The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

Received 14 October 2008
Revised 3 November 2008
Accepted 12 January 2009

Using the Robots.txt and Robots Meta tags to implement online copyright and a related amendment

Chyan Yang
Institute of Business and Management and Institute of Information Management, National Chiao Tung University, Taipei, Taiwan, Republic of China, and
Hsien-Jyh Liao
Taipei Prosecutor Office, Taipei, Taiwan, Republic of China

Abstract

Purpose – The Robots.txt and Robots Meta tags constitute a set of instruments that can be used to instruct software robots. However, the current version of Robots.txt and Robots Meta tags are both too simple and ambiguous in an internet world with many potential conflicts, especially in terms of copyright and trespass to chattels. This paper seeks to propose an amendment to the Robots.txt and Robots Meta tags to solve the problems.

Design/methodology/approach – Instead of following personal experience, this paper surveys several predominant cases in an attempt to find general principles that can be used as guidelines to amend the Robots.txt and Robots Meta tags.

Findings – According to several court cases, the Robots.txt and Robots Meta tags can not only be used to simply allow or refuse the software robots, but also expressing the online copyright authorization policies of webmasters. Any robot following the given policies can prevent possible conflicts, and undoubtedly, any robot ignoring these may be in breach of the law. In terms of adapting to their new roles successfully, the Robots.txt and Robots Meta tags need some supplements and adoption; as a result, the webmasters can express their will more explicitly and avoid unnecessary disputes in relation to copyright authorization scope and trespass to chattels as well.

Originality/value – This paper reveals the new function of the Robots.txt and Robots Meta tags. Based on this new function, this paper points out the disadvantages of the current Robots.txt and Robots Meta tags and proposes new a comprehensive amendment based on this new function.

Keywords Systems software, Tagging, Copyright law, Search engines

Paper type Research paper

1. Introduction

The dramatic increase in web sites, to around 172,338,726 by June 2008 (Netcraft, 2008), make searching and accessing specific materials harder and almost impossible without any aid. The software robot is one of the most wide spread tools used for such data searching services. For example, search engines like Google (Google, 2008a), Yahoo (Yahoo, 2008a), MSN (MSN, 2008a), etc., all use software robots to collect data on the internet and then provide that information to the public. Whilst having obvious benefits, the software robots do have some underlying disadvantages: the unlimited access from robots may overload the Web server and the re-use of the data collected by the robots may result in copyright and other legal disputes. To diminish such possible

Library Hi Tech
Vol. 28 No. 1, 2010
pp. 94-106

0737-8831
DOI 10.1108/07378831011026715

conflicts, the Robots.txt and its supplement, Robots Meta tags, are the most common and straightforward instruments that can be used by webmasters to exclude unwelcome robots or, as some recent legal cases suggested, grant a license in terms of digital copyright to permit legitimate access from specific robots.

Even though Robots.txt and Robots Meta tags are taking on more significant roles today, they have not been fully investigated by researchers. Only a few peer reviewed academic papers in relation to this topic have been released (Chau and Chen, 2003) and, as a result, sporadic amendment proposals are based on personal experience rather than general principles (Conner, 1996; Koster, 1994). In fact, with the current popularity of Robots.txt and Robots Meta tags, it is time for a wide-ranging review of the Robot.txt and Robots Meta tags and to recommend a comprehensive mechanism to regulate the relationship between robots and webmasters.

At the beginning of this paper, we will turn our attention to the software robot, with the Robot.txt and Robots Meta tags being viewed in Section 2 and Section 3. Before calling attention to the functions of the Robot.txt and Robots Meta tags, in Section 4, we raise some issues related to the software robots and the webmasters. In the following sections, we will review the original and newly developed functions – expressing online copyright authorization – of the Robot.txt and Robots Meta tags and reveal some uncertainties and a few disadvantages that have arisen. Next, in order to help the Robot.txt and Robots Meta tags meet their new roles, we propose an amendment before finally discussing some unsolved problems whilst suggesting additional issues that invite future research.

2. Introduction of the software robot

2.1 Software robots

A software robot, also called a spider, crawler, web robot, web agent, webbot, wanderer, and worm, can be defined as a software program issued by its user that traverses the web to collect data in compliance with standard HTTP protocol (Cheong, 1996). In the beginning of the process, a software robot will follow the initial URLs provided by user to retrieve the web pages. After parsing these collected pages, the robot will obtain more URLs and it can access to more pages consequently. Repeating this process over and over, a software robot will, theoretically, find most of the pages on the web. Software robots have been shown to be useful in various Web applications. There are four main areas where robots have been widely used (Chau and Chen, 2003). The first is "Building collections": software robots have been extensively used to access and collect data of web sites that are required to create an index for application programs, such as search engines. The second use is "Archiving": a few projects, like Citeseer.ist (1997), have tried to archive academic papers with regard to computer science from across the whole web. The third is "Personal search": a personal robot tries to search for web sites of interest to a particular user. The final use is for "web statistics": the large number of pages collected by robots is often used to provide useful, interesting statistics about the web, including the total number of distinct web sites on the web (Netcraft, 2008), the average size of a HTML document, etc. The complete process of how a robot collects data from the internet is shown in Figure 1.

In Figure 1, the first step involved is "accessing", where the robot users use their robots to collect data. Step two is "processing", where the robot offers the collected data for further processing, such as indexing, analysis, etc. As well as these two steps, some robot users, such as search engines or online archives, may provide the processed data to other online viewers in a last "distributing" step, but the last step is optional.

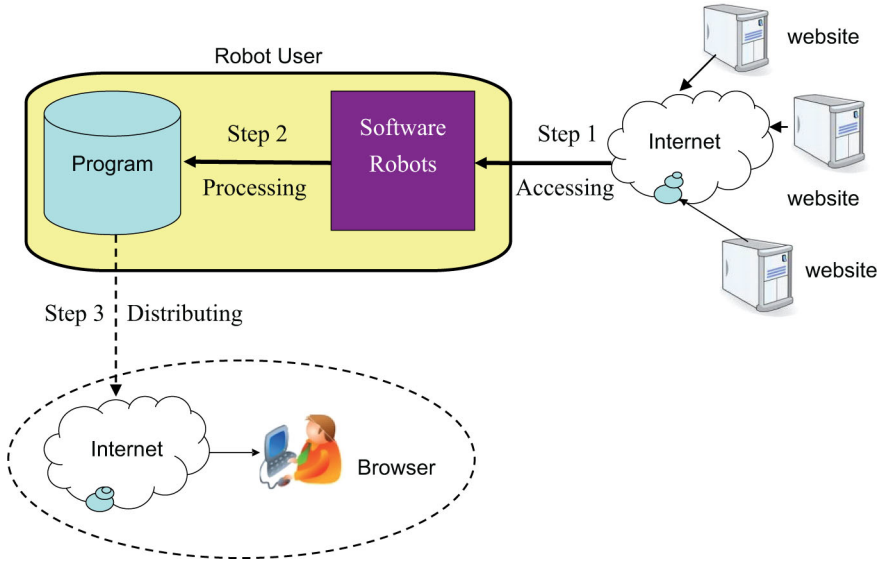

Figure 1.
The process of how a software robot works

2.2 Technical measures to exclude software robots

Software robots are not always welcomed by webmasters because some robots may disturb daily operations of web sites or divulge confidential information (Snyder and Rosenbaum, 1998). That is why webmasters may take some measures to exclude software robots in some cases. According to their effectiveness, two types of technical measures, active and passive, could be introduced to exclude unwanted software robots and, consequently, to eliminate any possible conflicts that can arise:

(1) *Active measure*. Webmasters might use a program which can distinguish robots from human viewers to prohibit access from "robots" (Tan and Kumar, 2002). Although this kind of program is helpful, its major flaw is, no matter how complicated it is, no one can guarantee that it can successfully detect any robot. That is to say, some "smart" or "disguised" robots may circumvent it and gain access to the Web pages, though it may violate some legal rules (Bainbridge, 2008). Other web sites may be password protected, but this strategy may seriously impede public access.

(2) *Passive measures*. Contrary to the costly active measures, some webmasters may try to adopt human-readable terms, like a "No access from Robots" term embedded in the pages, or machine-readable codes, such as Robots.txt in the server's root directory, in order to exclude robots. Nevertheless, even the human-readable terms may be readable to human viewers, it is almost impossible to correctly and explicitly interpret varieties of human-readable terms without human intervention (Feigin, 2004); that is to say, human-readable terms are incomprehensible to general software robots. Therefore, the Robots.txt and Robots Meta tags become the most important tools for use in robots exclusion mechanisms as explained in detail in the next section.

3. Introduction of the Robot.txt and Robots Meta tags

The Robots.txt and Robots Meta tags were both proposed in the 1990s. The Robots.txt is also called the "Robots Exclusion Protocol" (Snyder and Rosenbaum, 1998), "Robot Exclusion Standard" (Koster, 1994) or "Standard for Robot Exclusion"(Koster, 1994), though it is only a widely accepted convention consented by members of a robot mailing list (Koster, 1994), rather than an official standard with necessary official recognition (Feigin, 2004). Even so, most wide spread search engines, Google (Google, 2008b), Yahoo (Yahoo, 2008b), and MSN (MSN, 2008a) all support and fully recognize the original Robots.txt and Robots Meta tags; moreover, both Yahoo (Yahoo, 2008c) and MSN (MSN, 2008b) have tried to introduce some amendments to them. As far as web sites are concerned, research indicates that, in 2001, around 40 per cent of the web sites owned by the global high-rank companies adopted the Robots.txt and Robots Meta tags (Drott, 2002).

3.1 The robots.txt

The Robots.txt is a file that should reside in the root directory and must be named "robots.txt". A robots.txt file located in a subdirectory or named as something else is invalid, as software robots only check for this file in the root (Koster, 1994). The examples in Table I illustrate several common uses of the Robots.txt:

3.2 Robots Meta tags

Sometimes, the page creators do not administer their own web sites. For example, a staff member in a university creates his personal webpage on the web site of his department. In this circumstance, it is someone who works in the computer center of the university that is the webmaster having the authority to access the root; the staff member is neither able to access the root directory nor use the Robots.txt to exclude software robots. This disadvantage has been reduced by the use of Robots Meta tags: the "[No]index" tag and "[No]follow" tag, which should be within the page codes (Koster, 1997). Some examples are as shown in Table II.

In case the page creator has the right of access to the root directory, he can adopt the single "Disallow" directive to exclude robots, instead of exhaustively embedding redundant "Noindex" tags in all pages hosted in the same server.

	Examples	Meaning
1	User-agent: * Disallow:	Allow all robots complete access
2	User-agent: * Disallow:/	Exclude all robots from accessing the entire server
3	User-agent: lycra Disallow: User-agent: * Disallow:/	Only exclude the access from the robot called "lycra"
4	User-agent: * Disallow: /tmp Disallow:/log	Exclude all robots from the /tmp and the /log folder.

Table I.
Examples about Robots Meta tags

Table II.
Examples about Robots Meta tags

	Examples	Meaning
1	< Meta Name = "MY_ROBOTS" content = "noindex" >	Restrict the software robot called "MY_ROBOT" from indexing a page
2	< Meta Name = "ROBOTS" content = "noindex" >	Restrict the all robots from indexing a page
3	< Meta Name = "MY_ROBOTS" content = "nofollow" >	Restrict MY_ROBOT following links on a page
4	< Meta Name = "ROBOTS" content = "noindex,nofollow" >	Block all robots from both indexing and following links

4. Potential conflicts that arise due to software robots

4.1 The conflict in relation to data collected by robots – online copyright

According to the general principle of copyright laws, such as 17 USC 106, the UK CDP Act 1988, the creators of a copyrightable work automatically own the copyright of the works upon completion; and no one can reproduce, modify or distribute such works without the owners' consent (Rao, 2003; Spinello, 2007). However, the internet boom has given rise to certain ambiguities and contradictions within the internet copyright law. For example, when a web site owner makes his site available to the public via the internet, any person who views that site may come to the conclusion that some common copyright laws do not apply as information is made available for reading and temporarily copying by any person visiting that site (Reed, 2004). Consequently, some thoughtful viewers may have doubts about permanent copying, redistribution and modification of such works, since some authors may authorize the online viewers to read, reproduce and modify works, but others may not and, some works may be protected by copyright regime and some may be not (Seadle, 2006). However, even for experienced information professionals or attorneys, figuring out what is and is not allowed is difficult and time-consuming.

This problem becomes more complex after the emergence of online digital libraries (Citeseer.ist, 1997), internet archives (Blake, 2004) and even search engines (Gorman, 2006), which all use software robots to collect copyright documents, images and other information objects available on publicly accessible web sites. Up until now, few software robots are able to distinguish between a fully authorized work and a work posted only for limited use. This means that most software robots simply collect all works without any reference to copyright ownership. In this condition, when viewers search online libraries or archives for works pertinent to their areas of research and, download one or more works as required even without any definite consent or permission, there may arise some serious legal concerns. These include that redistributing any copyright work without the rights owner's consent may infringe the distribution rights of the owner and, under some special circumstances, that normal access to a web site may infringe the reproduction rights of the right owner. In Sections 5, 6 and 7, we will explain when these special circumstances occur and how to cope with such problems.

4.2 The conflict in relation to the various ways robots collect data – trespass to chattels

The load of a server is one of the main concerns of its administrator; with too many requests causing possible overload of the server, there are inevitably some conflicts.

(Koster, 1993) This situation can be called "denial of service", by analogy to the denial of service attacks that are sometimes the effect of computer viruses (Thelwall and Stuart, 2006). The leading case related to this issue is a US case: *eBay, Inc.* v. *Bidder's Edge, Inc.*[1]. The defendant, Bidder's Edge, Inc, was an aggressive web site whose main work was gathering prices and all other important information items for sale on auction web sites. Customers who wanted to bid on an item could visit this site to find out the best price instead of surfing the whole internet themselves. Even though eBay employed a robots.txt file in its server to try to prohibit such unwelcome robots activity, the defendant, Bidder's Edge, still used its software robot to collect information. Bidder's Edge, which issued a great deal of requests every day, actually interfered with the daily operation of the plaintiff's servers. On this ground, eBay launched a lawsuit against Bidder's Edge. The US district court finally announced a verdict stating that the conduct of the defendant constituted a common law tort, the trespass to chattels, because eBay had wasted bandwidth and processing power when dealing with the unsolicited intrusion from the defendant. Since then, numerous successful lawsuits have relied upon this verdict, trespass to chattels, to challenge unwelcome access to servers (Samuelson, 2003). In light of this US case, it has been mentioned that not only the common law countries which share the same heritage of trespass to chattels doctrine can reference this case (Adam, 2002), but also the courts of other countries should pay heed to the principle of this case in digital environments (Wong, 2007).

5. Two functions of the Robots.txt and Robots Meta tags

5.1 The original function: voluntary advice

The original idea of the Robot.txt and Robots Meta tags is to offer a common facility provided by the majority of robot authors to the internet community to protect web sites against unwanted access from their robots (Koster, 1994). They are not "enforced by anybody and no guarantee that all current and future robots will use them" (Koster, 1994). In other words, in respect of this design concept, the Robot.txt and Robots Meta tags are only a voluntary code; no one will be punished for breaching the access policy.

5.2 The new function: expressing online copyright authorization

Apart from mere advice, based on a recent noticeable US federal case, *Field* v. *Google, Inc.*[2], the Robot.txt and Robots Meta tags have both found their new roles. This case related to the "Cached link" of Google. In order to allow access when the original page is temporarily inaccessible, or allow viewers to compare changes made to pages during a specific period, Google's search results always includes a link to its own cached copy, which is a temporary repository consisting all source codes of indexed web sites[2]. The plaintiff, Mr Field, who posted 51 copyright works on his web site and "created a robots.txt file for his site, and set the permissions ... to allow all robots to visit and index all of the pages on the site"[2] and, with the knowledge of using Robots Meta tags could "instruct Google not to provide Cached link to a given Web page", Mr Field consciously decided to use none of them[2]. As a predictable result, Google routinely used its software robot, GoogleBot (Google, 2008a), to retrieve the plaintiff's web site, indexed his works and provided the Cache link as well as the search results. Based on these facts, Mr Field "alleges that Google directly infringed his copyright when a Google user clicked on the Cached link to the Web pages containing Field's copyrighted works and downloaded a copy of those pages from Google's

computers"[2]. After taking into account the fact that Mr Field did not take any measure, even though he had the opportunity and ability to employ the Robots.txt and Robots Meta tags to exclude any possible software robots or to instruct the search engine to not provide the Cached link, the federal district court in Nevada held, since Mr Field "knows the use" and "encourage it", that he has granted an implied license to Google according to his conscious silence (Sieman, 2007). As a result, Google did not infringe Mr Field's copyright at all[2].

It is notice that the court in this case suggested that the license from absence of the Robots Meta tags based on two facts: the first one is that, based on the fact that the defendant actually set the Robots.txt, accordingly, Mr Field, had fully ability and opportunity to employ the tags to prevent Google and, a more important one, Google will stop indexing the web sites in terms of the tags employed by the webmasters (Google, 2008a). That is to say, without the above two conditions, a mere absence of the tags could not directly induce an implied license. On this ground, in a recent Belgian case, *Copiepresse* v. *Google*[3], the court found that the newspaper publishers' failure to use standard technical exclusion methods such as the "Robots.txt" and Robots Meta tags did not amount to an implied license (Smith, 2008).

No matter the absence of the tags can be seen as a implied license, according to the forthcoming cases, we can make a conclusion that, although the original idea of Robot.txt and Robots Meta tags was to set up a code of voluntary advice, based on these verdicts, it is quite clear that the Robots.txt and Robots Meta tags have been far from the "voluntary recommendations without any enforcement"; and they have their new roles in the context of law. A webmaster who adopts the Robot.txt or Robots Meta tags to set permissions to allow robots to visit should absolutely be regarded as granting a license to robots, on the other hand, a webmaster who adopts the "Disallow" directive or the "Noindex" tag should be regarded as expressing his explicit will to exclude the robots; in addition, a webmaster who "consciously" does not use them may also be regarded as granting "a implied license" to such robots. As a result, any software robot which follows the license to gain access to the web site or index the collected data does not infringe any webmaster's copyright and, any robot which disregards the "Disallow" directive or "Noindex" tag but still accesses the web site may breach the copyright law in terms of this new function. To sum up, the appearance or absence of the Robots.txt and Robots Meta tags represents the webmasters' wishes; any robots deliberately ignoring these wishes may be in breach of the law. That is to say, the court in this case considered the Robots.txt and Robots Meta tags as instruments which can be used by the webmasters to express their will about what kind of robots are allowed, what are excluded and what kind of links should not be followed.

6. A few deficiencies of the present Robots.txt and Robots Meta tags in respect of potential conflicts and the new function

6.1 Some uncertainties with respect to the new authorization function

This new function of the Robots.txt and Robots Meta tags has transferred them from ethnic advice to a set of powerful tools; the webmaster can rely on these tools to obtain a more secure guarantee. On the other hand, even this new function conferred by case law is so imprecise that there are a few uncertainties that need to be clarified.

6.1.1 About "[No]index" tag. The first point that should be noticed here is that the "[No]index" tag may give rise to some misunderstanding: as we have seen in section 2. Software robots are used in many different areas; some may only use robots to

maintain links instead of making an index of the collected data. Therefore, a "[No]index" tag may cause some doubts as to what the webmaster's real wish is. Does he want to exclude all robots or just exclude the robots used by search engines? Since the tag may lead to legal consequences, we believe that it is safer to explicitly explain will expressed within the "Noindex" is only for excluding search engines' robots or, more broadly speaking, excluding all robots with further indexing possibility. In other words, the absence of a "Noindex" tag can not definitely result in a conclusion that the webmaster grants a license to "all" robots because the "Noindex" tag does not explicitly represent the copyright owner's wish in this situation.

6.1.2 About "[No]follow" tag. When the page containing a "[No]follow" tag and the pages followed by links are owned by the same person, any robot that disobeys this tag and copy the next page may infringement the copyrights of the page owner. However, sometimes, these two pages are not owned by the same person; on such an occasion, any robot that ignores the "Nofollow" tag and follows the link to access to other pages may not violate the copyright law, especially when the owner of next page dose not explicitly exclude software robots by employing any tag: because tags employed by any other but the copyright owner of the page are meaningless.

6.2 No appropriate tags to cover all copyright rights possibly infringed by the software robots

In section 2, we have demonstrated that the complete access process of a robot can be divided into three steps: the accessing, the processing and the distributing step. In each step, the robots or the robot users could infringe the webmaster's copyright without proper authorization. As we have seen above, the Robots.txt and Robots Meta tags are the best potential tools to be used for such a purpose as they are simple and widespread. In terms of the scope of authorization, it is the rights holders who have the right to decide the scope. But as we have shown in Section 3.1, in case the page creators or the right holders do not have their own servers, the Robots.txt can not represent the real wish of the rights holders since the right holders have no right of access to the root directory. From this perspective, the Robots Meta tags, to speak more specifically, the "[No]index" tag, is the only tag which can be adopted to represent the scope of the authorization, as another tag of the Robots Meta tags, the "[No]follow" tag, is useless in terms of authorization as we mentioned in section 6.1.2.

Nevertheless, unfortunately, in terms of the original meaning of "[No]index" tag, it can only be used to exclude software robots with further indexing possibility, rather than excluding all types of robots (Koster, 1997) and, furthermore, it can not cover all three steps the software robots involving and all copyright rights possibly infringed in these three steps.

The rights referred in all three steps are different, as in Table III.

It should be noticed that, in respect of the "accessing" step, that reproduction of software robots does not infringe the reproduction right of the right owners. The first and most obvious reason of this conclusion is: the contents of the web site, at least in most circumstances, are authorized all viewers, including the software robots, on the internet to access; the accessing here inevitably reproducing the contents of the web site into the memory or the disk of the viewers' computers and, as a result, the reproduction is lawful. Even though in some limited circumstances, the right owners try to exclude some viewers and software robots, the limitations and exceptions appearing in copyright laws still form possible executions of viewers and software robots (Sterling, 2003). The last reason is that, even the software robots are excluded by

the "[No]index" tag, a robot would have to access at least part of a page before reading the instruction not to access it, and many robots probably download entire pages before processing any instructions contained in them. Accordingly, it is unreasonable to allege that the software robots infringe the owner's copyright in this step. To sum up, with regard to the copyright authorization of software robots, the reproduction, adaptation rights in the "processing" step and the distribution rights in the "distributing" step are the rights we should concern.,

However, the "[No]index" tag, which is only mapped to the second "processing" step, at most can be used to express authors' will in respect of reproduction and nothing of adaptation. In addition, as for distribution of the work in the third "distributing" step, the will expressed in this tag is ineffectual in resolving the potential infringements resulted from ambiguous authorization scope.

6.3 No directives about access time in the Robots.txt to prevent potential trespass to chattels disputes

Since the Robots.txt can be used to express a webmaster's will, on the same grounds, it is reasonable to extend this new function further by using the Robots.txt to express the webmaster's preference, especially for what types of access are preferable. For example, how many times per day are suitable? When is the best time of access during a whole day, etc? However, the present version of the Robots.txt lacks any directive to instruct software robots to avoid possible torts.

7. A new comprehensive amendment to the Robots.txt and Robots Meta tags

We have considered several insufficiencies and ambiguities of the current Robots.txt and Robots Meta tags. These drawbacks make us believe the current Robots.txt and Robots Meta tags require some adjustments. However, although more and more advice about further amendments to the Robots.txt and Robots Meta tags has been emerging (Barone, 2007; Conner, 1996; Koster, 1996) and each with a different perspective and reason, they are almost all based on personal experience rather than on any general principle. Instead, in this section, we will try to amend the Robots.txt and Robots Meta tags based on their new function; expressing the webmaster's will. Our goal is to attempt to make them become a more reliable set of tools which can be used to specifically express the webmaster's will and to resolve potential conflicts. On this ground, we suggest that the new version of the Robots.txt and Robots Meta tags should at least embrace the following topics.

Table III.
Possible copyright infringement caused by robots and the tag

	Step	Possible copyright infringement
1	Accessing	None
2	Processing	Infringement of the reproduction right, since the crawler user always need to store the data Infringement of the adaptation right since the crawler user may modify the work
3	Distributing	Infringement of the distribution right since the distribution may be unauthorized

7.1 Adding new tags to fully express authorization scope and dismiss the ambiguous "[No]index" and "[No]follow" tags

In order to avoid unnecessary ambiguities about the scope of authorization, it would be useful to have a set of tools which could be used to present the webmasters' explicit wish about authorization. Based on the above discussion, it is quite clear that the current Robot Meta tags are insufficient in terms of authorization. To improve all the disadvantages, we recommend two new tags as follows. The tags mapping to the "distributing" step is a totally new tag and, with regard to "processing" step, a more general "[No]process" tag replaces the "[No]index" tag in the old version. The types of copyright which are covered by these two tags, to speak more specifically, the reproduction right, the adaptation right and the distribution rights, are all copyright rights of authors which could be infringed in the context of internet, especially in respect of crawler access (Sterling, 2003). The two new tags are as shown in Table IV.

The ways in which how these two new tags are used is similar to the [No]index and [No]follow. Some examples are shown in Table V.

7.2 Adding some new directives about access time in the Robots.txt to prevent potential trespass to chattels disputes

In the light of the eBay case explained previously, an instrument of this kind would be very helpful as the robots could rely on the preference given by the webmaster via this instrument to avoid possible trespass to chattels disputes. As we mentioned in section 4, the victim of the trespass by the software robot is the server, therefore, only the Robots.txt employed by the server administrator, rather than Robots Meta tags embedded in the page codes, can be used for this purpose. This requires further revision, for instance, adding a "crawler-delay" directive to delay the robots. In fact, the two important search engines, Yahoo (Yahoo, 2008d) and MSN (MSN, 2008b), have both declared some extensions with regard to this very topic. For example, both of them recommend webmaster's use the "crawler-delay" directive to reduce the number of requests made by robots. In addition, the "access-time" directive, which can also be used to request robots only accessing to the pages residing this server during the given period, has also been recommended (Conner, 1996).

Table IV. Two suggested new tags

	Steps	Tags	Meanings
1	Processing	[No]process	Allow or block any further processing
2	Distributing	[No]distribute	Allow or block any further redistributing

Table V. Examples of three new Robots Met tags

	Examples	Meaning
1	< Meta Name = "MY_ROBOTS" content = "nodistribute" >	Restrict the software robot called "MY_ROBOT" from distributing a page
2	< Meta Name = "MY_ROBOTS" content = "noprocess" >	Restrict the robot user who uses the robot called "MY_ROBOT" from processing data from this page
3	< Meta Name = "ROBOTS" content = "nodistribute,noprocess" >	Block all robot users from both processing and redistributing this page

8. Conclusion and future work

With the rapid expansion of the internet, software robots will inevitably become one of the most necessary and useful tools in organizing the tangled internet world. How helpful these robots are to the search engines and online archives is a good example in justifying this prediction. Accompanying this trend, a vital problem is how to effectively and lawfully use the robots to complete their tasks. Compared to other measures, the Robots.txt and Robots Meta tags are the most commonly tools to help robots and webmasters to cooperate with each other to achieve this goal. Furthermore, based on the cases, the Robots.txt and Robots Meta tags have been gradually evolving from merely voluntary advice to a set of potentially enforceable instruments which can be used to express a webmasters' will and preference. In respect of this new function and some necessarily clarified uncertainties, a new version of the Robots.txt and Robots Meta tags proposed in this paper will definitely play a more important role in the future internet world, since they can take more serious responsibilities in the resolution of future disputes.

Keeping all the advantages and features of the Robots.txt and Robots Meta tags in mind, in the future, combining Robots.txt, Robots Meta tags and Creative Common license (Creative Commons, 2008) and other online licenses altogether to form a more powerful tool of implementation of online copyright which can not only used by webmasters, but can also by each individual author of copyright works within the web site, such as the authors of video files on Youtube (Youtube, 2008) to express their complicated authorization scopes may form a new research direction. Apart from the copyright and trespass to chattels related subjects mentioned in this paper, the great power of software robots now also arise some concerns about revealing personal privacy and data protection on the internet (Thelwall and Stuart, 2006). Since the robots.txt and the Meta tags are only tools specifically designed for software robots, this tool may, hopefully, constitute a new regime in dealing with privacy and data protection issues.

Notes

1. *eBay Inc.* v. *Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058 (ND Cal. 2000).
2. *Field* v. *Google, Inc.*, 412 F. Supp. 2d 1106 (D. Nev. 2006), available at: http://w2.eff.org/IP/blake_v_google/ google_nevada_order.pdf (accessed March 11, 2008).
3. *Copiepresse* v. *Google, Inc.*, No. 06/10.928/C (February 2, 2007).

References

Adam, J.N. (2002), "Trespass in a digital environment", *The Intellectual Property Quarterly*, Vol. 6 No. 1, pp. 1-17.

Bainbridge, D.I. (2008), *Introduction to Information Technology Law*, Pearson Education, Harlow.

Blake, M. (2004), "Archive of Web sites", *The Electronic Library*, Vol. 22 No. 5, pp. 462-6.

Barone, L. (2007), "Robots.txt summit", available at www.bruceclay.com/blog/archives/2007/04/robotstxt_summi.html (accessed March 11, 2008).

Chau, M. and Chen, H. (2003), "Personalized and focused web spiders", in Zhong, N., Liu, J. and Yao, Y. (Eds), *Web Intelligence*, Springer-Verlag, New York, NY, pp. 197-217.

Cheong, F.C. (1996), *Internet Agents: Spiders, Wanderers, Brokers, and Bots*, New Riders Publishing, Indianapolis, IN.

Citeseer.ist (1997), available at: http://citeseer.ist.psu.edu/ (accessed March 11, 2008).

Conner, S. (1996), "An extended standard for robot exclusion", available at: www.conman.org/people/spc/robots2.html (accessed March 11, 2008).

Creative Commons (2008), available at: http://creativecommons.org/license/?lang=en (accessed March 11, 2008).

Drott, M.C. (2002), "Indexing aids at corporate websites: the use of robots.txt and META tags", *Information Processing and Management*, Vol. 38 No. 2, pp. 209-19.

Feigin, E.J. (2004), "Architecture of consent: internet protocols and their legal implications", *Stanford Law Review*, February, pp. 901-42.

Google (2008a), "How Google crawls my site?", available at: www.google.com/support/webmasters/bin/topic.py?topic=8843 (accessed March 11, 2008).

Google (2008b), "Preventing content from appearing in Google search results", available at: www.google.com/support/webmasters/bin/topic.py?topic=8459 (accessed March 11, 2008).

Gorman, G.E. (2006), "Giving way to Google", *Online Information Review*, Vol. 30 No. 2, pp. 97-9.

Koster, M. (1993), "Guidelines for robot writers", available at: www.robotstxt.org/guidelines.html (accessed March 11, 2008).

Koster, M. (1994), "A standard for robot exclusion", available at: www.robotstxt.org/orig.html (accessed March 11, 2008).

Koster, M. (1996), "Evaluation of the standard for robots exclusion", available at: www.robotstxt.org/wc/eval.html (accessed March 11, 2008).

Koster, M. (1997), "HTML author's guide to the robots meta tags", available at: www.robotstxt.org/wc/meta-user.html (accessed March 11, 2008).

MSN (2008a), "Control which pages of your website are indexed", available at: http://search.msn.com/docs/siteowner.aspx?t=SEARCH_WEBMASTER_REF_RestrictAccessToSite.htm (accessed March 11, 2008).

MSN (2008b), "Limit crawl frequency", available at: http://search.msn.com/docs/siteowner.aspx?t=SEARCH_WEBMASTER_REF_RestrictAccessToSite.htm (accessed March 11, 2008).

Netcraft (2008), "Netcraft web server survey", available at: http://news.netcraft.com/archives/web_server_survey.html (accessed July 9, 2008).

Rao, S.S. (2003), "Copyright: its implications for electronic information", *Online Information Review*, Vol. 27 No. 4, pp. 264-75.

Reed, C. (2004), *Internet Law: Text and Materials*, Cambridge University Press, Cambridge.

Samuelson, P. (2003), "Unsolicited communications as trespass?", *Communication of ACM*, Vol. 46 No. 10, pp. 15-20.

Seadle, M. (2006), "Copyright in the networked world: using facts", *Library Hi Tech*, Vol. 24 No. 3, pp. 463-8.

Sieman, J.S. (2007), "Using the implied license to inject common sense into digital copyright", *North Carolina Law Review*, Vol. 85, pp. 885-930.

Smith, G.(.2.0.0.7.). (2008), "*Copiepresse* v. *Google* – the Belgian judgment dissected", available at: www.birdbird.com/english/publications/articles/Copiepresse-v-Google.cfm?RenderForPrint=1 (accessed July 2, 2008).

Snyder, H. and Rosenbaum, H. (1998), "How public is the web? Robots, access, and scholarly communication", *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, Vol. 35, pp. 453-62.

Spinello, R.A. (2007), "Intellectual property rights", *Library Hi Tech*, Vol. 25 No. 1, pp. 12-22.

Sterling, J.A.L. (2003), *World Copyright Law*, Sweet & Maxwell, London.

Tan, P-N. and Kumar, V. (2002), "Discovery of web robot sessions based on their navigational patterns", *Data Mining and Knowledge Discovery*, Vol. 6 No. 1, pp. 9-35.

Thelwall, M. and Stuart, D. (2006), "Web crawling ethics revisited: cost, privacy, and denial of service", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 13, pp. 1771-9.

Wong, M.W.S. (2007), "Cyber-trespass and 'unauthorized access' as legal mechanisms of access control: lessons from the US experience", *International Journal of Law & Information Technology*, Vol. 15 No. 2, pp. 90-128.

Yahoo (2008a), "Yahoo! Slurp – Yahoo!'s web crawler", available at: http://help.yahoo.com/l/us/yahoo/search/webcrawler/ (accessed March 11, 2008).

Yahoo (2008b), "How do I prevent my site or certain subdirectories from being crawled?", available at: http://help.yahoo.com/l/us/yahoo/search/webcrawler/slurp-02.html (accessed March 11, 2008).

Yahoo (2008c), "How do I keep my page from being cached in Yahoo! Search?", available at: http://help.yahoo.com/l/us/yahoo/search/webcrawler/slurp-05.html (accessed March 11, 2008).

Yahoo (2008d), "How can I reduce the number of requests you make on my web site?", available at: http://help.yahoo.com/l/us/yahoo/search/webCrawler/slurp-03.html (accessed March 11, 2008).

Youtube (2008), available at: www.youtube.com (accessed March 11, 2008).

Further reading

Koster, M. (2008), "Robot in the web: threat or treat?", available at: www.robotstxt.org/wc/threat-or-treat.html (accessed March 11, 2008).

Wikipedia (2008), "Web crawler", available at http://en.wikipedia.org/wiki/Web_crawler (accessed March 11, 2008).

About the authors

Chyan Yang is Professor at the Institute of Business and Management and Institute of Information Management, National Chiao Tung University, Taipei, Taiwan, Republic of China.

Hsien-Jyh Liao is Prosecutor at the Taipei Prosecutor Office, Taipei, Taiwan, Republic of China. Hsien-Jyh Liao is the corresponding author and can be contacted at: hjliao123@gmail.com