Academic Data Collection in Electronic Environments: Defining Acceptable Use of Internet Resources[1]

Gove N. Allen

Freeman School of Business
Tulane University
New Orleans, LA 70118
gallen@tulane.edu

Dan L. Burk

Law School
University of Minnesota
Minneapolis, MN 55455
burkx006@umn.edu

Gordon B. Davis

Carlson School of Management
University of Minnesota
Minneapolis, MN 55455
gdavis@csom.umn.edu

[1]An earlier version of this paper was presented at the 4th annual meeting of the Association of Internet Researchers in Toronto Canada, October 16-19, 2003

Academic Data Collection in Electronic Environments: Defining Acceptable Use of Internet Resources

Gove N. Allen received his Ph.D. from the University of Minnesota in 2001 and currently serves as an assistant professor of e-business and information systems at Tulane University's A. B. Freeman School of Business. He holds bachelors and masters degrees in Accountancy from Brigham Young University. Dr. Allen has consulted major corporations on the implementation of information technology including Sony, AT&T, Sprint, Hewlett Packard, Micron, Intel, 3M, American Express, and NASA. More recently, he has made presentations at academic gatherings regarding automated data collection in the North America, Europe, and Australia. His continuing research utilizes various configurations of automated Internet data collection agents.

Gordon B. Davis is the Honeywell Professor of Management Information Systems, Emeritus at the Carlson School of Management at the University of Minnesota.. He is a pioneer in management information systems research and education. Professor Davis is the author of 23 texts and over 200 articles in management information systems, data processing, programming, and EDP auditing. His book, *Management Information Systems: Conceptual Foundations, Structure and Development,* 1974, Second Edition (with M. Olson), McGraw-Hill Book Company, 1985, has been ranked as a classic in the field

Dan L. Burk is the Oppenheimer, Wolf and Donnelly Professor of Law at the University of Minnesota, where he teaches courses in Patent Law, Copyright, and Biotechnology Law. An authority on issues related to high technology, he is the author of numerous papers on the legal and societal impact of new technologies, including articles on scientific misconduct, regulation of biotechnology, and the intellectual property implications of global computer networks. He is perhaps best known for his work in the area of "cyberlaw," where he has been a leading figure in the debates surrounding Internet jurisdiction, trespass to computers, and the deployment of digital rights management systems.

Academic Data Collection in Electronic Environments: Defining Acceptable Use of Internet Resources

Abstract

Academic researchers access commercial websites to collect research data. This research practice is likely to increase. Is this appropriate? Is this legal? Such commercial websites are maintained to achieve business objectives; research access uses site resources for other purposes. Website administrators may therefore deem academic data collection inappropriate. Is there a process to make research access more open and acceptable to website owners and administrators? These are significant issues. This article clarifies the problems and suggests possible approaches to handle the issues with sensitivity and openness.

Research access to commercial websites may be manual (using a standard web browser) or automated (using automated data collection agents). These approaches have different effects on websites. Researchers using manual access tend to make a limited number of page requests because manual access is costly to perform. Researchers using automated access methods can request large numbers of pages at a low cost. Therefore, website administrators tend to view manual access and automated access very differently.

Because of the number of accesses and non-business purpose, automated research requests for data are sometimes blocked by site administration using a variety of means (both technological and legal). This paper details the pertinent legal issues including trespass, copyright violation, and breech of contract. It also explains the nature of express and implied consent by site administration for research access.

Based on the issues presented, guidelines for researchers are proposed to reduce objections to research activities, to facilitate communication with website administration, and to

achieve express or implied consent. These include **notification** to website administration of intended automated research activity, **description** of the research project posted as a web page, and clear **identification** of automated requests for web pages. In order to encourage good research practices with respect to automated data collection, suggestions are made with respect to disclosing methods used in research papers and for self regulation by academic associations.

Keywords: Internet, research, automated data collection

1. INTRODUCTION

The Internet is used for many academic research purposes. The focus of this paper is academic access to commercial website resources. Two examples of recent publications in this journal demonstrate some of the varied approaches to such data collection.

Ba and Pavlou (2002) conducted a study of trust building in electronic markets. Using a web browser, they collected data about 682 Ebay auctions and feedback information for their sellers over a 45 day period. They used this information to demonstrate that positive characteristics of seller feedback profile lead to price premiums for sellers. Bapna et al. (2004) wrote "an automated agent to capture information directly from auction websites" (p. 27). The agent collected data every five minutes over twelve months. The data was used to show five distinct classes of bidders and how their proportions changed from 1999 to 2000.

These articles are only two of many that rely on information from commercial websites. A review of three top outlets for information systems research (MIS Quarterly, Information Systems Research, and Management Science) for the years 2002-2004 indicates eight articles that make use automated data collection techniques from commercial websites (Clemons et al. 2002; Palmer 2002; Bapna et al. 2003a; Bapna et al. 2003b; Brynjolfsson et al. 2003; Easley & Tenorio 2004; Pavlou & Gefen 2004; Bapna et al. 2004). At least another twelve make some use of commercial website resources without automated interaction (Ba & Pavlou 2002; Zhu & Kraemer 2002; Chen & Hitt 2002; Kim et al. 2002; Koufaris 2002; Dewan et al. 2003; Dellarocas 2003; Pinker et al. 2003; Bolton et al. 2004; Bhargava & Choudhary 2004; Snir & Hitt 2003).

As the role of the Internet continues to increase, studies of commercial Internet sites are crucial to academic research. The conditions for collecting data required for such inquiry are

important for both practical and legal reasons. The benefit of using data from commercial websites for academic research is clear. Since the sites are publicly accessible, our field has apparently given little thought to issues of research access, especially automated access. However, commercial websites almost universally post terms of service (TOS) documents that expressly restrict the manner in which individuals my use the site. The strict adherence to these terms would often exclude academic inquiry of any kind. In fact, with one exception (Koufaris, 2002) each article cited in the prior paragraph, would violate the terms of at least one website TOS document if conducted today[2].

This paper seeks to illuminate the legal issues surrounding the use of commercial websites for academic research. It suggests actionable guidelines for researchers who make use of commercial websites. The rest of this paper is organized as follows: Section 2 summarizes the typical ways that academic researchers make use of commercial websites. Section 3 describes the means that commercial web sites employ to control how their data are accessed. Section 4 describes the current legal issues pertaining to unauthorized access of websites. Section 5 details our recommendation for the field for the appropriate use of commercial websites in academic research and Section 6 presents our conclusions.

2. DESCRIPTION OF ACADEMIC USE OF COMMERCIAL INTERNET RESOURCES

The academic use of commercial Internet resources varies widely. Some studies use commercial sites only to gain an understanding of a context they model analytically; some use them for background or general information; some study the site *per se*, sending subjects or researchers to evaluate specific functions within the site; and, others collect data publicly

[2] Because Snir and Hitt (2003) do not disclose site identity, their study excluded

available on the site for analysis. Table 1 categorizes in to three groups the 20 articles found in *MIS Quarterly*, *Information Systems Research*, and *Management Science* from 2002 to 2004 that made use of commercial Internet resources.

Use	**Study**
General information, background, context	Bhargava & Choudhary 2004; Chen & Hitt 2002; Bolton et al. 2004; Pinker et al. 2003; Dellarocas 2003; Dewan et al. 2003; Pavlou & Gefen 2004
Study website *per se*.	Koufaris 2002; Kim et al. 2002; Zhu & Kraemer 2002; Palmer 2002; Agarwal & Venkatesh 2002
Analyze data collected from site	Ba & Pavlou 2002; Brynjolfsson et al. 2003; Bapna et al. 2003a; Bapna et al. 2003b; Easley & Tenorio 2004; Clemons et al. 2002; Bapna et al. 2004; Sinr & Hitt 2003

Table 1. Use of commercial Internet resources in recent IS literature

More than a third of the research reported in these articles made use of automated Internet data collection agents to accomplish their research objectives. Such automated practices merit special discussion because their use raises additional issues. Automated Internet data collection agents are simply computer programs designed to interact with web servers via the Internet to collect various kinds of data (Kauffman et al. 2000). They retrieve web pages and parse them to find data to be stored for analysis, references to images to be downloaded, and links to other web pages that might contain useful information. Because the agents have the capability to simulate data being entered into a web form (such as a customer order form) and posted by a user, they can dynamically interact with electronic commerce web servers and collect detailed data about various practices and behaviors in the online environment. Many programming environments provide tools for the development of such agents, so they are increasingly simple to develop and deploy in meaningful ways (Allen and March 2000).

Automated data collection agents vary widely in their architecture and sophistication. They can retrieve documents in either a serial or parallel fashion. When an automated agent

requests a page from a web server, it typically takes much longer to retrieve the page than it does to parse it into data elements and process those data. An agent that processes its requests serially (waiting for the first request to be retrieved and parsed before the second request is issued), places a much lower burden on the web server with which it is interacting than an agent that issues concurrent requests for a large number of pages and parses them as they are received. Data collection agents can also be built to operate in a distributed manner in which several copies of an agent can be running simultaneously on different computers, all working together to complete the same data collection task. Because Internet data collection agents are designed to interact with web servers over the Internet, they are easily adapted to cooperate with each other over the same channel, making it relatively simple to deploy massively parallel, geographically dispersed data collection networks with hundreds or thousands of nodes. Typically, details about the particular data collection agent used are not disclosed in academic articles, although the personal interactions of the authors with several researchers actively using this technology indicate that current research employs the full range from the simplest to the most complex.

3. MEANS USED BY WEBSITES TO CONTROL ACCESS TO DATA

Although our experience indicates that few commercial websites have specifically considered the implications of either allowing or preventing access for academic research, virtually all have employed a variety of means to grant access to some while restricting access by others. For competitive reasons, commercial websites employ various technological and legal means to prevent unauthorized access to their data.

Technological Means to Control Access to Website Content

Technological mechanisms for such control include password protection to portions of websites with restricted data, encryption of data transmissions to prevent unauthorized access to

their data while it is sent across the public network, copy protection, and traffic monitoring coupled with page request refusal for access that may be undesirable because of its volume or because the site objects to a particular use such as a competitor accessing information about product offerings.

The technological mechanisms work directly to prevent unwanted access, that is, they actively deny access to content unless the requests present the proper credentials. The use of technological controls alone is unsatisfactory for four reasons. First, there is typically some way to circumvent the technology: encryption can be cracked, copy protection is regularly defeated, and passwords can be obtained under false pretenses. Second, even if access is correctly granted, once content has been delivered, there may be inadequate means to control its subsequent use. For example, a user may purchase a copy of a movie intended for in-home use and later use it in a public setting. Third, the implementation of technical means may place extra burden on users, as for example, requiring a user to read and enter a code rendered as a graphic with irregular type in order to access specific content. Finally, technical means generally place increased demands on the infrastructure. For example, encrypted transmissions are larger (sometimes by a factor of two) than their unencrypted counterparts.

Legal Means to Control Access to Website Content

To complement the performance of technological means of controlling the access and use of data, websites have several legal means at their disposal. Legal means currently in use include seeking a court order to enjoin individuals or organizations against access or seeking the redress of damages on the basis of trespass, copyright infringement, or breach of contract. Unlike their technological counterparts, legal means work indirectly, relying on court injunctions, the threat

of court action, and the costs of potential litigation to deter individuals from inappropriate use of system data and resources. There are three legal claims site owners typically make to control the use of their resources: copyright infringement, trespass to chattel, and breach of contract. Although issues surrounding copyright infringement are significant, they are typically of small concern to academic researchers as will be discussed in the next section. Trespass to chattel and breach of contract both hold significant implications for academic research and both depend heavily on the terms specified in two documents maintained by sites concerned about access. These are a site's terms of service (TOS) document and a robots.txt file.

Terms of Service Document (TOS)

The home page of a commercial website generally has a link to the site's terms of service (TOS) document (also often called "terms of use" document). Along with specifying the details of various site policies, this document describes acceptable use of the site.

Because of the central role that TOS documents play in asserting the legal rights of a website, and because they are written with the intent of forming a binding contract with users of the site, they are typically lengthy and written in precise language. Moreover, they tend to be written as adhesion contracts (drafted entirely by one party and heavily restrict the other party) and allow only a few very specific uses of the site. Because of the inherent inequality in such contracts, courts tend to interpret any ambiguity in favor of the non-drafting party. Accordingly, their terms often appear impractical. Consider the following excerpt from the walmart.com TOS document:

> Your use of the Site following any such change constitutes your unconditional agreement to follow and be bound by these Terms of Use as changed. For this reason, we encourage you to review these Terms of Use whenever you use this Site. (http://www.walmart.com/catalog/catalog.gsp?cat=119985, 3/28/2005).

Considering that when printed single-spaced in 12 point type, the walmart.com TOS document is 12 pages, the statement that users should read it each time they visit the site is unreasonable. In fact, companies must know that almost no users have even seen the site's TOS document. An analysis of the web logs of one major corporation's web site indicated that out of every one million visits to their home page, fewer than five followed the link to the TOS document. This demonstrates one of the very interesting characteristics of the TOS document: It is intended to support litigation rather than to enlighten users with respect to the proper use of the site. When a website wants to teach users the appropriate way to make use of site resources, it does so with techniques that are much more user accessible, such as frequently-asked-question pages (FAQs) and Help pages.

The role of the TOS document as a legal tool rather than the definitive document that details acceptable use of site resources is demonstrated in the following excerpt also taken from the TOS document at walmart.com.

> You may download or copy the Contents and other downloadable materials displayed on the Site for your personal use only. No right, title or interest in any downloaded materials or software is transferred to you as a result of any such downloading or copying. You may not reproduce (except as noted above), publish, transmit, distribute, display, modify, create derivative works from, sell or participate in any sale of or exploit in any way, in whole or in part, any of the Contents, the Site or any related software (http://www.walmart.com/cservice/terms.gsp, 3/28/05)

This policy clearly permits copies of the site's data for some personal use, and indicates that commercial use of the data is prohibited. However, this statement directly conflicts with another document that describes allowable use of the site: robots.txt.

Robots.txt File

The robots.txt file plays a role that is similar to TOS documents; however, its use is limited to automated retrieval tools. Written in compliance with Standard for Robot Exclusion

(Koster 1994), the robots.txt file provides specific instructions for the kinds of automated retrieval tools (called robots or bots) used by Internet search sites like google.com to index Internet resources. By reading the file, which is located in the root directory of a web server, a bot can determine what parts of the site it is allowed to access. Like a TOS document, robots.txt does not exert any technological control over the use of site resources; individuals who write bots may or may not implement the functionality to read and respect the restrictions it contains. The robots.txt file for walmart.com follows:

```
# go away
User-agent: *
#Disallow: /
Disallow: /solutions
Disallow: /cservice
Disallow: /reflect.gsp
```

(http://www.walmart.com/robots.txt, 3/28/05)

This policy indicates that all bots are specifically restricted from requesting documents that begin with "walmart.com/solutions," "walmart.com/cservice," or "walmart.com/reflect.gsp," but that all other areas of the site are available for automated retrieval. In this case, walmart.com has specifically allowed bots access to all product and pricing information located under "walmart.com/catalog." At the same time, it restricts bots from accessing parts of the site dealing with order tracking, the store locator, and product recall information, all of which are located under "walmart.com/cservice." Clearly, the intent of the robots.txt file is to allow bots to index the content of the site to make it available to be searched by users of Internet search engines such as google.com, yahoo.com, and altavista.com—all commercial endeavors, even though a strict reading of the TOS document clearly proscribes this kind of access.

Thus, walmart.com allows itself to be indexed by search engines while preserving the strength of its TOS document. The contradiction is likely intentional and demonstrates that the wording of TOS documents should not be taken as the absolute position of a site with regard to the use of its resources.

4. LEGAL ISSUES PERTAINING TO INTERNET ACADEMIC DATA COLLECITON

The legal tools that websites employ to control access to their resources largely rely on three separate legal principles: trespass, copyright infringement, and breech of contract.

Trespass as a Basis for Legal Challenge to Academic Use of Commercial Websites

The unauthorized use of website resources by commercial entities has been the subject of considerable case law and courts have begun to formulate legal standards for such activity (Burk 2000; O'Rourke 2000; Elkin-Koren 2001). By far the most successful legal claim asserted against on-line data gathering has been the theory of trespass to chattels, an ancient legal doctrine intended to protect the owners of moveable property (chattels) against interference with the use of their property. At common law, a claim of trespass to chattels requires contact be made with the chattel, notification that contact was unwelcome, and interference with or dispossession of the chattel, resulting in some harm or damage to the chattel, or pecuniary loss to the owner. This definition of trespass has been reformulated for networked computers to hold that electrical impulses satisfy the common law requirement of physical contact, and the increased load on the networked system qualifies as interference or dispossession, and terms posted in the TOS document or robots.txt fill the requirement of notice. Harm or pecuniary loss is often presumed from the loss of processing cycles, the use of network transmission capacity, or the diversion of data storage capacity because the owners of web sites bear the cost of responding to requests made to their servers. While any single incremental request does not bear a direct variable cost,

clearly the infrastructure needed to respond to 10 million requests per day is more costly than the infrastructure needed to respond to a thousand requests per day.

Several courts have now embraced this renovated legal theory. For example, in *eBay v. Bidder's Edge*, a United States District Court relied on a theory of trespass to enjoin Bidder's Edge, an aggregator of on-line auction data, from collection of data from the eBay site (eBay 2002). Similarly, in *Register.com v. Verio*, a trespass theory was used to penalize the collection of ownership data from a publicly accessible domain name database (Register 2000). Most recently, in *American Airlines v. Farechase*, a Texas state court enjoined the producer of software used to search in real time for air fares on airline websites (American 2002).

Several trespass cases have as an alternate theory relied upon federal statutes prohibiting unauthorized access to computers. Section 1030 of the U.S. criminal code penalizes unauthorized access to a networked computer when damage results. This statute was originally enacted in response to the Robert Morris Cornell "worm" incident. Unfortunately, Congress failed to define "authorization," allowing this statute to be applied to a broad range of Internet activity where permission to engage in the activity has been implicitly or explicitly denied. For example, in the case of *EF Cultural Travel v. Explorica*, automated retrieval of travel pricing data by a bot was held to constitute unauthorized access to a travel agency's website (EF Cultural Travel 2001). Similarly, in the recent *Farechase* decision mentioned above, access to the airline website was also enjoined on an alternate theory of unauthorized access under a state computer crime statute. Many states have such statutes that may be applied in addition to the federal unauthorized access statute.

These cases are significant for academic researchers because they establish an almost absolute right for website owners to exclude others from their publicly available equipment and information. Although these cases do not deal directly with the academic use of commercial Internet resources, the legal theories advanced do not provide for any research or public benefit exceptions. Any load placed on a networked server without prior consent may be considered trespass or unauthorized access. Although liability for trespass or other legal claims depends to a great extent on the terms of the researcher's employment contract, both the researcher and the sponsoring institution could be targets for legal action.

Copyright as a Basis for Legal Challenge to Academic Use of Commercial Websites

In the commercial context, copyright claims have also been asserted against some undesired use of system resources. However, this claim cannot extend to the mere collection of facts. Under the United States Supreme Court holding in *Feist v. Rural Telephone*, facts are generally not copyrightable because they lack originality (Feist 1991). Compilations of facts may be protected by copyright, if the selection and arrangement of the facts is original. Extracting the facts from their matrix of selection and arrangement takes nothing copyrightable, so long as the selection or arrangement is not also taken. Consequently, copyright claims asserted to stop unwanted access of system resources have not been successful to date.

A more credible copyright claim might be made for copies of web pages generated in the course of access by automated agents. Computer networks function by reproducing digitized files and transmitting them piecemeal to other computers where they are reassembled. Along the way, all or part of the files are reproduced by routers and other intermediate computers. When automated data agents request web pages from networked servers, the pages served typically are

reproduced in several places, including the recipient computer cache file and in the RAM of the recipient computer. In operational terms, the Internet comprises a massive, distributed copy machine and any access to files on the network produces copies.

Additionally, automated search agents may be designed to request a page and store that page as originally downloaded, in addition to parsing it for the information the researchers wants. This practice has significant academic value. When longitudinal data is analyzed, some observations may be puzzling; by keeping an original record of the web pages, a researcher can refer back to the source data in its original context to examine it more closely. The anomalous observations may be accurate, they may appear because of an error in the program that generated the file at the website, or they may appear because of an error in the data collection agent. Without the original HTML file, determining the cause of the anomaly is difficult at best, and most often impossible. Perhaps more importantly, when the original HTML file is stored, it can be reparsed by a researcher without placing any load on the originating website. Accordingly, this practice benefits the research process as well as the website owners. However, storage of original HTML pages constitutes a conscious form of archiving or copying beyond the copies made in the normal course of the Internet's technological function.

Under United States law, copies that are fixed long enough to be perceived, either directly or with the aid of a machine, are subject to copyright law. The Copyright Act indicates that copies existing only temporarily in computer memory fall outside the Act, but copies that exist in more durable media, such as those written to a hard drive, are likely subject to copyright law. Some courts have held that even RAM copies may be relevant for copyright purposes.

Thus, copies made when web pages are served up could constitute copyright infringement if they are not authorized by the owners of the pages. Typically, such authorization is implied from the public availability of the pages. Due to the nature of the technology, they can only be viewed if copies are made, so we infer that the owners knew or intended that such copies should be made in order for the pages to be viewed.

However, implied consent can usually be revoked by some sort of notice to users. Continued use of the copyrighted material in the face of explicit notice that access is unauthorized could then constitute infringement. In such a case (Kelly 2002), at least one court has held that display of images served to a search engine constituted copyright infringement; however, academic research agents are unlikely to publicly display the pages they search.

Alternatively, the creation of such copies in the course of Internet access may constitute fair use. In the United States, unauthorized uses of copyrighted materials may be permissible under the fair use provisions of the Act, which particularly favors uses for non-commercial educational or research purposes. Several courts have held that creation of temporary or intermediate digital copies in the course of a permissible end use may constitute fair use. Fair use principles may also tend to favor academic archiving of web pages that are under study.

It must be remembered, however, that the Internet is international in its reach, and fair use does not exist in the copyright law of any country other than the United States. Many other countries have limited exceptions or “fair dealing” provisions that may permit certain otherwise unauthorized uses. These provisions often include exceptions for educational or research purposes. Not only do the exceptions vary by country, but copyright law in some regions is currently undergoing substantial change. For example, the March 2001 European Union

Copyright Directive required member states to make changes to their national copyright laws. Although the directive allowed member countries substantial latitude in their national adoption, many member states have still not legislated the required changes but continue to debate over various drafts. Thus, not only is the storage of web pages for analysis not permissible in some countries from which the pages originate or to which they may be transmitted, but for many countries, that status may be changing.

Contract Breech as a Basis for Legal Challenge to Academic Use of Commercial Websites

The terms posted in TOS documents purport to form a binding contract with those using the site where the TOS document is posted. Under this theory, the TOS document is held to act in the same manner as "shrinkwrap" or "clickwrap" licenses accompanying software. Under a legal fiction of agreement, users are said to manifest assent to the terms of the contract by the act of making use of the site. However, formation and enforceability of such contracts is problematic, both because consumers may not have seen the terms, and because terms of the agreement may be unconscionable or surprising. Courts have been split on the enforceability of such purported contracts. Virginia and Maryland have recently amended their contract law by adoption of the Uniform Computer Information Transaction Act, or UCITA, to make this type of contractual assertion legally enforceable. There has been significant opposition to adoption of this act in other states, including enactment by a few states of legislation explicitly prohibiting the application of UCITA to their citizens. Thus, the contractual implications of posted TOS document may vary not only from country to country, but from state to state in the United States

Achieving Express Consent to Avoid Legal Challenges

As the discussion of TOS documents demonstrates, each legal objection to automated access hinges upon the degree of access consented to by the owner of a website, and the inferences that can be drawn from website owner actions regarding consent. The challenges to automated data collection agent access for research can be removed by obtaining express consent to access the site. Although conceptually desirable, express consent is frequently impractical.

Specific express consent can be obtained by a researcher for a particular research project. In any situation where a researcher interacts with a certain web server, the owner of the server clearly has the right to expressly allow the researcher's data collection activity. It has been our experience that when owners of small commercial sites are informed of an intention to request small (and even moderate) amounts of data from their servers for purposes of academic research, some are very willing to work with the researcher. Accordingly, one approach for a researcher to take is to seek express consent. When express consent has been granted, many of the other legal issues cease to be concerns.

Obtaining express consent may not be a satisfactory approach for several reasons. First, in some situations, the resources consumed in the process of obtaining express consent (both on behalf of the researcher and on behalf of the organization whose consent is sought) may far outweigh the resources that would be used in obtaining the data. Consider, for example, what the process might entail in gaining express consent from any substantial player in the online community. Once an initial request for express consent is sent, it may several days (or weeks) for the request to even reach the person or committee with the authority to grant consent. Once the decision maker has the request there may be other issues that prohibit consent from being

granted. Decision makers may be concerned about setting a particular precedent within the organization, or they may be unsure of the legal ramifications of granting express consent. Moreover, since there is likely very little direct benefit to the organization, even the process of considering such requests may outweigh the perceived benefits, leading to a policy of summarily rejecting all such requests. One researcher sought express consent to use automated data-collection agents to gather data from a major e-commerce site, even visiting the company headquarters. Although upper management of the site was interested to learn of the results of the studies the researcher had already conducted using data from their site, they would not grant permission for continued automated access. However, after the meeting, one member of the management team gave additional guidance to the researcher in a separate communication. The executive told the researcher that the company's site had mechanisms in place to monitor automated data access and that anytime automated data access became onerous, site administrators take corrective action. He explained that because the researcher's prior data collection activity had not been "noticed" by the site's administration, he could reasonably expect to continue his activity at similar levels without objection by the site's management. Although anecdotal, this incident demonstrates the tendency for organizations to be conservative in formal decisions, even as informal channels may be less so.

A second concern with requesting express consent is that once the dialogue is opened, it is entirely possible that, beyond not granting express consent, an officer of the company may expressly forbid a researcher from conducting the study. In this case any implied consent that might have reasonably been inferred by the connection of the server to a public network would clearly be revoked.

A third concern with seeking express consent is that such contact with web site administration may threaten the validity of some research studies. In much the same way that people behave differently when they know they are being observed, websites may present different information if they know they are being examined for purposes of an academic study. In 2000, Amazon.com came under heavy fire for giving different prices to customers depending on which browser was used to issue the request as well as the age of Amazon cookie in the browser's cache on the customer computer (Rosencrance, 2000). Clearly the technology exists to relay different information based on the characteristics of the request. If there is reason to believe that a site, once aware of the study, may tailor the information sent to an academic researcher, it may contaminate the data collection to notify the organization of the study by a request for express consent.

There are significant differences between seeking express consent from a website to use their data in academic research and negotiating research access to the resources of organizations in a traditional research setting. The differences are based on the fact that in automated site access, the data to be accessed for the research activity has already been placed before the public by posting it on a website. In a traditional research setting, an organization being asked for access to processes or data will typically have five concerns: access will disrupt normal operations, direct cost associated with granting access, known or unknown legal issues with releasing some information, publication or leaking of information may have adverse effects on the company's competitive position, and allowing researchers access to particular locations may expose the company to undue liability. In the case of automated access to websites, organizations address each of these issues before making data and processes publicly accessible.

Implied Consent when Express Consent is Absent

In some cases, commercial websites post neither a TOS document nor a robots.txt file. This is the reason that one of the research articles listed in Table 1 (Koufaris, 2002) would not violate the terms of service if it were conducted today—it made use of booksamillion.com, which posts neither a TOS document nor a robots.txt file. When such is the case, one might reasonably infer that an organization grants "implied consent" for the public to access its web servers, even through the use of heavily automated means, because they are connected to a public network in which the use of automated retrieval mechanisms is pervasive.

5. RECOMMENDATIONS FOR RESEARCHERS COLLECTING DATA FROM COMMERCIAL WEBSITES

The following guidelines for researchers reflect our views regarding manual website access and researcher circumvention of technological restrictions. There are three recommendations for actions to communicate with website administration when using automated access to websites. The discussion of the recommendations reflects the perspectives of both website administration and academic researchers. The views and recommendations are intended to guide academic researchers conducting studies that make use of publicly available, commercial Internet resources or are contemplating such research.

Views on Manual Access to Websites and Circumvention of Technological Restrictions

Manual, non-automated access of information on publicly available web pages should be acceptable without special permissions or actions. Even though the website may not expressly permit such access for research, the load on the website is negligible. Such limited access outside the directly intended purpose of a website is expected by site owners. Therefore, limited access by those doing research fits within normal website expectations. Moreover, the strict

enforcement of TOS document terms in this situation would virtually close commercial websites to any examination by academia. This would be bad public policy and may lead to corrective legislation, weakening the efficacy of existing legal means for controlling access to site content.

Any action taken by an academic researcher to bypass technological means of controlling access to website resources is unacceptable. Such actions show that the researcher is acting in bad faith and will reduce the credibility that other academic researchers hold with commercial websites. Perhaps more importantly, the circumvention of technological measures to gain access to data clearly violates the Digital Millennium Copyright Act (DMCA) in the United States, as well as provisions set forth in the European Union Copyright Directive (2001/29/EC). Accordingly, such behavior may have criminal consequences. If access to data is restricted using technological means, the only acceptable way for a researcher to access the data is to seek express consent in cooperation with website management.

Recommendations for Automated Access to Websites

Automated access of information on publicly available web pages can place a substantial load on site resources. For such access that complies with a site's robots.txt file, or in its absence, specific terms about automated retrieval in the TOS document, we suggest two principles to facilitate communication between the researcher and the website. These are *description* (providing site administration a document that describes the research) and *identification* (providing site administration with the ability to easily determine the effect of the research on the site). In addition, when the robots.txt file (or in its absence, the TOS document) does not expressly permit the planned access a recommended procedure is *notification* of research activity.

Provide Description of Research Activity as a Webpage

Commercial websites are becoming increasingly aware of the need to limit competitor access to their site resources (YiHua, et al. forthcoming). Unfortunately, the activity of a competitor using an automated, data-collection agent often bears strong resemblance to a researcher using such technology. For this reason, it is important to clearly describe that the data collection activity is strictly for academic research purposes. A detailed description of the research study (including potential findings) can be posted as a web page. This may make a system administrator more likely to allow access than if no information is provided about the research. The study's description document should provide the email address of the appropriate member of the research team and perhaps a telephone number. It may also include a statement of the researcher's intention to not disrupt the site's operations and to comply with the wishes of site management, as well as a statement about whether the identity of the site will be disclosed in any research report.

In a particular longitudinal study in which no description of the academic nature of the study was provided to website administrators, one website blocked access to the researcher by restricting the range of IP addresses allowed to access the site. Later when the researcher apologized for causing trouble at the site, the site administrator indicated that if he had known that the requests were part of an academic study, he would not have terminated access. There are different approaches to making a website aware of the academic nature of a particular study; one is to include a reference to the descriptive webpage with each automated request.

Provide Identification of Source with each Automated Request

A website administrator can usually determine the source of accesses on the site through the analysis of log files stored by web servers. Commercial web servers have the functionality to record information about each request made of the server. They record specified portions of each HTTP (hypertext transfer protocol) request. The information in the request header is used to specify the requested resources on the server and also to indicate information about the program (e.g., a bot or a browser) that issued the request.

The User-agent field of the HTTP request header is used to uniquely identify the program (and version) that is requesting the resources of the server. The field may also include a comment enclosed in parentheses that can be used for research access identification. For example, Netscape Navigator version 6.2 sends the following in the User-agent field of the request header:

Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:0.9.4.1) Gecko/20020508 Netscape6/6.2.3

This indicates that the user agent is based in the open source Mozilla code base and also supplies information about the current operating environment in the comment portion of the field (enclosed in parentheses). Finally, it has information regarding the version of Navigator.

Researchers should therefore use the User-agent field of the HTTP request to identify their data-collection agent. Moreover, a researcher should use the comment portion of the field to direct a site administrator to the web page that describes the current research endeavor. For example, a researcher might formulate the user-agent field for all HTTP requests of a particular study as follows:

AcademicAgent/1.0 (http://research.tulane.edu/~gallen/Study7.html)

This helps the site administrator identify how much traffic is being generated by the particular agent and allows easy access to the document that describes the research project. This

approach works well because the User-agent field is almost universally logged by all web servers, so when an administrator examines the logs, the required information is readily available. By making requests identifiable in this manner, a researcher facilitates communication with a concerned administrator.

Following these suggestions to describe the academic intent of the research access along with the practice of identifying each automated request places some burden on the researcher but it also provides them the benefit of initiating an open, cooperative relationship with the website rather than a closed, adversarial one. Clearly, in this setting, website administration will favor these recommendations because it provides more information as they decide which automated requests to allow and which to deny using various technological tools for blocking access.

In Some Cases, Send Notification of Research Activity

If a researcher wishes to pursue a data collection protocol that is not expressly permitted by the robots.txt file (or in is absence, TOS documents), he or she should send a message to the webmaster of the site indicating his or her intentions. This notification should indicate the expected duration of the data collection, the intended frequency of requests, a description of the kinds of data to be collected, a means to contact the appropriate member of the research team, and a statement of the researcher's intention to cooperate with the site's administration. The notification need not be phrased as a request for permission, and a researcher should not expect to receive a response. This way, the administrator can easily express any concerns to the researcher; however, if the amount of data to be requested is of no concern, the administrator may simply ignore it. From the perspective of the academic researcher, the requirement to send notification is a burden, and may result in express notification that the research activity is

unwanted; however, it demonstrates that the researcher is acting in good faith which may be important to the resolution of a website's concerns without legal action. From the perspective of the website, this is an agreeable arrangement for several reasons. First, it removes any burden the site would otherwise have in detecting unwanted automated access. Second, it allows research activity with limited infrastructure impact to be allowed without the time required for a formal approval process. Third, it allows the site to permit some research data collection activity without expressly setting a precedent. Fourth, it reduces uncertainty by allowing site to observe the effect of the data collection activity before rendering an official decision. Fifth, it preserves the site's absolute ability to revoke access at any time because formal permission is never granted.

We should mention that suits have been filed alleging violation of the DMCA because provisions in the robots.txt file were ignored. However, none have been successful to date and precedent is emerging (e.g. Lexmark 2005) that more clearly excludes specifications of a robots.txt file from being considered "effective technical protections" as specified in the DMCA.

Recommendations for a Generally Accepted Policy for Disclosing Automated Data Collection Procedures in Research Papers

Rather than waiting for legislation to define research access, it may be helpful for academic researchers to agree on acceptable behavior. For example, a formal policy of accepted behavior for automated data collection could be adopted by an appropriate organization such as the Association for Information Systems (AIS) or the Association of Internet Researchers (AoIR). This would provide the basis for wording in research papers indicating researcher compliance with the policy or if the policy was violated, an explanation of why. Until such a policy is in place, reviewers may reasonably require authors to provide a brief description of the

automated data collection system (perhaps in a footnote) and list any communication between the researchers and web site administrators—such as requests for consent made by the researcher or inquiries made by website administration into the data collection activity.

Establishing such a policy which could be broadly adopted on a voluntary basis may hold significant future implications. Just as industry self-regulated adoption of accepted privacy principles in the context of Fair Information Practices has led to proposed legislation that complements self-regulation programs rather than supplants them, our field has the opportunity to establish the standard that could influence potential future legislation in a manner that may favor academic research as a preferred activity—especially if we as academic researchers can demonstrate effective self-regulation.

6. CONCLUSION

Researchers are currently making substantial academic use of commercial Internet resources. Such research activity is important in developing our understanding of the many organizational aspects that are so deeply affected by the Internet. Because these commercial Internet resources are publicly available, researchers have given little thought to the processing load that their activity places on website infrastructure. Commercial sites have begun to exercise various legal means to limit individuals from making automated access of their resources. Academic research is not exempt from the legal arguments that have been successfully advanced.

Commercial websites define acceptable access of their resources in their terms of service document often in conjunction with the robot.txt file located at the root directory of a web server. Although the terms of service document may be useful for defining research access, it is often written to support legal action against any possible site misuse. Consequently, its terms should not be viewed as absolute with respect to academic research. The robot.txt file, when present,

defines specific areas of a site that are available to or restricted from automated interaction. Accordingly, the robots.txt file should be viewed together with the terms of service document in assessing the appropriateness of automated academic access.

Given some ambiguity in a website's position on allowing academic research access, we recommend three polices to assist researchers in dealing with the practical problems of automated data collection for research purposes: description (provide a webpage that describes the research activity), identification (identify source of each page request along with a reference to description), and notification (if research access is not expressly permitted) by sending a message indicating purpose, scope, and frequency of requests.

If researchers can arrive at reasonable policies for automated research access of Internet sites, then self regulation may be useful in dealing with site administration, as well as important in influencing emerging legislation pertaining to academic use of commercial internet resources. A professional organization is the appropriate place for such a policy to be formulated.

While the practices that we outline here will not grant academic researchers legal carte blanche to crawl Internet websites or immunize researchers from legal liability, we expect that adherence to these guidelines may allow researchers to avoid or defuse most legal conflicts. Formal adjudication is costly, and is unlikely to be pursued if reasonable informal solutions are available. In a situation where informal channels are not explored before legal action is taken by a website owner, adherence to these practices demonstrates that a researcher has acted in good faith, which may be critical in obtaining a favorable outcome.

REFERENCES

Agarwal, R. and V. Vankatesh. 2002. "Assessing a Firm's Web Presence: a Heuristic Evaluation Procedure for the Measurement of Usability." *Information Systems Research*. 13 (2), 168-186.

Allen, G. and S. March. 2000. "Developing Internet agents: a Tutorial using Visual Basic 6.0." *The Proceedings of the International Conference on Information Systems 2000*. Brisbane, Australia.

American Airlines v. Farechase, No. 067-194022-02 (Tx Dist. Ct. Tarrant Cty., March 8th 2002)

Ba, S. and P. A. Pavlou. 2002. "Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior. " *MIS Quarterly*, 26(3), 243-268

Bapna, R., P, Goes, and A. Gupta. 2003a. "Replicating Online Uankee Auctions to Analyze Auctioneers' and Bidders' Strategies." *Information Systems Research*. 14 (3), 244-268.

Bapna, R., P, Goes, and A. Gupta. 2003b. "Analysis and Design of Business-to-Consumer Online Auctions." *Management Science*. 49 (1), 85-101.

Bapna, R., P, Goes, and A. Gupta. 2004. "User Hetrogeneity and its Impact on Electronic Auction Market Design: An Empirical Exploration." *MIS Quarterly*, 28(1), 21-43

Bhargava, H. K. and V. Choudhary. 2004. "Economics of an Information Intermediary with Aggregation Benefits." *Information Systems Research*. 15 (1), 26-36.

Bolton, G. E., E. Katok, and A. Ockenfels. 2004. "How Effective are Electronic Reputation Mechanisms? An Experimental Investigation." *Management Science*. 50 (11), 1587-1603.

Brynjolfsson, E., Y. Hu, and M. D. Smith. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers." *Management Science*. 49 (11), 1580-1596.

Burk, D. L. (2000) The trouble with trespass. Journal of Small & Emerging Business Law, 4, 27-56.

Chen, P. Y. and L. M. Hitt. 2002. "Measuring Switching Costs and the determinants of Customer Retention in Internet-Enabled Businesses: A Study of the Online Brokerage Industry." *Information Systems Research*. 13 (3), 255-274.

Clemons, E. K., I. Hann, and L. M. Hitt. 2002. "Price Dispersion and Differentiation in Online Travel: An Empirical Investigation." *Management Science*. 48 (4), 534-549.

Dellarocas, C. 2003. “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms.” *Management Science*. 49 (10), 1407-1424.

Dewan, R., B. Jing, and A. Seidmann. 2003. “Product Customization and price Competition on the Internet.” *Management Science*. 49 (8), 1055-1070.

Easley, R. F. and R. Tenorio. 2004. “Jump Bidding Strategies in Internet Auctions.” *Management Science*. 50 (10), 1407-1420.

eBay Inc. v. Bidder’s Edge, Inc., 100 F.Supp.2d 1058 (N.D. Cal. 2000).

EF Cultural Travel BV v. Explorica Inc, 274 F.3d 577 (1st Cir. 2001)

Elkin-Koren, N. 2001. Let the Crawlers Crawl: On Virtual Gatekeepers and the Right to Exclude Indexing. Univ. Dayton Law Rev. 26, 179-209.

Feist Publications, inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991)

Kauffman, R., S. March, and C. Wood. "Mapping Out Design Aspects for Data-Collecting Agents." *International Journal of Intelligent Systems in Accounting, Finance, and Management* 9 (4), 217-236.

Kelly v. Arriba, 366 F.3d 811, 822 (9th Cir. 2002)

Kim, J. , L. Jungwon, K. Han, and M. Lee. 2002. “Businesses as Buildings: Metrics of the Architectural Quality of Internet Businesses.” *Information Systems Research*. 13 (3), 239-254.

Koster, M. 1994. “A Standard for Robot Exclusion.” http://www.robotstxt.org/wc/norobots.html. (on June 1, 2005).

Kougaris, M. 2002. “Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior.” *Information Systems Research*. 13 (2), 205-223.

Lexmark v. Static Control 387 F.3d 522, 6th Circuit (2005)

O'Rourke, M.A. 2000. Shaping Competition on the Internet: Who Owns Product and Pricing Information?. *Vanderbilt Law Rev*. 53, 1965-2006.

Palmer, J. W. 2002. “Web Site Usability, Design, and Performance Metrics.” *Information Systems Research*. 13 (2), 151-167.

Pavlou, P. A. and D. Gefen. 2004. “Building Effective Online Marketplaces with Institution-Based Trust.” *Information Systems Research*. 15 (1), 37-59.

Pinker, E. J., A. Seidmann, and Y. Vakrat. 2003. “Managing Online Auctions: Current Business and Research Issues.” *Management Science*. 49 (11), 1457-1484.

Register.com, Inc. v. Verio, Inc., 126 F. Supp.2d 238 (S.D.N.Y. 2000).

Rosecrance, L. 2000. "Amazon charging different prices on some DVDs." *Computer World*, (Sept. 5).

Sheng, Y., P. Mykytyn, and C. Litecky, Forthcoming. "The Internet, Intelligent Agents, Competitor Analysis and its Defense in the E-marketplace." *Communications of the ACM.*

Snir, E. M. and L. M. Hitt. 2003. "Costly Bidding in Online Markets for IT Services." *Management Science*. 49 (11), 1504-1520.

Zhu, K. and K. Kraemer. 2002. "e-Commerce Metrics for Net-Enhanced Organizations: Assessing the Value of e-Commerce to Firm Performance in the Manufacturing Sector." *Information Systems Research*. 13 (3), 275-295.