as agent of the coauthors, should not have to worry about any liability deriving from compilation copyright.[115]

B. *Different Types of Copyright Infringement*

58. To win a copyright infringement suit, a plaintiff must prove that the plaintiff holds a valid copyright, and that the defendant violated one or more of the plaintiff's exclusive rights by copying the plaintiff's protected expression.[116] In addition to direct copyright infringement, the spider-plaintiff may also pursue contributory and vicarious copyright infringement claims.[117]

59. If an intermediator makes an unauthorized copy of a spider's protected expression and passes that expression verbatim to the end-user, then the intermediator is liable for direct copyright infringement.[118] If the intermediator does

[112] *See* Feist Publications, Inc. v. Rural Tel. Serv., 499 U.S. 340, 363 (1991).

[113] Match scores rank pages based on how well those pages match the user’s query.

[114] *See* 17 U.S.C. §§ 106, 201.

[115] The selection of advertising provided by the spider could be a result of joint authorship. At least one spider, Lycos, selects the advertisement based on the user's query. *See* Levine Interview, *supra* note 65. For example, if the user's query mentions "router boxes," Lycos would not pick an advertisement at random, but rather it would select one from a manufacturer of router boxes. *See id.*

[116] *See Feist*, 499 U.S. at 361.

[117] For this subpart, assume that the spider holds a valid copyright in the material being copied by the intermediator.

[118] *See* 17 U.S.C. §§ 106(1), 501.

not pass the expression directly to the user but makes modifications first, the intermediator still may be a direct infringer.[119] The amount and nature of modification by the intermediator will determine its liability.[120]

60. If a potential infringer could escape liability by making minute and insubstantial changes to a work of authorship, the protection offered by copyright law would be worthless.[121] On the other hand, if copyright protection extends too broadly, it stifles creativity and runs counter to the constitutional mandate behind copyright law, to "promot[e] the sciences and the useful arts."[122]

61. The intermediator has a right to copy the spider's facts if the intermediator expresses those facts originally. To prove infringement, the plaintiff must have a valid copyright in the work and must show that the defendant unlawfully copied the work. If the spider shows actual copying, the next step is to decide if the intermediator's copying was unlawful because it copied more than unprotected facts and thereby infringed the spider's copyright. The copying need not be literal; infringing copying can be found if the two works are substantially similar.[123]

62. Although a spider can sue an intermediator for damages based on specific cases of infringing copying, the spider may shut down the intermediator permanently via injunctive relief by pursuing a case of contributory infringement. Contributory infringement requires the plaintiff to prove that the intermediator has no substantially noninfringing use.[124]

63. In *Sony Corp. of America v. Universal City Studios*, the motion-picture studios pursued a contributory infringement case against Sony to enjoin the introduction of the VCR into the United States.[125] The studios argued the VCR was nothing more than a tool for copyright infringement.[126] The Court allowed VCRs to be marketed in the United States by decreeing that the practice of private, non-commercial "time-shifting"--recording a program and watching it later--was not

[119] *See* § 106(2).

[120] *See* § 504(c) (willful infringement carries a higher penalty); § 506 (willful and for purposes of commercial advantage carries criminal penalties).

[121] *See* Nichols v. Universal Pictures Co., 45 F.2d 119, 121 (2d Cir. 1930) (infringing copying is not limited to literal, verbatim copying or "else a plagiarist would escape by immaterial variations").

[122] U.S. CONST. art. I, § 8, cl. 8.

[123] For a discussion of the ways courts have defined and used "substantially similar", see 3 NIMMER ON COPYRIGHT, *supra* note 94, § 13.03 [A], at 13-28 to 13-58.

[124] *See* Sony Corp. of Am. v. Universal City Studios, 464 U.S. 417, 442 (1984).

[125] *Id.* at 420.

[126] *See id.*

copyright infringement.[127] Thus, the Court reasoned, VCRs were acceptable because they had a "substantial noninfringing use."[128]

64. Spiders could sue intermediators on the theory of contributory infringement. To be successful, spiders would have to prove that the intermediators have no substantially noninfringing use. For the intermediators to prevail, courts must declare that the practice of intermediating is like time-shifting, and although it meets all the previously understood elements of copyright infringement, it does not, in fact, infringe.

65. A spider may also pursue a vicarious infringement case. In the standard case of vicarious infringement, the plaintiff sues a principal whose agent is doing the actual infringing.[129] Here, the spider would sue the end-user for vicarious infringement. To succeed on a vicarious infringement claim, the spider must prove that the intermediators induced or promoted the infringing acts.[130] Such a course of action would be unwieldy and unproductive, but it might be a useful legal threat in a publicity campaign designed to scare end-users away from meta-spiders.

IV. Licensing as an Alternative to Copyright Protection

A. *Software Licensing*

66. If copyright law fails to sufficiently protect the spider, then the spider can try licensing its output. To do this, the spider would have to make enforceable agreements with its users to prevent otherwise permissible copying. This Part examines the analogy between a spider's results and consumer software, how copyright law has fallen short in providing adequate protection, and how solutions used for the latter can be adopted for the former. This Part further discusses the prerequisites to making an enforceable agreement, and the possibility that the validity of such agreements may be preempted by federal copyright law.

67. Standard copyright law protections apply to computer software.[131] For example, if an employee makes a copy of Microsoft Word at work for use on her home

[127] *See id.* at 455.

[128] *Id.* at 442.

[129] *See* Gershwin Publ'g Corp. v. Columbia Artists Mgmt., 443 F.2d 1159, 1161-62 (2d. Cir. 1971).

[130] See *id.* at 1162.

[131] *See* Apple Computer, Inc. v. Franklin Computer Corp., 714 F.2d 1240, 1246-55 (3d Cir. 1983) (discussing copyrightability of various programs); 17 U.S.C. § 101 (1994) (defining computer program). *But see* § 117 (copying permitted for proper functioning of the program and for archival uses).

computer, she is infringing Microsoft's exclusive right of copying.[132] If someone lends a friend her licensed copy of Microsoft Word with the knowledge that her friend will make an illegal copy, that person could be found guilty of contributory or vicarious infringement.[133] Due to the ease of copying and the concomitant belief that it would lead to wholesale copyright infringement, there is a statutory provision against renting software.[134]

68. Copyright protections regarding derivative works also apply.[135] If a computer software developer incorporates someone else's object code verbatim into her own product without the consent of the original developer, then that product is an infringing derivative work.[136] Fair use allows the same engineer to examine that same object code, discern the algorithm it is performing, and duplicate its functionality,[137] so long as the engineer commits no other misdeeds in the process.[138]

69. Though copyright law currently contains narrow exceptions that allow decompilation for reverse engineering, commentators recognize that software developers have a legitimate interest in preventing their object code from being closely examined.[139] In addition, software developers wish to bypass limitations imposed by the first sale doctrine,[140] although specific legislation addressed their

[132] *See* § 106.

[133] *See Sony Corp.*, 464 U.S. at 434-42; *Gershwin*, 443 F.3d at 1162.

[134] 17 U.S.C. § 109(b)(1)(A).

[135] *See* § 106(2).

[136] *See* Computer Assoc. Int'l, Inc. v. Altai, Inc., 775 F. Supp. 544, 560 (E.D.N.Y. 1991), *aff'd*, 985 F.2d 693 (1992).

[137] *See* Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1520 (9th Cir. 1992) ("Where there is good reason for studying or examining the unprotected aspects of a copyrighted computer program, disassembly for purposes of such study or examination constitutes fair use.").

[138] *See* Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 844-46 (Fed. Cir. 1992) (holding that Atari had likely committed misdeeds and infringed Nintendo's copyright in its efforts to reverse engineer Nintendo's lockout system). In a related area, the Semiconductor Chip Protection Act sets out a specific fair use exemption for copies made while understanding the operations of an otherwise protected integrated circuit. 17 U.S.C. § 906 (1994).

[139] *See* Maureen A. O'Rourke, *Drawing the Boundary Between Copyright and Contract: Copyright Preemption of Software License Terms*, 45 DUKE L.J. 479, 509 (1995); *see also* Miller, *supra* note 100, at 1026-27 (arguing against allowing decompilation).

[140] *See* 17 U.S.C. § 109(a) (allowing "the owner of a particular copy . . . to sell or otherwise dispose of possession of that copy . . . without the authority of the copyright owner"); *see also* David A. Rice, *Licensing the Use of Computer Program Copies and the Copyright Act First Sale Doctrine*, 30 JURIMETRICS J. 157, 162 (1990) (discussing how the first sale doctrine does not protect the interests of software developers).

major concerns.[141] This is where software licensing enters.[142]

70. Software licensing began before the widespread availability of consumer software, when it was feasible to conduct full negotiations to make licensing agreements between informed and sophisticated parties.[143] Software companies developed shrinkwrap licenses to get the benefits of software licensing without incurring the cost of individual negotiations with many consumers.[144] For example, a purchaser of a copy of Microsoft Word buys only a license to use the product, not a copy of the product itself.[145] Microsoft affixes the license to the envelope containing the disks.[146] By opening the envelope and breaking the seal, the purchaser accepts the terms of the license.[147] These terms include a prohibition on disassembly and decompilation of the program's object code.[148]

71. Courts and commentators have debated the validity of shrinkwrap licenses.[149] In the most recent case, ProCD, Inc. v. Zeidenberg, Judge Easterbrook upheld the validity of a shrinkwrap agreement designed to protect the paradigmatic

[141] Computer Software Rental Amendments Act of 1990, Pub. L. 101-650, § 802(2)(b)(a)(A), 104 Stat. 5134, 5134 (1990) (codified as amended at 17 U.S.C. § 109(b) (1994)) (forbidding the rental, lease, or lending of a computer program).

[142] *See* Rice, *supra* note 140, 176-78 (1990).

[143] As an acknowledgment of the maturity of the field of software licensing, the National Conference of Commissioners of Uniform State Laws is developing an entire article of the U.C.C. governing software licensing. *See ULC at Chicago-Kent: UCC Article 2B* (visited Feb. 10, 1997) <http://www.kentlaw.edu/ulc/uniform/uccart2/ucc2b296.html>.

[144] *See* O'Rourke, *supra* note 139, at 495.

[145] Shrinkwrap license on file with *The Boston University Journal of Science & Technology Law* [hereinafter Shrinkwrap License]; *see also* Rice, *supra* note 140, at 157 ("Software companies purport to license the use of, rather than to sell, computer program copies.").

[146] Shrinkwrap License, *supra* note 145.

[147] *Id.*

[148] *Id.*

[149] *Compare* Vault Corp. v. Quaid Software Ltd. 847 F.2d 255, 268-70 (5th Cir. 1988) (striking down a state law clarifying the enforceable provisions of shrinkwrap licenses as preempted), *with* ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1452-53 (7th Cir. 1996) (upholding shrinkwrap license on compact disk telephone directory under common law contract). Shrinkwrap licenses have generated extensive academic debate. *See generally* Mark A Lemley, *Intellectual Property and Shrinkwrap Licenses*, 68 S. CAL. L. REV. 1239 (1995) (criticizing the proposed Uniform Commercial Code Article 2B); O'Rourke, *supra* note 139 (arguing for upholding shrinkwraps in certain circumstances); Robert W. Gomulkiewicz & Mary L. Williamson, *A Brief Defense of Mass Market Software License Agreements*, 22 RUTGERS COMPUTER & TECH. L.J. 335 (1996) (arguing for upholding shrinkwrap licenses).

example of unprotectable facts, a telephone directory.[150] It follows that if a manufacturer can protect a telephone directory through a shrinkwrap agreement, then developers can probably devise a method to protect a spider's results by an analogous method.

B. *Making an Enforceable License to Protect Spiders' Output*

72. Spider operators must meet all standard requirements of contract law to make an enforceable agreement. This Part will focus on enforceability of the terms and the mechanism of acceptance.

73. For the purposes of this Part, assume that the spider wishes to enforce the following license provisions: (1) in consideration of the spider's services, the user[151] agrees to use the spider's results only for his or her personal use;[152] (2) the user agrees not to forward the results, in whole or in part, modified or unmodified, to anyone else.

74. These terms require users to give up their statutory rights to copy facts,[153] and thus meta-spiders can argue that the terms are unenforceable due to preemption[154] or unconscionability.[155] *ProCD* recently upheld the enforceability of exactly this sort of provision.[156] Even if these provisions are fully enforceable, there remains the question of what constitutes acceptance.

75. The practice of shrinkwrap licensing has evolved to the point where the user takes more positive action than merely opening an envelope to agree to terms. For example, Netscape distributed Navigator 2.01 for the Macintosh over the Internet via a downloadable installer program.[157] As part of the installation process, when Navigator is first launched, it displays a license agreement and asks the user to read it carefully before continuing. The user signifies agreement to the

[150] *ProCD,* 86 F.3d at 1452-53.

[151] Assume that "user" is defined in the license as anyone connecting to the spider, whether end-user or intermediator, who wishes to avail themselves of the spider's services.

[152] This term would probably be limited by section 117 of the Copyright Act which allows copying necessary for proper functioning of the program and archives. 17 U.S.C. § 117 (1994).

[153] *See* § 102(b).

[154] *See* § 301.

[155] *See* O'Rourke, *supra* note 139, at 529-32.

[156] ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1455 (1996) (two-party contract not preempted by the Copyright Act).

[157] The installer program is available at *Welcome to Netscape*, *supra* note 14.

terms set forth by clicking on a button labeled "Accept,"[158] and the setup process would continue.

76. Suppose a spider employs a similar approach, requiring a user to click on an analogous "Accept" button before performing a search. This approach becomes problematic if the user is not human. Would the agreement be enforceable against an intermediator who would never "read" the terms to which it was "agreeing?" In order for the spider to hold the intermediator to its terms of service, the spider would have to show that an agreement was reached. Presumably, the spider could argue that an intermediator knew or should have known that they designed the spider's interface to be human-readable, that the disclaimer and acceptance button were part of this interface, and therefore, that availing oneself of the spider's services constitutes acceptance of the spider's terms. This argument does not address the difficult issue of when, or how quickly, an intermediator should be aware of changes to the terms of the license,[159] but it should suffice in the main goal of providing enforceability.

77. Another approach might be to pursue a technical fix. On one level, machines make agreements all the time, such as when two modems "negotiate" the protocol and speed they will use to speak with each other. Possible solutions for providing notice to an intermediator of the spider's licensing terms could be the definition and acceptance of an addition to the HTTP protocol,[160] a new HTML META[161] tag, or a new standard entry in a site's robots.txt[162] file. This site could then alert anyone, human or machine, that the site does not wish its contents to be freely redistributable.

V. UNFAIR COMPETITION

78. Even if the URLs returned by a spider are unprotectable facts, and even if

[158] *Id.*

[159] *See* Arizona Retail Sys., Inc. v. Software Link, Inc., 831 F. Supp. 759, 766 (D. Ariz. 1993) (shrinkwrap license unenforceable when consumer and buyer had negotiated agreement previously).

[160] HTTP stands for "hypertext transfer protocol," and is the method of communication between Web clients and servers. *See* WIRED STYLE, *supra* note 1, at 134.

[161] HTML stands for "hypertext markup language." *See id.* Web pages are written in HTML. *See id.* For a discussion on how META tags can be used to communicate with spiders, see *AltaVista Search: Help for Simple Query* (visited Feb. 10, 1997) <http://altavista.digital.com/cgi-bin/query?pg=h#meta>

[162] For a full explanation of robots.txt, see *Web Robots FAQ, supra* note 24. The concept of robots.txt was adopted in response to the excessive traffic caused by spiders. *See id.* The file robots.txt sets out what areas of a site are off-limits to the spider's harvesting engine, as well as other parameters setting out what constitutes accepted use of the site by non-humans. *See id.* Adherence to a machine's robots.txt file is purely voluntary. *See id.*

licensing proves unworkable, the spider may use unfair competition law to protect its results. This Part will examine the applicability of an equitable notion of unfair competition and its codification in section 43 of the Lanham Act.[163]

A. *International News Service v. Associated Press*

79. The practice of intermediating bears a striking resemblance to the circumstances in *International News Service ("INS") v. Associated Press ("AP")*, in which the Supreme Court conferred protection to otherwise unprotectable facts through the application of the equitable notion of unfair competition.[164]

80. INS pilfered AP's news by obtaining early editions of AP newspapers on the East Coast and telegraphing their contents for use in INS papers on the West Coast.[165] Although acknowledging that AP's news consisted of unprotectable facts, the Supreme Court recognized a quasi-property right in news,[166] due to its "novelty and freshness."[167] The Supreme Court's opinion of INS's practices could also apply to an intermediator that strips a spider's advertisements and replaces them with its own:

> Stripped of all disguises, the process amounts to an unauthorized interference with the normal operations of complainant's legitimate business precisely at the point where the profit is to be reaped, in order to divert a material portion of the profit from those who have earned it to those who have not.[168]

From this general principle, it becomes necessary to look at the specific business to determine what constitutes unfair competition.[169]

81. Courts could apply *INS* to the spider's case because the spider's results resemble news. Like a newspaper, the spider may also return works of authorship. Unlike news, the spiders' main assets are thoroughness, completeness, ease of use, and response time, but not so much the "novelty and freshness"[170] that made the

[163] Lanham Act § 43, 15 U.S.C. § 1125 (1994).

[164] 248 U.S. 215, 241-42 (1918).

[165] *See id.* at 231.

[166] *See id.* at 236.

[167] *Id.* at 238.

[168] *Id.* at 240.

[169] *See id.* at 236 ("Obviously, the question of what is unfair competition in business must be determined with particular reference to the character and circumstances of the business.").

[170] *Id.* at 238.

news sufficiently unique to confer protection.[171] Despite the differences in the media, the repeated pilfering of facts by a meta-spider is closely analogous to *INS*, where "[i]n effect, going to the well once may not be actionable; only frequent return trips to another's product may enable the state claim to persist."[172]

B. *Section 43 of the Lanham Act and the AltaVista Mixup*

82. Intermediating may be actionable under section 43 of the Lanham Act, which codifies the federal tort of passing off by prohibiting "false designation of origin, false or misleading description of fact, or false or misleading representation of fact."[173] A meta-spider could conceivably escape section 43 liability by clearly labeling the sources from which it gathered its results, thus avoiding the likelihood of confusion as to the origin of the data. That, however, would not necessarily excuse the meta-spider from other elements of section 43 liability. If a meta-spider, or other intermediator were to strip a spider's advertisements and replace them with its own, they would probably be found liable for attempting to "deceive as to . . . sponsorship."[174] A recent mixup involving a spider and an intermediator illustrates this sponsorship element.

83. When the Digital Equipment Corporation ("Digital") wanted to name their new spider "AltaVista," they found that AltaVista Technology, Inc. ("ATI"), a small software company, was already using the name.[175] Digital purchased the name outright and granted ATI a limited license in return.[176] Soon after the launch and attendant publicity surrounding Digital's AltaVista search engine, hundreds of thousands of users went to ATI's web site[177] by mistake.[178] Rather than display a prominent disclaimer explaining the difference between the two AltaVistas and

[171] The protection is limited to "the extent necessary to prevent that competitor from reaping the fruits of complainant's efforts and expenditure." *Id.* at 241. The Supreme Court refused to modify the lower court injunction, which set enjoined INS from copying AP's reports "until its commercial value as news to the complainant and all of its members has passed away." *Id.* at 245.

[172] Jane C. Ginsburg, *No "Sweat"? Copyright and Other Protection of Works of Information after* Feist v. Rural Telephone, 92 COLUM. L. REV. 338, 357 (1992).

[173] Lanham Act § 43(a)(1), 15 U.S.C. § 1125(a)(1) (1994).

[174] *Id.*

[175] *See* Digital Equip. Corp. v. AltaVista Tech., Inc., No. 96-12192NG, 1997 WL 136437, at *1 (D. Mass. Mar. 12, 1997); Kirkpatrick, *supra* note 52, at B18.

[176] *See Digital*, at *1; Interview with Steve Bauer, Partner, Testa, Hurwitz & Thibeault, in Boston, Mass. (Dec. 14, 1996) [hereinafter Bauer Interview].

[177] *AltaVista Technology, Inc.* (visited Mar. 31, 1997) <http://www.altavista.com/> [hereinafter *ATI*].

[178] *See Digital*, 1997 WL 136437, at *3; Kirkpatrick, *supra* note 52, at B18.

redirecting users to Digital's site, ATI set up a front-end to Digital's AltaVista search engine.[179] Then they sold advertising,[180] leading Digital to bring action against ATI to cease and desist.[181] Digital then filed suit. The district court granted Digital a preliminary injunction, noting that Digital would probably succeed on the merits as to the trademark license breach, trademark infringement, and unfair competition claims.[182]

VI. CONCLUSIONS

84. Strictly construed, copyright law does not provide adequate protection to spiders, but licensing, the Lanham Act's Section 43, and unfair competition law may provide alternatives.

85. Licensing is one possible way to work around this limitation, but the standards for determining agreement have not yet been decided. Given the commercial potential of the Web, these standards should develop rapidly.

86. Section 43 of the Lanham Act currently provides the best statutory protection for spiders' search results, but there remains the slim possibility that an intermediator could escape liability by providing conspicuous disclaimers about the exact nature of their activities.

[179] *See Digital*, 1997 WL 136437, at *2-3.

[180] *See Digital*, 1997 WL 136437, at *2-3; Kirkpatrick, *supra* note 52, at B18.

[181] Bauer Interview, *supra* note 176.

[182] The ATI site provided a front-end to Digital's site, but gave notice that the front-end uses Digital's AltaVista search engine. *See Digital*, 1997 WL 136437, at *4-5.