Agent Exclusion on Websites

M.L. Boonk1, D.R.A. de Groot2 A. Oskamp1,
and F.M.T. Brazier2

1 Computer/Law Institute, Faculty of Law, Vrije Universiteit Amsterdam,
de Boelelaan 1105, 1181 HV, Amsterdam, The Netherlands
Phone: +31 - 20 - 598 6215; Fax: +31 - 20 - 598 6230
a.oskamp@rechten.vu.nl
m.boonk@rechten.vu.nl
http://www.rechten.vu.nl/~CLI
2 IIDS Group, Faculty of Sciences, Vrije Universiteit Amsterdam,
de Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
Phone: +31 - 20 - 5987434; Fax: +31 - 20 - 5987653
{frances, davidra}@cs.vu.nl
http://www.iids.org/

Abstract. This paper focuses on legal aspects of intelligent search agents, with respect to the status of exclusion clauses often found on websites. These clauses were initially meant to stop search bots from websites that are not meant for public access. It is a question whether these clauses also hold for intelligent search agents.

1. Introduction

This paper focuses on legal aspects of intelligent search agents, continuing the research that has already been done in the ALIAS-project.[1]
The use of agent technology in the process of searching for information on the Internet could significantly decrease the amount of time it takes users to find relevant information. However, a number of legal issues need to be resolved before we can truly benefit from the advantages of intelligent search agents. For these agents to effectively search the Internet, website access is essential. And it is exactly here that an important issue arises.

In the 1990's, the use of search engines became very popular and many people built their own search bots. These bots, also named "wanderers", "gatherers", "spiders", "crawlers" or "harvesters", traverse the Internet, requesting and processing every available webpage and following every available link on a webpage, usually for indexing purposes. For various reasons, many website owners were not (and are still not) too happy about all these bots visiting their websites.[2] Some bots fire too many get-requests in too little time, which significantly slows down or may even crash a system. Many webpages are unsuitable for bots, e.g. because they require an interactive response or contain dynamic data so the page will have changed before it has been indexed. Machine readable exclusion clauses (further: *no-robots clauses*) are designed to keep search bots from visiting certain areas of the Internet. A bot can be programmed to "read" these no-robots clauses so it "knows" it should not access a particular webpage. Currently, many website administrators have a no-robots clause added to their website.

Also, an increasing number of website administrators add General Terms and Conditions to their websites, which often prohibit the use of bots, agents or all automated means to access their websites.

---

1 Brazier, F.M.T., Oskamp, A., et. al. (2003) "ALIAS: Analysing Legal Implications and Agent Information Systems", p. 5-7 Computer Science, Faculty of Sciences, Vrije Universiteit Amsterdam, http://www.iids.org.
2 By "website owner" we define anyone who is in charge of / responsible for a website.

Intelligent search agents may be equipped with features that enable them to effectively search the Internet. To find information on a particular subject, a search agent may need to search exactly those webpages that are labelled as "unsuitable for bots" by a no-robots clause or have some form of General Terms and Conditions. Whether these no-robots clauses apply to search agents as well and what the presence of General Terms and Conditions on a website entails for search agents, is as yet unclear.

This paper discusses no-robots clauses and General Terms and Conditions that an agent can come across when visiting a website.

First, no-robots clauses are explained and the question as to why they were originally designed is answered.

Second, the reasons for using no-robots clauses are explained, along with the legal status of no-robots clauses.

Third, the behaviour of a search bot and a search agent when accessing a particular website is compared.

Fourth, the implications of General Terms and Conditions for agents are discussed.

Fifth and last, the conclusions that can be drawn from this paper are presented.

2 Definition of agents and bots

As exact definitions of a bot and an intelligent search agent do not exist, we first define both a bot and a search agent. The so-called "Robot Exclusion Standard", a document that proposes a method for excluding bots (the so-called "robots.txt"-file) defines a bot as follows:

> *"A*
>
> *robot is a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced. (…) Normal Web browsers are not robots, because they are operated by a human, and don't automatically retrieve referenced documents (other than online images)."* [3] [4]

An intelligent search agent is *"a computer program which performs tasks (mainly searching) on behalf of another entity, possibly over an extended period of time, without continuous direct supervision or control".* [5] Although there is much discussion on what an intelligent search agent is exactly, researchers do agree that for a process to be called an intelligent search agent, it should have at least the following characteristics: it should be autonomous, proactive, reactive, communicative, have a certain level of intelligence and it may be mobile. [6]

---

[3] http://www.robotstxt.org/wc/faq.html#what ("What is a WWW robot").

[4] It is important to note that the Robot Exclusion Standard was designed in a time when documents were less advanced.
Nowadays, a document can also contain other files than images, e.g. audio and video files.

[5] This definition is based on Krupansky, J.: "What is a Software agent?" in: http://www.agtivity.com/agdef.htm.

[6] See Franklin, S. and Graesser, A. (1996), Is it an Agent or a Program?: A Taxonomy for Autonomous Agents, Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages, Springer-Verlag. See also Wooldridge, M.J. and Jennings, N.R. (1995), Intelligent Agents: Theory and practice. The Knowledge Engineering Review, 10(2):115–152. See also Brazier, F.M.T., Oskamp, A., et. al op.cit.

3 What can an agent come across while searching the web?

A search agent acting as a representative of a human being while searching the web could come across the following "obstacles": no-robots clauses, contractual exclusion in General Terms and Conditions, and technical security measures.

In this paper, we discuss when and how agent access to websites can be restricted. As this paper merely focuses on "non-technical" measures to prevent agents from accessing a website, we mainly discuss the implications of no robot clauses and General Terms and conditions for agents and only make a few remarks on technical security measures.[7]

4 "No robots- clauses"

4.1 What are no-robots clauses?

In this paper, we define a no-robots clause as a "no access" notice that can be "read and understood" by automated means which have been programmed to do so.

In the 1990's, many bots caused inconvenience while trying to access webpages that were unsuitable for them or by firing too many get-requests in too little time. A Dutch programmer, Martijn Koster, came up with a means for website owners to "say no" to wanderers, spiders, web crawlers and other bots that travel the Internet to gather webpages for indexing purposes and thus wanting to access a particular website. He wrote the "Robot Exclusion Standard".[8] This document, which is not an Internet standard, but "only" widespread practice, explains how a robots.txt file can be implemented to prevent bots from causing inconvenience. [9] [10] The Robot Exclusion Standard demands that before a robot requests a particular page in a "web domain", it should first request the robots.txt file in the root of the web server. This file sums up pages or areas within the web domain that are not supposed to be visited by bots. An example of a robots.txt file is the following:

```
User-agent: *

Disallow:
```

The asterisk (*) in the User-agent field refers to "all robots". Since nothing is disallowed, bots are allowed to visit every available web page on the website to which the robots.txt file applies.[11]

An owner of webpages who does not have the rights to specify a robots.txt file, has alternative means to notify bots that certain pages should not be indexed or that the links on the page should not be followed by bots. A so-called "Robots metatag", placed in the HTML <HEAD> section of a page, can specify either or both of these actions. When the appropriate metatags are added to a html page, a bot will not index the page or follow the links on that page. For example, an effective no robots-metatag to instruct a bot not to index, would be:

```
<META name="ROBOTS" content="NOINDEX">.
```
[12]

---

[7] In doing this, we assume that the agent has been programmed to read any machine readable code.

[8] See http://www.robotstxt.org.

[9] See http://www.robotstxt.org/wc/norobots.html (introduction) and http://www.searchtools.com/robots/robots-txt.html (Search Indexing Robots and Robots.txt).

[10] http://www.robotstxt.org/wc/faq.html (What if I can't make a /robots.txt file?).

[11] See http://www.searchtools.com/robots/robots-txt.html.

[12] See http://www.searchtools.com/robots/robots-meta.html.

Note that, as no restrictions have been placed on the option to follow links specified on the site, this implies that the owner has no objections to this use.

4.2 Why implement no-robots clauses?

There could be many reasons for a website owner not to want bots accessing particular webpages or an entire website. First, a webpage could be unsuitable for bots, because it contains (dynamic) content which is of no use to bots, for example pages with forms, or "special offers" that frequently change so the information stored in the search engine's index never corresponds with the actual information on the webpage. In these cases, a no-robots clause often serves as a *guide*, showing bots that particular webpages are not suitable for access by bots.

Second, a website owner may not want his/her website to be accessed by bots because they cause inconvenience. A website may be visited by bots that do too many get-requests in too little time, which may cause the web server to malfunction. A website owner could also add a no-robots clause to his/her website because visiting bots generate too much traffic as they simply request every available webpage and follow every available link on a webpage instead of searching for specific information. In these examples, the no-robots clause functions as an *obstacle* for access to non-humans.

Over the past decade, no-robots clauses have been used quite successfully as guides. All major search engines (including Google, Altavista and Yahoo) adhere to them and every day countless bots are saved from needless visit to pages that are of no avail to them anyway. The system appears to work more or less to everyone's satisfaction.

Unfortunately, the same does not hold for no-robots clauses that are meant to function as "obstacles". Website owners appear to have quite a problem maintaining the obstacle function of no-robots clauses. An important reason for this is that they lack legal status. Although so-called Internet standards exist, these are no more than a set of criteria, voluntary guidelines, and best practices for the Internet. There is no law or treaty obliging anyone (or anything) to adhere to these Internet standards.[13] Not adhering to current best practices on the Internet may lead to interoperability problems, but violating these standards does not have any legal consequences. Moreover, there is no Internet standard that demands adherence to no-robots clauses.[14] [15] [16]

Also, no-robots clauses are not to be regarded as effective technological measures that prevent bots from accessing a webpage. An automated device that has not been programmed to detect and respond to either a robots.txt or a "no robots notice" in metatags and thus ignores a no-robots clause, will not be stopped: it will request the page as if there were no "no-robot" clause and follow the links on that page.[17]

Article 6.1 of the EU Copyright Directive orders member states to protect against circumvention of effective technological security measures. Article 6.3 of this Directive defines technological security measures as follows:

> *"Technological measures shall be deemed 'effective' where the use of a protected work or other subject matter is controlled by the rightholders through application of an access control or protection process, such as encryption, scrambling or other transformation of the work or other subject-matter or a copy control mechanism, which achieves the protection objective."[18]*

---

[13] http://www.jamesshuggins.com/h/oth1/search_engine_disputes.htm,
http://searchenginewatch.com/resources/article.php/2156541,
http://searchenginewatch.com/sereport/article.php/2165861,
http://www.linksandlaw.de/geschichtedeslinking22.html.

[14] Anyone who does not wish to conform to Internet standards as designed by the Internet Community simply risks not being able to communicate with other members of the Internet community. Although there have been several Request for Comments (RFC) to come to an Internet standard for the robots.txt file, as yet this has not resulted in an Internet standard. The W3C does have a standard describing the implementation of a no-robots clause in metatags, but the actual use of no-robots clauses in metatags is not subject to any Internet standard. See ftp://ftp.rfc-editor.org/in-notes/rfc2026.txt, particularly p. 8-9 and ftp://ftp.rfc-editor.org/in-notes/rfc3700.txt. See also Kleve, P., "Juridische Iconen in het Informatietijdperk" , Kluwer 2004, p. 82.

[15] See for RFC's: http://www.rfc-editor.org/.

[16] See for the RFC's for robots.txt: http://www.robotstxt.org/wc/norobots-rfc.txt

[17] Besides, many bots have not been programmed to recognise "no robots-metatags".

A no-robots clause is not an effective measure to control access to a website, because it does not enhance a website owner's control over access to his/her website. A website which only has a no-robots clause can still be accessed by an electronic means which has not been programmed to read and understand a no-robots clause. The given definition of "effective technological security measures" clearly excludes no-robots clauses as "effective technical measures".[19]

The only way to technically prevent bots from visiting a webpage is by protecting a page with "technological security measures", e.g. only allow visitors who can produce credentials that are relevant and necessary for accessing the website.

A website owner who does not want a particular bot visiting his/her website, can request the program's user to stop visiting the website, regardless of whether the website has a no-robots clause. Moreover, the Ebay/Bidder's Edge case shows that ignoring the no-robots clause of a particular website is not enough for assuming trespass to chattels.[20] The website owner must make it plausible that he suffers damages due to the program visiting the website. In the latter case however, the issue is not whether the website has a no-robots clause, but whether the website owner suffers damages because of the visiting program.

4.3 Do no-robots clauses apply to agents?

Although no-robots clauses were originally designed to prevent bots causing too much inconvenience, it has been argued that they should also apply to agents. In General Terms and Conditions of websites, the use of agents is often explicitly prohibited, usually in the same sentence as the prohibition of "bots, spiders, wanderers, etc." In our opinion, agents should not be regarded as "yet another kind of bot", because a well-trained search agent may act in a completely different way. A website owner with a no-robots clause on his/her website may, however, benefit from being visited by a search agent.

Whether search agents should adhere to no-robots clauses does in fact depend on: (a) whether a search agent behaves like a bot and (b) the legal status of no-robots clauses.
   The fact that a search agent can operate autonomously and on behalf of a particular person, entails that it has the ability to operate differently from search bots. As a search agent is capable of actually searching for information, it can be used for finding specific information, e.g. the cheapest airline ticket for a flight to India that can be found today. A bot, on the other hand, can be quite suitable for gathering data e.g. to build an index, but is completely incapable of finding an answer to a given question.
   What do these differences imply for the website owner who wants to avoid certain behaviour on his/her website? A closer look at how agents, bots and humans using a web browser act when visiting a particular website is needed.
   A human using a web browser searches with a specific purpose in mind. He/she accesses the website to search for certain information that he needs at that moment. He/she instructs his browser to collect one webpage, reads it and then decides what to do next. Either he has found the requested information, or he has not, and decides whether to continue searching and where. The human will only look at pages he regards as relevant. The fact that a website is generally designed for access by humans, implies that the actual (proper) use of a website by a human does not inconvenience the website owner.
   A bot does not search with a specific purpose. In fact, it does not search for information, it just gathers all available data on the page for indexing purposes. There is no connection at all between an actual question and the data gathered by a bot. Most bots are programmed to systematically process everything that is on a webpage and follow every link on that page. To gather data, a typical bot starts somewhere on the Internet, simply requesting every available webpage (either depth first or breadth first) and follow every link on that page. Unless explicitly stopped, a bot will not stop until it has processed every available

---

[18] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain
   aspects of copyright and related rights in the information society.
[19] Groom expresses a different opinion on the legal status of no-robots clauses. In his opinion, no-robots clauses are to be considered as "effective technological measures" as explained in article 6.3 of the Copyright Directive. However, his argumentation for this is not convincing: Groom J. Are 'Agent' Exclusion Clauses a Legitimate Application of the EU Database Directive?", (2004) 1:1 *SCRIPT-ed*, @: <http://www.law.ed.ac.uk/ahrb/script-ed/docs/agents.asp>.
[20] http://www.tomwbell.com/NetLaw/Ch06/eBay.htm.

website. It quickly and systematically gathers all available data on a website, thereby potentially overloading a web server. This is one of the reasons why a bot visiting a website can cause much inconvenience for a website owner.

An agent has been given a specific item for which to search, e.g. a specific telephone number or "all available pictures of a dog on a purple bicycle". As an agent searches with a specific purpose in mind, it will often try to find a relevant "starting point" first.

Unlike a bot, an agent will often not request and process every webpage it can find and follow all available links, but it will try and search pages it assumes to be relevant. After all, it accesses the website to search for certain information that it needs at that moment. It will search a page for clues to the requested information. In doing this, it could look at the context of a page. For example, an agent that has been asked to find a telephone number could have a notion of "telephone number" being related to "people" or "researchers" inserted in its ontology.

If the agent has found the requested information, it will stop searching and return its results to its user. If it has not found an answer to the question posed by its user, the agent reasons whether to continue searching and where.

In sum, whereas search agents go and search for the answer to a given question, bots do not search: they only process webpages which are then indexed. An intelligent search agent's behaviour resembles that of an extremely rapid human being executing a search on the Internet and therefore does not inconvenience the website owner. Its behaviour on a website therefore differs substantially from that of a bot accessing the same website.

As for the legal status of no-robots clauses, it has already been mentioned that a no-robots clause itself can not stop an automatic program from accessing a website. no-robots clauses are to be perceived as voluntary guidelines rather than as regulations that can be enforced by law.

5 "Contractual Exclusion" of agents in General Terms and Conditions

A second obstacle for agents accessing a website could be the General Terms and Conditions published on the website. Generally, one is only permitted to access the website when one has agreed to conform to these terms. The obligation to conform to a website's no-robots clause can be included in these terms. For example, the company Health to go- has added the following phrase to its General Terms and Conditions:

> *"You understand that the robots.txt file is the only means by which robots are authorized to access our web site. You agree not to violate any of the robot access policies."* [21]

If a website has General Terms and Conditions, but can be accessed without having read them, it remains unclear how an agent should adhere to them. According to Dutch legislation, one must have been given the possibility to read the General Terms and Conditions in order to be bound by them.

This means that if an agent requests a webpage which is part of a website that has General Terms and Conditions, for these terms to be valid, their existence needs to be made known to an agent or its user. An agent requesting a webpage could, for example, be automatically redirected to the webpage containing the website's General Terms and Conditions. Whether an agent can be expected to "understand" these terms is unclear. Perhaps these terms need to be "translated" into code that can be read and understood by agents to be binding.

As General Terms and Conditions are renown for their nuances or ambiguities, whereas in its essence a computer program cannot understand ambiguity, it seems impossible to translate these terms "as is" into code that can be read and understood by agents. Further research is needed to investigate the possibilities concerning alternative terms.

Additionally, the presence of terms and conditions implies that there is going to be a contract between parties. The question "under which circumstances can an agent conclude a contract?" needs further research as well.

---

[21] http://www.health2go.com/terms.php.

Let us assume the agent can indeed "read and understand" the General Terms and Conditions, then what should it do after processing them? If General Terms only prohibit the use of *bots*, it could be argued that the use of *agents* is allowed.

If the General Terms explicitly prohibit the use of agents, agents should not access the website: Even if the agent's user has been given the necessary credentials to access the website, the mere fact that he has agreed to the General Terms of Use implies that his personal agent cannot access the website.

In

practice, General Terms hardly ever prohibit the use of "bots" or "agents" as such. They either prohibit the use of any automated means, or use such vague notions that it remains unclear whether they allow agents. Some examples are:

USF Corporation:

> *"(…)You agree that you will not use any robot, spider, or other such programmatic or automatic device, including but not limited to automated dial-in or inquiry devices, to obtain information from this General Website or otherwise monitor or copy the Materials (…)."* [22]

Reliance India Call:

> *"When using this Website, you expressly undertake, represent and covenant that you shall not: (…) Use any robot, spider, scraper, site search/retrieval application, or other manual or automatic device or process to retrieve, index, "data mine," or in any way reproduce or circumvent the navigational structure or presentation of the Website or the Information provided on this Website, without our written authorization; or (….)"* [23] [24]

Lycos:

> *"You agree that you will not use Lycos Network Products and Services to: (…) Use automated means, including spiders, robots, crawlers, or the like to download data from any Lycos Network database."* [25]

If the use of *any automated means* is prohibited by General Terms, it seems obvious that the use of agents on that particular website is not allowed. However, the exclusion of *any automated means* suggests that the use of a web browser (which is an indispensable tool for human users accessing the Internet) is prohibited as well!

If one has agreed to the General Terms and Conditions of a website which oblige adherence to a no-robots clause and one then does not adhere to this particular no-robots clause, one could be held liable for not adhering to these terms. However, if the General Terms and Conditions do not explicitly prohibit agent access to a website, it remains unclear whether agents are allowed to access a particular website or not.

Further research is needed to resolve this issue, both with regard to the development of General Terms and Conditions that can be accessed and processed by agents and the design of General Terms and Conditions that are unambiguous enough to be understood by agents.

---

[22] http://www.usfc.com/common/termsofuse.jsp.

[23] http://www.relianceindiacall.com/US/termsofuse.asp.

[24] Other examples: http://www.marqueedomains.com/privacy.html : (Prohibited Conduct You agree that you will not use MarqueeDomains.com Network Products and Services to: (…) Use automated means, including spiders, robots, crawlers, or the like to download data from any MarqueeDomains.com Network database"), see http://www.idxmanager.com/policy.asp: "Prohibited Conduct (…) Use automated means, including spiders, robots, crawlers, or the like to download data from any IDX Network database."

[25] http://info.lycos.com/legal/legal.asp.

6 Concluding remarks

Whether agents should adhere to no-robots clauses depends on the reason why a no-robots clause has been added to a website. If a no-robots clause is merely added as a guide to notify bots that certain pages are unsuitable for them, this does not imply that those pages would also be unsuitable for search agents. On the contrary, the information on those webpages may be particularly suitable for search agents to visit. The search agent acts as a representative of a human (who would normally use a web browser) and its behaviour on an actual website resembles that of a human using a web browser, rather than that of a bot. Since no-robots clauses do not apply to humans using a web browser (which is an automated device), it would be irrational for agents to adhere to no-robots clauses. When the no-robots clause functions as a guide, there will be no need for agents to adhere.

If a no-robots clause is used as an obstacle, there is no legal obligation to adhere to the no-robots clause itself, but adherence to a no-robots clause could be enforced as part of General Terms and Conditions of a website. However, this does not imply that non-adherence to a no-robots clause itself would be legally enforceable.

For General Terms and Conditions to be legally binding, certain conditions need to be fulfilled, for example the agent or its user needs to be aware of their presence. For that reason, the General Terms and Conditions should be added to a website in code that can be read and understood by agents or accompanied by effective security measures. Whether an agent is allowed to access a website will thus depend on what is actually said in General Terms and Conditions.

A website owner could take away many uncertainties on this matter by protecting his/her website with technical security measures. The website owner could require credentials to access a website and only grant credentials to people that respect the General Terms and conditions of that website, which oblige adherence to the website's "no-robots clause."[26]. Circumvention of effective technical measures is prohibited in all European Union countries.[27] Technical security measures are also extremely effective to ensure that a website's General Terms and Conditions are read and agreed upon in advance.

Acknowledgements

---

[26] This becomes clear when subscribing to news sites, e.g. http://www.volkskrant.nl, or recht.nl (http://www.recht.nl). For American jurisprudence: see http://www.tomwbell.com/NetLaw/Ch06/eBay.htm (Ebay/Bidder's edge), http://searchenginewatch.com/sereport/article.php/2165861and and http://www.linksandlaw.the/geschichtedeslinking- 22.html (both: Ticketmaster/Microsoft). For England, see e.g. http://www.law.gwu.edu/facweb/claw/linking.htm and Verbiest, Th. « Entre Bonnes et mauvaises références, à propos des outils de recherche sur Internet » http://www.club-Internet.fr/cyberlexnet/COM/A990225.htm (both concerning Shetland Times/Shetland News), Verbiest, Th. « Journalisme et droit d'auteur en Belgique : http://www.robic.ca/cpi/Cahiers/12-2/VerbiestThierry.html (News Index/Sunday Times) See for a survey of jurisprudence on this matter Verbiest, Th., « La responsabilité des outils de recherche sur Internet en droit français et droit belge » : http://www.juriscom.net/pro/1/resp19990430.htm.

[27] See for a survey on technical security measures: Koelman, K.J., "Auteursrecht en Technische Voorzieningen" Amsterdam, 2003, p. 7-140.