Legal Concerns of Web Site Reverse Engineering

Holger M. Kienle, Daniel German, and Hausi Müller
Computer Science Department
University of Victoria
Victoria, Canada
{kienle,dmgerman,hausi}@cs.uvic.ca

Abstract

Researchers that are involved in Web site reverse engineering are often not aware of potential legal implications of using someone else’s Web site for experimentation. Even if researchers are concerned with legal problems, there is little guidance available. This paper explores the legality of Web site reverse engineering with the intent to raise awareness among researchers about this issue. The discussed legal issues encompass copyright, contract, and trespass law.

1. Introduction

“No longer an information ‘wild, wild west,’ the Internet increasingly is influenced by legal considerations.”

– Baker in [22]

A broad definition of reverse engineering (RE) is the process of extracting know-how or knowledge from a human-made artifact [48]. An alternative definition is provided by the U.S. Supreme court, who defines it as “a fair and honest means of starting with the known product and working backwards to divine the process which aided in its development or manufacture” [4]. This process typically starts with lower levels of abstraction to create higher levels of understanding. RE has a long-standing tradition in many areas, ranging from traditional manufacturing to information-based industries. This paper discusses RE of software in general and RE of Web sites in particular.

From an economic perspective, RE can have both negative and positive effects. On the one hand, RE can reduce incentives to innovate. On the other hand, it can increase competition and spur innovation in the marketplace. If RE is undertaken by a competitor for the purpose of developing an alternative product, the original inventor has still sufficient incentive if its development cost and lead-time result in a competitive advantage compared to the costs of the competitor’s RE effort. Inventors can also try to make their product harder to reverse engineer. For instance, obfuscation techniques can be used for software [36]. On the extreme side, prohibition of RE can result in de facto monopolies. In general, the effect of RE on the market depends on factors such as its required time, difficulty, cost, and purpose of the RE effort.

RE is constrained—implicitly and explicitly—by law. There are laws in some domains that explicitly address RE. For example, the chip industry has the Semiconductor Chip Protection Act (SCPA), which was passed by the U.S. Congress in 1984. The SCPA protects chips from RE for the purpose of manufacturing cloned chips. However, it also contains a RE provision that makes it possible to reproduce the mask work, to analyze it, and then to incorporate the obtained results in one’s own original design [39]. In contrast, software has no legal protection of its own kind (i.e., sui generis). The U.S. Copyright Act of 1976 included software as literary works that are protectable by copyright. As a consequence, RE of software has to be implicitly addressed within the general framework of copyright law.

Motivation for RE of software is varied, ranging from trouble-shooting of bugs and ensuring interoperability to cracking and pirating. Because of the latter, RE is sometimes perceived as illegal or un-ethical as a whole. However, RE of software is a standard industry practice and many practitioners believe it makes economic sense, promoting competition and innovation [5, 48].

RE of software falls into two distinct groups: Binary RE (or low-level RE) recovers source code from software that is only available in binary form, whereas high-level RE is typically concerned with the analysis of source code with the objective of recovering its design and architecture [6].

RE of software can employ white-box as well as black-
box techniques [7]. In white-box RE, the code itself is analyzed to gain understanding. If the code is in binary form, disassembly or decompilation is often a necessary first step. In black-box RE, understanding is achieved by observing the behavior of the code via inputs and outputs. The latter technique is especially common when an interface is reverse engineered (e.g., to achieve interoperability). For example, to understand the functioning of a Web site, the reverse engineer can vary the inputs (e.g., via modification of the requesting URLs, including PUT and GET requests, cookie data, etc.) and observe the corresponding outputs (i.e., responses of the Web server).

Proprietary software is shipped in binary form; developers keep the source code as trade secrets [5]. Thus, binary RE is typically performed by competitors, but also if one’s own source has been lost. In this case, techniques such as emulation, disassembling, and decompilation are used to gain insight into the object code [6]. In contrast, high-level reverse engineering is typically conducted with one’s own source code or with the permission of the code owner. As a consequence, researchers of high-level RE did not have to deal with legal issues, whereas researchers that develop tools to facilitate binary RE had to deal with the legal ramifications of their research results. Cifuentes, a researcher in the binary RE field, has published extensively about legal issues [6, 7, 9, 8].

Web site reverse engineering (WSRE) is pursued by researchers in the high-level RE field,3 which might explain why there have been no concerns voiced about the legal ramifications so far. Furthermore, Web sites are often perceived as “open” source and “public domain,” reflecting remnants of the belief that all information on the Web is free and unregulated [30, 35].

Web site owners, however, have become more concerned about protecting their investment (e.g., intellectual property such as content and design). This is especially the case for commercial sites. As a consequence, some Web site owners try to use legal means to control the way in which others are utilizing their site. Examples are the explicit prohibition of RE activities via the site’s terms of use (contract law) and of crawling of the site (trespass law).

These legal developments potentially affect RE research. Researchers routinely use commercial Web sites for case studies.4 While researchers can hope for a sympathetic reaction of Web site owners if their site is picked for a case study, it should not be taken for granted. This paper explores the legality of WSRE with the intent to raise awareness among researchers about potential legal problems.

The paper is organized as follows: Section 2 explains WSRE in more detail and states what forms of WSRE raise legal concerns. Section 3 explores these concerns in more detail, addressing copyright, contract, and trespass law. Section 4 draws some conclusions.

2. Web Site Reverse Engineering (WSRE)

Since WSRE is a new form of RE and not as well established as traditional RE, we give a brief overview of the environment, process, and (research) tools.

We define WSRE as the process of gaining understanding about the functioning and structure of a Web site and the information contained in it. This definition is rather broad, covering activities such as

- extraction of information to obtain Web site metrics or reports.
- combination of information from different Web sites to create new services.
- understanding the (link) structure, contents, and implementation techniques of a Web site.

WSRE is performed by different parties, ranging from academics as part of their research activities to better understand the Web to organizations that want to understand other Web sites (including competitors’) to leverage that knowledge for their own site.

When discussing WSRE it is useful to distinguish between the different views of a Web site [27]:

client view: The view of the Web site that a client (typically using a Web browser) sees.

deployment view: The view of the Web site that a Web server (accessing the local file system) sees.

developer view: The view of the Web site that a developer (using a Web development tool such as Dreamweaver) sees.

A particular WSRE tool can operate on one or several of these views. For example, a Web crawler operates on the client view, a JSP extractor can operate on the deployment view, and a HTML extractor can operate on both client and deployment view. So far, there are few WSRE tools targeting the developer view.

WSRE tools that operate on the client view can be used without prior consent of the Web site owner as long as the site is publicly available. This is similar to a binary RE tool, which can be run on someone else’s binary code assuming the code is legally obtained. In the client view, reverse engineers can also make use of black-box RE techniques to better understand the functioning of the site. In contrast, WSRE tools that operate on the deployment or developer view need full access to the code base, which is only available by consent of the Web site owner. Thus, such tools are

---

3This observation is based on several conferences that publish RE research such as CSMR, IWPC, CSMR, and WSE.

4Some researchers pick their own university Web sites (e.g., see [33]) or demo sites (e.g., see [19]) for their case studies. However, to avoid bias, commercial sites should be used as well.
comparable to high-level RE tools, which operate on source code.

Client-view WSRE has some interesting new properties compared to traditional software systems: (1) the code on the server side (source as well as binary form) is not accessible, (2) one part of the Web site is executed by the server (e.g., CGI scripts and servlets), the other part by the client (e.g., HTML rendering, JavaScript interpretation, and Applet execution), (3) client side code can be in binary form (e.g., Flash and Applets) or source (e.g., HTML, XML, and JavaScript), and (4) accessing a Web site (typically) does not require agreement to a license.

This paper concentrates on an environment in which RE is conducted on the client view without consent (similarly to traditional white-box binary RE). In this case, several legal issues can arise between the reverse engineer and the Web site owner (discussed in Section 3).

2.1. Client-View WSRE Process

The client-view WSRE process can be structured into several distinct activities:

downloading: The content of a Web site is retrieved from a Web server (via HTTP requests). This is the raw, unaltered data (mostly in the form of HTML pages). Many researchers have developed their own tools (e.g., see [33] [61] [31] [19] [57]), but general-purpose tools such as GNU Wget, cURL,5 and Lynx6 can be utilized as well. Usually the whole content of a site is downloaded. For some reverse engineering activities it is sufficient to download only a certain subset of the site. In this case downloading is governed by configuration files that provide field input values and cookie parameters [41, 23].

During downloading, Web proxies such as Proxomitron (http://www.proxomitron.info/) can be used to monitor and analyze the flow of information between browser and Web site.

extracting: In this step, artifacts (such as links and HTML document structure) are extracted from the raw data obtained via downloading. This step requires some form of HTML extractor and possibly other extractors (e.g., for JavaScript) [25].

Some extractors parse the complete source (e.g., [61] [33]) while others use lightweight, lexical approaches (e.g., [57] [19] [23]).

5cURL (http://curl.haxx.se/), a.k.a. curl, is a command line tool for downloading of files using URL syntax.
6Lynx (http://www.lynx.browser.org/) is a text-based Web browser. Its -traversal option allows to download a Web site’s pages.

analysis: Sophisticated analyses can operate on the extracted artifacts to (semi-automatically) produce new, more abstract artifacts.

Examples of analyses are detection of navigational patterns of a site [42], clustering of pages [33], tracking of the evolution of a site [40], matching contents in a multilingual site [58], clone detection [56], and various site and page metrics [62] [31].

archiving: After downloading, the contents can be archived locally for future reference and repeatable experiments. Some analyses need local copies of a site. For example, one study of the evolution of Web pages downloaded about 151 million URLs; 0.1% of those were randomly chosen for archiving for detailed analysis [14].

For some analysis it is sufficient to archive derived information from a page. The Web evolution study mentioned above archived a word-based feature vector of every downloaded page. A rudimentary example is the tracking of whether a certain page changed after the last download with the archiving of only a single checksum of the page’s contents.

reporting: The result of analyses have to be reported in a suitable format to the reverse engineer for exploration. RE results are also reported in published case studies by researchers.

RE information is often represented in a graph model. In the ReWeb tool, results of analyses are visualized with dotty, a scriptable graph editor [41]. Similarly, the Rigi tool has been customized to visualize the structure of Web sites [33] [27]. Reports can be also provided in HTML and displayed on a browser. Such reports typically contain links that point to the corresponding Web site artifacts.

The discussed RE activities are not sequential. Black-box WSRE is often unstructured and ad-hoc because the reverse engineer explores the site interactively. For example, Udell discusses WSRE using lightweight tools such as telnet, Perl, stunnel (a SSL decryptor), and Proxomitron [59, 60].

There are other RE activities such as crawling and (screen) scraping. Crawling can be characterized as downloading and extraction of a Web site, possibly followed by archiving. Search engines such as Google use crawling to obtain information from the Web. Google archives the crawled version of the page as well (unless it is otherwise instructed by the page’s meta-tags).

Scraping is an ad-hoc technique that combines extraction and analysis to directly obtain content from pages formatted in HTML. Scripting languages are often used for simple scraping based on regular expressions [29]. For example,
O’Reilly’s Amazon Hacks book contains several examples of how information from the Amazon site can be scraped with Perl [3]. WebL is a programming language that has been specifically developed to facilitate the processing of structured Web documents such as HTML [28, 32]. Its markup algebra allows the programming of access patterns that reference certain parts in an HTML document. The Web Scraping Proxy (http://www.research.att.com/~hpk/wsp/) generates Perl snippets that can be used to develop a scraper for a certain site. Besides RE, scraping is also useful for automated regression testing.

3. Legal Concerns

The Web is increasingly influenced and affected by legal issues. The state and federal courts in the United States have published in 2001 more than 1,100 opinions involving the Web [22]. While many of these cases can easily apply existing law in the new context of the Web, some raise unique legal issues. Similar to other areas on the Web (e.g., domain names, meta tags, and linking [43, 13]), it can be expected that WSRE will become more of a legal issue in the future.

WSRE combines legal issues of software RE with the Web. Both of these areas are not well established and still evolving. As a consequence, the legal uncertainty is relatively great. Still, previous cases involving RE of software can be used as guidance for WSRE.

Figure 1 gives an overview of the legal issues. The WSRE activities introduced in Section 2.1 are shown along with the laws that potentially affect them. Because of space limitations, this paper can provide only a brief treatment. For this reason, important legal cases are given along with references that provide more in-depth discussion. Since we assume that the reader is a reverse engineer, weavored references from computer science publications.

In the following, we assume U.S. law and explicitly note when discussing laws in other countries. There are several international treaties (e.g., for copyright there is the Berne Convention of the World Intellectual Property Organization—WIPO) that are the bases for national laws, guaranteeing a certain uniformity. However, there are still many more or less subtle differences between countries. For example, the U.S. grants a longer copyright protection than most other countries—this is the result of the Sonny Bono Copyright Term Extension Act passed by Congress in 1998.

3.1. Copyright

A copyright applies to work that is expressed in a “tangible medium for the purpose of communication” [2]. Examples of such works are literature, music, paintings, movies, and television broadcasts. In some countries, copyright also includes computer programs and information embedded in databases. A copyright is subject to no formalities, which means that no action is required from an author to obtain a copyright.

The copyright law has to balance the rights of the content creator to be compensated adequately for his or her efforts, and the rights of citizens for access to information. Copyright grants several pecuniary rights, which are held exclusively by the owner. Among these are making and selling copies of the work and making derivative works. A consumer right is fair use, which allows the consumer to use copyrighted material in certain ways regardless of the copyright owner’s wishes. For example, limited quoting from a copyrighted book is permissible under fair use.

Copyright law can be directly applied to electronic copies of digital media [44]. Most digital copyright cases on the Web have involved images (e.g., Playboy vs. Sanfilippo and Kelly vs. Arriba Soft). The contents of a Web site (text, images, JavaScript, etc.) have a copyright even without a copyright notice. As with other works, Web sites can be also registered by the Copyright Office [11]. An exception are Web sites of the Federal Government, which are in the public domain. Web sites can also use an explicit open content license such as Creative Commons (http://creativecommons.org/) to replace copyright.

Archiving of copies of whole Web pages or sites can infringe on copyright. In practice, problems arise if such copies are further distributed or posted online without permission of the copyright owner. Careful consideration to copyright issues is necessary if an archived version is made publicly available, for instance, to facilitate a RE benchmark such as WebETS among researchers [26].

As part of the software RE process, the software is copied into machine memory (RAM) by a tool. Such RAM copies cannot be avoided. The same holds for WSRE when the contents of a Web site are downloaded. This raises the question whether this intermediate copy is a reproduction of the work and hence constitutes a copyright infringement.

Several cases have found that intermediate copies are indeed covered by copyright. In Intellectual Reserve vs. Utah Lighthouse Ministry, the owners of a Web site hosting copyrighted material were held liable of contributory infringement because of intermediate RAM copies in the users’ browsers [15]. Addressing the users’ RAM copies, the court stated [20]:

When a person browses a website, and by so doing displays the [copyrighted material], a copy of the [copyrighted material] is made in the computer’s random access memory (RAM), to permit viewing of the material. And in making a copy, even a temporary one, the person who browsed infringes the copyright.
<table>
<thead>
<tr>
<th>RE Activity</th>
<th>Law</th>
<th>Infringement</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>downloading, extracting</td>
<td>copyright</td>
<td>intermediate RAM copy</td>
<td><em>Intellectual Reserve vs. Utah Lighthouse Ministry</em> (1999) [20, 15],</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><em>MAI Systems vs. Peak Computers</em> (1993) [51], <em>Nintendo vs. Atari</em> (1992) [52, 4],</td>
</tr>
<tr>
<td>extracting, archiving,</td>
<td>contract</td>
<td>terms of use</td>
<td><em>Vault vs. Quaid</em> (1988) [45], <em>ProCD vs. Zeidenberg</em> (1996) [54, 12],</td>
</tr>
<tr>
<td>downloading</td>
<td>trespass</td>
<td>server access</td>
<td><em>eBay vs. Bidder’s Edge</em> (2000) [37, 46], <em>Register.com vs. Verio</em> (2004) [63, 1],</td>
</tr>
</tbody>
</table>

**Figure 1. Summary of laws and cases that affect Web site reverse engineering**

In *MAI Systems vs. Peak Computers*, an appellate court ruled that loading software into a computer creates a fixed copy of that software [51]. (Copyright law permits the software buyer certain copies, but the court ruled that this does not apply if the software is licensed.) Even though this interpretation of intermediate copies has been widely criticized, it seems now firmly established in case law. Since WSRE causes intermediate copies during downloading and extraction the same issues apply as for software RE.

U.S. courts have excused intermediate copies for certain RE purposes on the grounds of fair use. (The European Community’s Directive on the Legal Protection of Computer Programs explicitly allows black-box RE.) In *Nintendo vs. Atari* and *Sega vs. Accolade* decompilation of programs to achieve interoperability was ruled permissible. This decision carries over to white-box RE where no decompilation step is necessary [7]. Besides achieving interoperability, there might be other legitimate purposes such as correcting errors and improving performance or portability.

The *Ticketmaster vs. Tickets.com* case addresses several interesting legal issues, among them RAM copies [20]. Ticketmaster operates an event Web site that contains short descriptions about events and enables users to order event tickets. Tickets.com is a competitor that also advertises events and sells tickets. Tickets.com also lists events for which they do not sell tickets. In order to obtain certain event information (such as date, place, and time), Tickets.com downloaded Web pages from Ticketmaster and extracted event information from them. Among other claims, Ticketmaster alleged that Tickets.com committed copyright infringement. The district court ruled in March 2003 that “taking the temporary copy of the electronic information for the limited purpose of extracting unprotected public facts leads to the conclusion that the temporary use of the electronic signals was ‘fair use’ and not actionable.”

For a fair use defense, the court has to consider four factors in determining whether a particular use (here, RE) of the copyrighted work (here, Web sites) is fair: “(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.” The first factor weighs in favor of academic research, but courts have to consider all four. In *Ticketmaster vs. Tickets.com*, the court found that the first factor weighs against Tickets.com because it operates its site for commercial purposes, but found that the other factors weigh in its favor.

The *Digital Millennium Copyright Act* (DMCA), passed by Congress in 1998, amends the Copyright Act and is a response to digital media and the Web. The DMCA generally prohibits RE to circumvent copyright-protection mechanisms. There are a few exemptions that permit RE, most notably for achieving program-to-program interoperability and testing the security of computer systems. However, these exceptions are often quite narrow, limiting their applicability. There is no exception for academic research. Since most RE research is not concerned with interoperability or
testing issues per se, they cannot hope to fall under the exemption.

It is still unclear, how far-reaching the DMCA actually is. In Lexmark vs. Static Control, Lexmark tried to use the DMCA's anti-circumvention provisions to prevent competition for its toner cartridges. Lexmark sued Static Control because they reverse-engineered Lexmark's toner cartridges, which contain a certain code that is used by the printer to establish compatibility. A court granted Lexmark an injunction against Static Control in February 2003. However, the U.S. Copyright Office ruled in October 2003 in favor of Static Control [66].

A number of legal cases suggest that Web sites cannot expect to freely link to other Web sites as a general rule. This means that reporting of RE information using links that refer to the original Web site might be a copyright infringement under certain circumstances. For example, the framing of contents from another Web site might constitute a derivate work and as such a copyright infringement [53]. In this case, the user of the Web browser's user might be the direct infringer and the frame provider a contributory infringer. The first case that involved framing was filed in 1997 by a group of six content providers (among them the Washington Post) against TotalNews, which operates a Web site that frames the Web sites of online newspapers. TotalNews packaged the news stories with their own advertisement in a separate frame. In the complaint, the plaintiffs claimed, among other things, copyright infringement. The case was settled out of court in December 1997. As part of the settlement TotalNews stopped framing the suing newspaper sites.

Wood analyzes the legal risk surrounding linking and gives the following recommendations [67]: (1) do not deep link without permission, (2) do not frame other entities' pages without permission, and (3) do not use links to sites that encourage others to engage in copyright infringement.

Another recent aspect of law questions whether reporting of information obtained by RE could violate copyright. In 1996, the European Union adopted the European Parliament and Council Directive on the Legal Protection of Databases (EU Directive) [21]. Member countries of the EU had to implement corresponding laws by January 1998. (The U.S. currently does not protect databases, but several bills are pending in Congress [17].) The EU Directive defines a database as "a collection of works, data or other independent materials arranged in a systematic or methodical way and individually accessible by electronic or other means." This definition covers Web pages that provide a collection of information, such as real estate listings or news stories. A German appellate court accepted that any Web site qualifies as a database because it is a collection of Web pages.

The copyright holder of the database has the right "to prevent extraction and/or reutilization of the whole or of a substantial part." Thus, it might constitute a copyright infringement if substantial information is extracted (e.g., via scraping) from Web sites and subsequently publicly reported. For example, in NVM vs. De Telegraaf, De Telegraaf operated a real estate search agent that extracted listings from the Web site of a broker [21]. A lower court held that even extraction of small pieces of information can be infringing if that data is of great value to end users. Subsequently, the Dutch Supreme court ruled in the case that owners of online databases can prohibit deep linking to the contents of their database.

The Directive allows countries to define exceptions for scientists that download data for educational or research purposes, but not all countries have done so [34]. Even with that exemption, reutilization and republishing of the extracted information is still not permitted.

Assuming that a certain RE activity infringes on copyright, one might ask whether the tool that assists in the infringement is deemed illegal. The Supreme court ruling on Sony vs. Universal City Studios (1976) indicates that a tool is not illegal if there are also legitimate uses for it [44]. Sony, the manufacturer of the Betamax video recorder, was not held liable for contributory copyright infringement even though the machines certainly can be used for illegal activities. In Sony's favor weighted that the machines also have "significant noninfringing uses." Similarly, in Vault vs. Quaid (1988), Quaid reverse engineered Vault's copy protection system and marketed a circumvention tool. Vault alleged contributory copyright infringement of its customers' programs, but the court ruled against Vault because Quaid's tool enabled users to make back-up copies as copyright's fair use provision allows them to do.

However, since the decisions of Sony vs. Universal City Studios and Vault vs. Quaid, Congress has passed the DMCA, which is unique in the sense that "for the first time in history, it isn't the copyright violation that was the crime. It is the creation of the technological tools to violate copyright that became the crime" [10]. Isenberg concludes that had the DMCA's anti-tool rules been in existence at the time of the Sony vs. Universal City Studios ruling, "the entire home video recording industry might never have gotten off the ground" [22]. Currently MGM tries to overturn the Sony ruling in a case against peer-to-peer file-sharing software in MGM vs. Grokster [47].

### 3.2. Contracts

Often software is not owned, but licensed and in order to obtain and run the software it is necessary to accept the

---

Section 1201(a)(1)(A) of the United States Copyright Act: "No person shall circumvent a technological measure that effectively controls access to a work protected under this title."
terms and conditions stated in the license agreement. Licenses are governed by contract law, not copyright law.  

For mass-market software, so called shrink-wrap licenses are used, which are supposed to come into effect when the software is installed and not returned to the store. A variation is the click-wrap agreement, in which the consumer expresses consent by clicking “I agree” when viewing the license terms during installation or download from the Web.

A license often contains contractual restrictions that prohibit decompilation and RE. For example, the end user license agreement of Adobe Acrobat Reader 5.0 states:

You also agree not to reverse engineer, decompile, disassemble or otherwise attempt to discover the source code of the Software except to the extent you may be expressly permitted to decompile under applicable law, . . .

Are such anti-RE terms enforceable? Samuelson and Scotchmer stated in 2002 that “the caselaw in the U.S. is in conflict on the enforceability of anti-[RE] clauses in software contracts” [48]. Contract law is a state law, which can be preempted by federal copyright law. In *Vault vs. Quaid*, a court ruled that the anti-RE clause of a shrink-wrap license that Vault wanted to enforce under Louisiana’s License Act as preempted by copyright law and hence unenforceable. (In the EU the law explicitly states that anti-decompilation clauses in software contracts are void [48, n.228].) In *ProCD vs. Zeidenberg* a shrink-wrap license was held enforceable for the first time [54]. Since then, more courts have followed that decision, also for click-wrap licenses [12]. In *Harold Bowers vs. Baysate Technologies*, an appellate court did not accept the preemption argument for a shrink-wrap license and ruled that its anti-RE clause is valid [38].

Possibly encouraged by these developments in contract law, more and more Web sites try now to enforce their own rules via the site’s terms of use. Courts have labeled terms of uses that one can view, but need not explicitly agree to, as browse-wrap agreements. For example, the Maxim Web site (http://www.maxim-ic.com/) has a Legal Notices link on the bottom of every page. Clicking on it brings up a page containing the Terms of Use and User Conduct. The former states that if one does not agree with any terms “you are required to exit this Website immediately.” The latter reads:

You shall not in any manner: (i) distribute, use or make copies of, or otherwise duplicate any materials contained on the Website . . . (iii) cause or permit the reverse engineering, disassembly, decompilation of, or any other attempt to derive any software comprising or available through the Website, except to the extent permitted by law; In addition, without limitation, You agree not to use the Website: (a) to harvest, collect, gather or assemble any Materials for other than internal, noncommercial use; . . .

This contract significantly restricts RE activities. Archiving is explicitly prohibited even though already covered by copyright. Extraction and reporting are allowed, but publication of the results are prohibited.

Google prohibits scraping of their search results. Some sites (including Maxim) have link policies, striving to control incoming links from other sites. While the Maxim site is certainly an example of quite restrictive terms of use, it can be expected that this trend will intensify.

Even though click-wrap licenses are probably enforceable, it is unlikely that simply providing a link to them at the bottom of pages is sufficient. In *Ticketmaster vs. Tickets.com*, the court declined in 2000 to enforce Ticketmaster’s terms of use shown at the very bottom, which users were not required to review and explicitly accept. Subsequently, Ticketmaster changed the site, showing a more visible warning, which said that going beyond the start page meant that users accept the site’s terms of use. In this case, the court was more sympathetic and ruled in March 2003 that

“a contract can be formed by proceeding into the interior web pages after knowledge (or, in some cases, presumptive knowledge) of the conditions accepted when doing so” [20].

This seems to indicate that a contract might be valid even though the user has not explicitly agreed to it (e.g., by pressing an “I agree” button).

Browse-wrap agreements are the weakest form of contract. Still, reverse engineers—and Web site users in general—should probably take them seriously and check their contents. Courts might reason that reverse engineers presumably have knowledge of the terms of use because their usage of the site goes beyond mere browsing.

If a Web site has no terms of use, the site’s owner has given the user some implied rights (implied license doctrine) [20, 15]. Opinions about what is permissible by such an implied license differ widely. By making material available on a Web site, one can argue that the owner expects it to be downloaded for browsing and possibly archived. However, courts tend to interpret implied licenses narrowly.

---

8The Uniform Commercial Code (UCC) provides the legal framework for contracts in most states. Louisiana has a sales law similar to UCC, while Maryland and Virginia have adopted the Uniform Computer Information Transactions Act (UCITA) [24].

9Section 301 of the Copyright Act states that federal copyright law overrules state law claims when they are equivalent to copyright law claims, meaning that state law cannot regulate the same thing that copyright law does [51].
Hence, the lack of terms of use cannot be simply interpreted as implied acceptance of all RE activities.

### 3.3. Trespass

From a Web site’s viewpoint, not all visitors might be equally welcome. Some sites, for example, want to restrict crawling and/or scraping of their contents. Owners of sites have used illegal trespass claims in an attempt to prohibit unauthorized access of their site. Such owners view their Web sites as valuable “real estate” that they want to protect from intruders.

Trespass to chattel\(^\text{10}\) applies if there is substantial interference with chattel, such as damage or impairment. Trespass to land occurs when a person enters another’s property without the owner’s consent. Web sites can be treated as property and as a consequence allow to sue for trespass [18]. This means that accessing a Web site for downloading against the owner’s wishes could be illegal trespass.

In eBay vs. Bidder’s Edge, eBay sued Bidder’s Edge (BE) for trespass to chattel because of a Web crawler that frequently visited eBay’s site to scrape auction information [46]. The court ruled in favor of eBay:

> “it is undisputed that eBay’s server and its capacity are personal property, and that BE’s searches use a portion of this property. Even if, as BE argues, its searches use only a small amount of eBay’s computer system capacity, BE has nonetheless deprived eBay of the ability to use that portion of its personal property for its own purposes.”

Register.com vs. Verio is a similar case where a district court agreed to Register.com’s trespass claims and granted a preliminary injunction, barring Verio to crawl and obtain information from Register.com’s Whois database. The appellate court upheld that injunction [63].

However, not all courts agree that trespass law is so easily applicable. In Ticketmaster vs. Tickets.com, the court in its first ruling in 2000 stated that “it is hard to see how entering a publicly available web site could be called a trespass, since all are in invited to enter.” In a later ruling in March 2003 it elaborated [20]:

> “Since the spider does not cause physical injury to the chattel, there must be some evidence that the use or utility of the computer (or computer network) being spiderized is adversely affected by the use of the spider. No such evidence is presented here. This court respectfully disagrees with other district courts finding that mere use of a spider to enter a publicly available web site to gather information, without more, is sufficient to fulfill the harm requirement for trespass to chattels.”

So far there are few cases involving trespass law against crawlers, and courts have responded differently to such claims. To be on the safe side, reverse engineers should refrain from downloading if the Web site owner does not consent to crawling (e.g., by restricting access via the Robots Exclusion Protocol (http://www.robotstxt.org/)).

### 4. Conclusions

This paper has explored legal issues of Web site reverse engineering (WSRE). WSRE can work on three different views of a Web site. Of these, only the client view is of legal concern and hence was further discussed. The RE activities of the client view (i.e., downloading, extraction, analysis, archiving, and reporting) have been identified. Potential legal problems of these activities have been discussed in the context of copyright, contract, and trespass law.

Other laws as the ones we discussed (e.g., patents and trade secrets) might constrain WSRE as well. We did not discuss patent law because in order to obtain a patent, the inventor has to disclose significant details about it to the public. With this assumption, RE is not strictly necessary and as a consequence no RE right exists in patent law. If a patent does not reveal everything, RE activities should not infringe a patent. Conversely, if an inventor does not want to disclose the invention, it can be protected as a trade secret. Using RE to obtain a trade secret is legal if the product has been acquired “by fair and honest means, such as purchase of the item on the open market” [48].

While one can hope that legal problems will seldomly arise for academic researchers, we believe that reverse engineers need to know the constraints that laws place on their work. This is the case for RE of both traditional software and Web sites. As the case of Edward Felten [49, 16] exemplifies, legal issues threaten more and more to constrain academic research [48]. It might be the time for reverse engineers to become more vocal and try to influence future legal developments that negatively affect the legality of RE.

Besides legal issues, other dimensions that should determine RE activities are ethical considerations. Some RE actions might be legal, but not ethical and vice versa [35]. If reverse engineers believe that a Web site indeed constitutes (intellectual) property, ethical behavior requires from them to act accordingly. Spinello, who discusses ownership issues from an ethical perspective, states that

> “the bottom line is if a web site is to be regarded as property with a legitimate owner, that owner

\(^{10}\)Chattel is a legal term for personal property such as books, cars, and computers [46].
has the right to control his or her intellectual product, that is, to set the rules and conditions for how that web site will be accessed and used by others” [50].

It is reasonable to argue that ownership claims should be limited in the case of “serious disutilities” to users of the site. Furthermore, Web site owners have a moral responsibility as well and should respect the “common good of the web.” Whether a certain WSRE activity is ethical is a question that has to be resolved on a case-by-case basis.

While this paper has exclusively addressed legal issues, the role of technology to mediate between Web site owners and users should not be overlooked. The Robots Exclusion Protocol allows site owners to control the extent to which downloading of contents by Web crawlers is permissible. Robots can ignore the protocol, but ethical robots will respect it. The protocol is a de facto standard and is respected by all major search engines. The W3C’s Platform for Privacy Preferences (http://www.w3.org/P3P/) realizes machine-readable privacy policies. Similar schemes could be used by the site owner to communicate to what extent RE activities are permissible. Reverse engineers could then pick sites for their research without fearing litigation. Lastly, contacting the owner of the site and trying to solicit approval should be the first steps of every WSRE effort.

Acknowledgments

Many thanks to the anonymous reviewers for pointing out mistakes and making valuable suggestions. Thanks to Crina-Alexandra Vasiliu for proofreading.

This work has been supported by the Toronto Center for Advanced Studies (CAS), IBM Canada Ltd.

References
