國立交通大學

資訊管理研究所

博士論文

可用於自動蒐集開放網路內容之著作權授權表達法

Expressions of Copyright Authorization Used for Automatically Acquiring Free Internet Materials

研 究 生：廖先志

指導教授：楊 千 教授

中 華 民 國 九 十 八 年 五 月 十 日

可用於自動蒐集開放網路內容之

著作權授權表達法

Expressions of Copyright Authorization Used for Automatically Acquiring Free Internet Materials

研 究 生： 廖先志 Student: LIAO, Hsien Jyh

指導教授： 楊千 Advisor: YANG, Chyan

國立交通大學

資訊管理研究所

博士論文

A Dissertation

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy in Information Management

May 2009

Hsinchu, Taiwan, the Republic of China

中華民國 98 年 5 月

可用於自動蒐集開放網路內容之
著作權授權表達法

學生：廖先志　　　　指導教授：楊 千

國立交通大學資訊管理研究所

摘要

網路圖書館在近幾年日益重要，而自由內容或開放內容這類概念的出現，也深深影響了網路圖書館典藏內容的豐富程度。但是著作權問題一直是成功建立網路圖書館的一個重要議題。實際上，如何合法而有效率地在網路上收集館藏是網路圖書館管理者的一項挑戰。雖然使用軟體機器人在網路上自動蒐集館藏是一種有效率的作法，但因爲軟體機器人通常無法正確地自動判讀在網路上所蒐集的作品其著作權授權範圍爲何，故使用該機器人之網路圖書館管理者在散佈或重製蒐集回來的作品時，會有潛藏的侵害著作權風險。因此，建立一種授權表達機制，一方面可以被作者用來充分表達著作權授權的範圍，一方面又可以自動地被軟體機器人判讀，是一項合理的解決方案。

本文試圖提出二種可以達到上述要求的授權表示方式：其中第一種是改良CC授權，稱爲CCFE，另外一種則是Robots標籤（Robots.txt and Robots Meta tags）的擴充版。CCFE可以很輕易地使可被機器人自動判讀的授權資訊與任何格式的檔案連結在一起。另一方面，經由擴充部分的標籤以後，Robots標籤也可以負擔起表達著作權授權範圍的工作。

關鍵字：著作權、網際網路、圖書館、軟體機器人、自由內容

Expressions of Copyright Authorization Used for Automatically Acquiring Free Internet Materials

Student: LIAO, Hsien Jyh Advisor: YANG, Chyan

Institute of Information Management
National Chiao Tung University

Abstract

Internet libraries have been gradually popular in recent years. The appearances of "Free Content" and "Open Content" actually affect the amounts of Internet libraries materials. However, copyright is one of the most important issues of construction of a successful Internet library. In fact, how to legally collecting works in an economic way is a great challenge for librarians. Launching software robots to automatically acquire works on the Internet is efficient but with high potential legal risks, because the robots can not automatically comprehend the real copyright authorization scope. As a result, the libraries distribute or reproduce the collected works may infringe the copyrights of authors. To solve this problem, an ideal solution is designing a scheme which can be identified by software robots and can be used to fully express copyright authorization scope.

In this thesis, we propose two mechanisms which both fulfill the two requirements above: one is an expansion of the Creative Commons license, the CCFE, and another is a revised edition of the Robots.txt and Robots Meta tags. The CCFE can reduce one of the main disadvantages of the original CC: machine-readable metadata can not be easily embedded in digital files. In addition, with some extra commands and tags, the Robots.txt and Robots Meta tags can also be used to express copyright authorization scope as well.

Keywords: copyright, Internet, library, software robots, free content

ACKNOWLEDGEMENT

經過六年的奮戰，終於拿到了這個博士學位，箇中辛苦，非親身經歷者，無法體會。

這一路走來，要感謝的人非常多，首先，要感謝指導教授楊千老師，沒有他的體諒與包容，恐怕我也走不到今天。其次，要感謝我的好同學陳鍾誠，他給予我許多支持與協助，論文的完成，他實在功不可沒。還有同班同學楊耿杰和盧浩鈞，也提供了我許多寶貴的協助。此外，其他許許多多的師長與朋友，也都提供了我不可或缺的幫助，此處要一併申謝。

當然，家人的支持更是必要的，我只能說，自己非常幸運，在求學的過程中，有如此支持我的家人，使我能在忙碌的工作之餘，能順利完成學業。

總之，回思這六年的求學過程，感謝上帝對我的眷愛，賜我一段如此豐美的歷程。

Contents

List of Tables

List of Figures

1. Introduction

1.1 Research Background

Libraries are important to culture development and its influence is gradually increasing in the today's Internet age; because the Internet effectively widens acquisition of libraries materials (Hundie, 2003), broadens the accessibility of libraries (Barker, 2001) and encourages communities to share information, rather than restricting access to it (McCray and etc., 2001). For example, the Citeseer (Citeseer, 1997) is a well known and popular online digital library. A large number of academic papers related to computer science can be searched on it (Giles et al, 1998). One important part of Citeseer is the software robot ("crawler" or "spider"), which can retrieve and store all related papers in Adobe Portable Document Format (PDF) or PostScript (PS) format from other Web sites (Raghavan et al, 2001). Citeseer then indexes these documents. Users may search Citeseer for documents pertinent to their area of research, and users may download one or more documents as required.

The first possible concern of an Internet librarian or library constructor is the amount of collections in the library. For example, Citeseer only focuses on the research papers in relation to computer science and, in order to acquire as many papers as possible, it employs software robots rather than manually collecting papers on the Internet. Generally speaking, an Internet librarian or a library constructor prefers collecting the largest amount of collections subject to the budget limit and the subjects. In the Internet world, software robots which can automatically acquire materials are a popular choice to achieve this goal. Moreover, a software robot with screening ability, such as keywords selection, can also help the library constructors to choose the works belonging to the preset subjects.

The next concern for an Internet library, along with the growing of the collections, is the copyright issue which is very essential to libraries; in fact, it may be the one which librarians most concern about, no matter for a traditional mortar-and-brick library or a digital library (Lopatin, 2006; McCray and etc., 2001). The copyright issue is arose when the collection of the library is still copyright protected. According to modern copyright laws, such as 17 U.S.C. 106 and the WIPO copyright treaty, 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). That is to say, copyright is one of the important issues which could impede the development of digital library because the dissemination of copyrighted works, one of the basic functions of a library, could result in copyright infringements (Bolin, 2006). In fact, subject to other same conditions, the amount of collections in a library free from copyright infringment allegations is definitly less than it of a library disregarding any copyright issues.

Before deeply discussing the collecting methods and copyright issues, it will be very helpful to examine several illustrative websites or libraries which acquire their collections via the Internet. We especially focus on what kinds of works in these sites, how these sites acquire collections and how they circumvent possible copyright infringement allegations.

The first example is the Internet Archive, also called as "WayBack Machine" which is an archive mainly consisting of copies of past Web pages on the Internet with the use of software robots (Internet Archive, 2009). Due to the fact that the Internet Archive is an non-commercial organization and its main purpose is reserving the historical data on the Internet rather than launching time-consuming negotiations with authors, the Internet Archive relies upon the 'fair use' and other related copyright law exemptions for libraries to be the defenses against potential copyright infringement allegations (Hirtle, 2003).

The next example is the websites which provide a Web space for authors to upload their own articles and for contributors to publish others' works with fully permissions, such as the Scribd and the Issuu (Scridb, 2009; Issuu, 2009a). In fact, a website, like Scribd or Issuu, is an agent or mediator, which only offers an platform where right owners and users could interchange with each other: right owners could release their works on the library site as long as grant some copyrights and, accordingly, the users could choose the works not only meeting their specific purposes but within the scope of authorization as well. As soon as these uploaded files are alleged to infringe any copyrights, the webmasters will instantly remove all suspected materials whenever receiving notices (Issuu, 2009b). In other words, an library adopting this strategy counts on the licensing from authors as well as the Safe Harbor exemption, such as 17 U.S.C. 512, as it does not precisely examine whether the contributors have real authority or not.

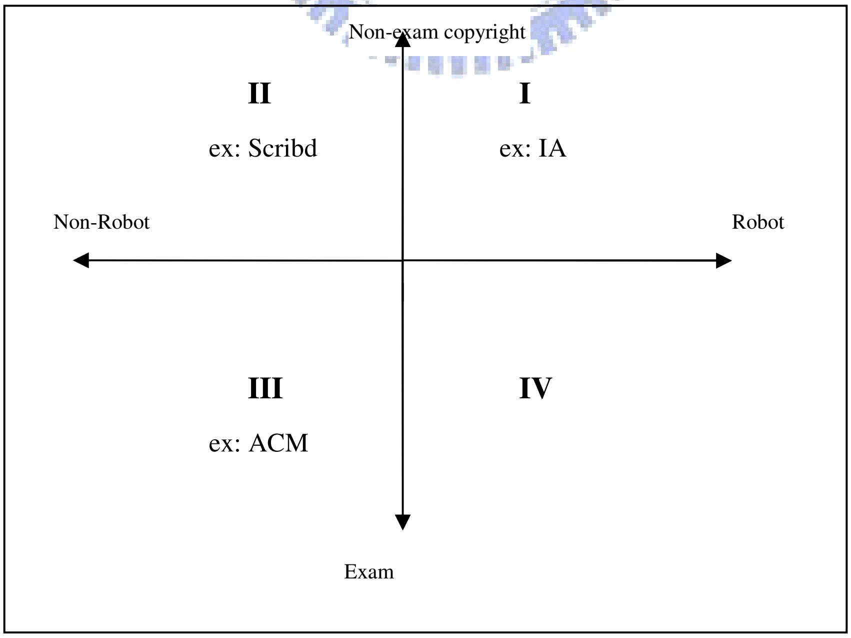
We can find out that the first two examples both rely upon the exemptions of copyright laws. Another straight way to avoid potential copyright infringement allegations is constructing a website where all collections are owned by him and, no one, except the librarian himself, could have rights against him. In other words, the librarians may contract with the content owners or the right holders and make a proper arrangement of the benefits. For instance, ACM Digital Library only collects all articles subject to its copyright terms (ACM Digital Library, 2009). Nevertheless, because the negotiation process may be costly as well as direct communication to the numerous authors on the Internet is almost impossible; Internet libraries belonging to this model are all business, main-stream publishers or media. For instance, BBC built a trial site, BBC Creative Archive, to release more than 500 full TV programs (BBC, 2006).

Moreover, a similar example is only focusing on the work without copyright protection. For example, Project Gutenberg announces to encourage the creation and distribution of eBooks, mainly the works in public domain (Hart, 2004). That is to say, all collections in this Website merely consist of public domain or out-of-copyrighted works and, as a result, no one could challenge a depository of this kind about the copyright.

In fact, the present Internet libraries may adopt one or more strategies rather than a pure one. For example, the main materials of the Project Gutenberg are in public domain under US Copyright law, as long as few materials are subject to authors' permission[1]. The Citeseer is another example, which not only employs software robots to collect articles on the Internet, but also allow authors to submit their article to this library (CiteseerX.ist, 2009).

1.2 Research Motivation

As we being above-mentioned, the two concerns--how to collect works and how to circumvent potential copyright infringement allegations--are very important to the Internet library constructor. The foregoing examples demonstrate several strategies adopted by the website constructors in respect of these two essential concerns. In terms of the first concern, there are two choices available to the website operators: one is employing software and another is collecting works manually. As to the copyright issue, specifically examining copyright to make sure how he can use the works is one option; another option is relying upon the copyright exemptions. In fact, an Internet library is a website from the users' viewpoints. That is to say, an Internet library constructor may adopt the strategies similar to the website operators. Therefore, if we focus on "employing software robots to collect works" as well as "examining the copyright" and use these two as the vertical and horizontal axes, the websites can



[1] http://en.wikipedia.org/wiki/Project_Gutenberg

be placed in the one of the four quadrants in the following diagram:

Figure 1: Libraries in the four quadrants

In the first quadrant, a library (Model I Library) relies on the traditional library exemptions to avoid potential copyright infringement allegations. The second (Model II Library) and the third (Model III Library) both depend on the licensing of authors, but the Safe-Harbor exemption is more important to Model II Library the because the libraries of this model, at most times, do not explicitly monitor the correctness of the licenses it obtained, rather disseminating works in good faith. Moreover, libraries merely focus on the materials of the public domain should be placed in the third quadrant as well.

In light of the various strategies, the risks of copyright infringement allegations are different as well: not surprisingly, libraries belonging to the first and second models have the highest risks; the reason will be rendered in the following sections. On the contrary, the risk of a model III Library is relatively low. However, in the real world, the lower risk is not free at all and, in fact, is relatively expensive: As to a library of the third model, the time and money spent in completing the negotiations between the publishers and authors are quite significant. On the other side, the cost in respect of confirming copyright authorization scopes of the other two models are relatively low: libraries of the first model do not pay any attention on it and, libraries belonging to the second model almost pay nothing neither because a Model II library only removes works whenever it receiving notices.

Apart from the concern about the copyright infringement, another important concern is about the way to create collections in the library. As we have seen, the libraries of the first one model clearly face a higher legal risk than libraries of other three models. In general, the reason of taking such high risk is that, subject to the same budget, the total amount of collections in a model one library is higher than the other three models and, at the most times, the amount of collections is one of the most critical issues to a library which may actually affect the users' favors. The reason why a library of the first one model can acquire more works than the others is that it employs software robots to collect works on the Internet. In respect of the huge number of collections in the libraries of the first model, fair use, or other general copyright exemptions, is the only effective way which libraries of this model could account on because the total number of works is too massive to explicitly identify the scope of copyright authorization.

On the other hand, libraries of the other two models collect works without any software robot. Since a library of the model II depends on favors of the contributors or authors, the constructors of a library of this model can not passively decide the total amount of its collections; as a result, in general, the total amount of collections in a Model II library is less than it in a Model I library.

As to the other the third model, the amount of collections of a library of this model is relatively limited, because it collects works by hand and, in general, the human's work speed is less than an unstopped software robots. For example, in spite the amounts of collections in some present libraries, such as ACM Digital Library (ACM Digital Library, 2009), are relatively large; however, comparing to the total number of works on the Internet, the collections of a library belonging to this model are still relatively limited because such libraries have to be subject to their budgets. A summary of these three models are also shown as follows:

Table I. A summary of the four models in risks, costs and amounts of collections

	Copyright Policy	Risk	Cost	Number
I	General copyright exemptions: fair use etc.	High	Low	Almost unlimited
II	Licensing from authors and the Safe Harbor exemption	Medium	Low	Limited
III	Licensing from authors	Low	High	Relatively limited

The Model I and Model II libraries both depend on copyright exemptions, however, the traditional library exemptions could not directly and clearly apply under this circumstance since the conditions are not satisfied (Bolin, 2006). Furthermore, great diversities appearing in the copyright limitation and exception rules in different national laws increase the risks. For example, the scope of “fair dealing” in the UK is much narrower than “fair use” in the US, as the former has no general exception of the later (Cornish and etc., 2003a). Moreover, the “private use” exceptions in the civil law countries is much common than it in the common law countries, as the civil law countries respect the intelligence in the work rather than the exploitation benefits in it. On these two grounds, the Internet libraries can not firmly rely upon the limitations and exceptions to lawfully access to, reproduce, even redistribute as the exceptions of individual national legislations are diverse and, under some circumstances, unpredictable. Even though, ignoring the diversities and uncertainties of copyright limitations and exceptions, the copyright exceptions could be applied, the ‘fair use’ or other similar exceptions inevitably undermines the quality of contents because the future uses of the contents are bounded because users of the library can not be sure what the exact authorization scope of the work is.

On the other hand, the simplest solution to reduce such high legal risks is to explicitly examine the copyright of each work, such as getting license from the authors or only collecting public domain works. However, a specific analysis of the copyright of a work is very difficult and needs a lot of human and financial resources. As a result, the number of the libraries belonging to this model is quite limited.

1.3 Possible Ways to solve the Copyright Problem

Instead of expensive human intervention, there are other two main possible useful ways to avoid the potential copyright infringement allegations (Lessig, 2006a): the first approach is definitely the law. For example, a government can grant a totally new copyright exemption which only applies to the Internet library or, directly amplify the reach of fair use exemption. The next useful way is the code. In the context of the Internet, the code, which, more specifically, is software or hardware, forms cyberspace what it is and constitutes a set of constraints on how you can behave (Lessig, 2006b). On this ground, designing a new software robot which can precisely identify the authorization scope is a possibly useful way to reduce the risk on copyright infringement allegations.

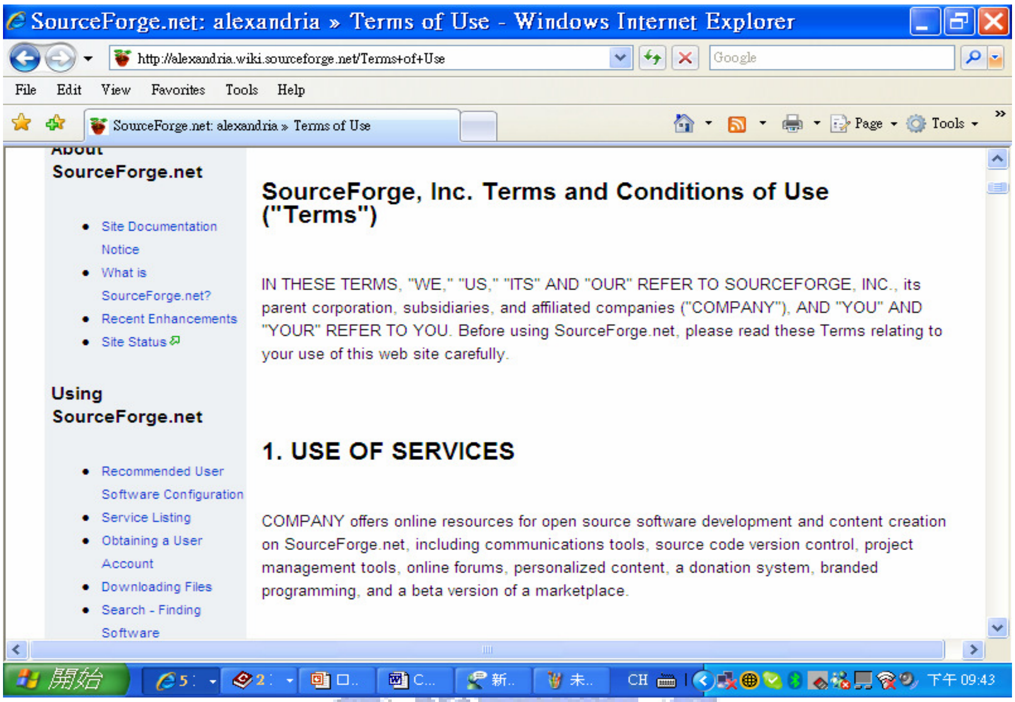
Even though these two ways can both effectively solve the copyright infringement problem. However, a new exemption may inevitably conflict with the present rules of copyright laws; therefore, it is not a proper choice for the unpredicted consequences. Moreover, a new exemption needs a lot of researches and discussions; in other words, it is very time-consuming. On the other hand, in general, the change in architecture of the Internet may be fiscally cheaper than granting a new exemption, because the process of getting a segment of new code is much easier.

On all the reasons above, employing software robots to automatically collect works, including copyrighted and out-of-copyrighted works, and identifying the explicit authorization scope of the collected works is the best strategy for an Internet library. That is to say, a library belongs to this model, in quadrant IV, could achieve the goal of broadest collections as well as facing a low risk of copyright infringements.

Nevertheless, this mixed strategy is nothing more than an ideal one in the current time and, in fact, no Internet library so far could launch a software robot with an ability to automatically collect works as well as explicitly identify copyright authorization scope. In fact, just few software robots are able to differentiate between a copyrighted document and a document that has been posted by an author for general use; as a result, they simply automatically retrieve all papers via the Internet.

Some technical hurdles actually impede the advance of the Internet library, especially in respect of the ability to automatically identify authorization scope: The first one is: the real meaning of such information, especially in terms of the legal meaning, is not easy to understand without human beings interferences. To speak more explicitly, there are two kinds of difficulties involved: at first, the information, especially expressed in natural language, could not be perfectly identified and comprehended by software robots. Consequently, the misunderstandings by software robots could inevitably lead a misjudgment of the copyright authorization scope. Secondly, the vague expression could also result in some misunderstandings. For example, a common jargon “Under Copyright Law Protection”, without specifically indicating under which nation’s copyright laws, may mislead software robots and result in ambiguities to some extents.

The second difficulty is that, even though the right meaning of authorization information could be specifically understood by software robots, the exact location of the authorization information of a particular work is not easy to be determined. For example, in SourceForge, all programs are under the same GPL license, which is expressing in the “Term of Use” section of the website, as shown in Figure 2.


SourceForge.net: alexandria » Terms of Use - Windows Internet Explorer
http://alexandria.wiki.sourceforge.net/Terms+of+Use
Google
File
Edit
View
Favorites
Tools
Help
SourceForge.net: alexandria » Terms of Use
Page
Tools
開始
CH
下午 09:43

About SourceForge.net

- Site Documentation Notice
- What is SourceForge.net?
- Recent Enhancements
- Site Status

Using SourceForge.net

- Recommended User Software Configuration
- Service Listing
- Obtaining a User Account
- Downloading Files
- Search - Finding Software

SourceForge, Inc. Terms and Conditions of Use ("Terms")

IN THESE TERMS, "WE," "US," "ITS" AND "OUR" REFER TO SOURCEFORGE, INC., its parent corporation, subsidiaries, and affiliated companies ("COMPANY"), AND "YOU" AND "YOUR" REFER TO YOU. Before using SourceForge.net, please read these Terms relating to your use of this web site carefully.

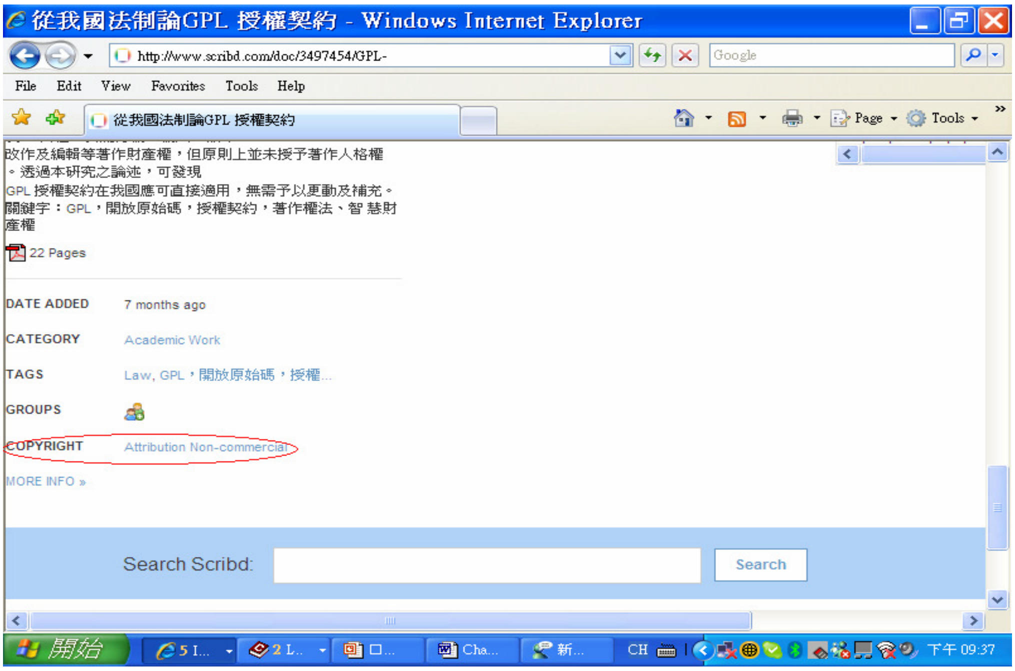
1. USE OF SERVICES

COMPANY offers online resources for open source software development and content creation on SourceForge.net, including communications tools, source code version control, project management tools, online forums, personalized content, a donation system, branded programming, and a beta version of a marketplace.

Figure 2: A snapshot of SourceForge’s “Terms and Conditions of Use”[2]

On the other hand, every document in Scribd is licensed under the same Creative Common license, as shown in Figure 3. However, as illustrated in these two figures, the locations of the authorization information are different: one is on another page and one is in the same page.

[2] http://alexandria.wiki.sourceforge.net/Terms+of+Use



從我國法制論GPL 授權契約 - Windows Internet Explorer
http://www.scribd.com/doc/3497454/GPL-
Google
File
Edit
View
Favorites
Tools
Help
從我國法制論GPL 授權契約
Page
Tools
開始
CH
下午 09:37

改作及編輯等著作財產權，但原則上並未授予著作人格權
。透過本研究之論述，可發現
GPL 授權契約在我國應可直接適用，無需予以更動及補充。
關鍵字：GPL，開放原始碼，授權契約，著作權法、智 慧財
產權

22 Pages

DATE ADDED	7 months ago
CATEGORY	Academic Work
TAGS	Law, GPL，開放原始碼，授權...
GROUPS	
COPYRIGHT	Attribution Non-commercial

MORE INFO »

Search Scribd: Search

Figure 3: A snapshot of a document in Scribd[3]

In order to solve the two difficulties above, the first suggestions is offering a much complex software robot: a robot with great artificial intelligence as well as high-level information retrieval technology to find out which piece of information is the real one and to comprehend the legal meaning of information in natural languages. Nevertheless, technologies in these two areas--artificial intelligence and information retrieval, are very complex and, in fact, a software robot with such ability has not existed yet.

Therefore, the next suggestion is offering the authors of the works a mechanism which could be easily understood by the robots, as well as could be used to properly express the copyright authorization scope should be a more practice measure. To speak more explicitly, a mechanism which fulfills two minimum requirements could be used in such circumstances: the first requirement is that the mechanism should be fully identified by software robots and, the second one is that this mechanism should

[3] http://www.scribd.com/doc/3497454/GPL-

have flexible ability to express the copyright authorization scope of works, no matters what types of works.

Furthermore, we hope to construct a library not only acquiring collections by software robots, but also focusing on free and open works. The reason is that a library with free and open works can effectively encourage exchanges of all works on the Internet and, as a result, stimulate more developments and reservations of cultures. We hope the mechanisms proposed in this thesis can be useful to achieve this goal.

Based on the foregoing discussion, a fixed term expression, rather than natural languages is a more ideal proposal. Moreover, the popularity of a fixed term expression is very important, because search engines, the most common users of robots, only support several popular fixed-term expressions and this fact will finally decide the number of users of the proposed expressing methods. In other words, a well designed but unpopular fixed-term expression is nothing but an unrealistic imagination.

In the present Internet world, there are two popular fixed-term expressions: the Creative Commons (CC license thereafter) and the Robots.txt and Meta tags. These two mechanisms are dedicatedly designed for software robots, that is to say, any further modification of these two could easily be understood by software robots. More importantly, these two approaches are all supported by popular search engines' robots, such as Google (Google, 2008b), Yahoo (Yahoo, 2008b), and MSN (MSN, 2008a). However, with regard to expressing the copyright authorization scope, some drawbacks appear to these two schemes: even the CC license covers several common copyright authorization choices, it still have some disadvantages and needs further modifications, especially for works in some kinds of digital forms. On the other hand, the Robots.txt and Meta tags are purposely designed for software robots and very easy to use, but do not focus on expressing explicit copyright authorization. As a result, all these two candidates need some modifications. Furthermore, in respect of the licensing on the Internet, there are two kinds of people in need of expressions of copyright authorization scope. The first one, not surprisingly, is the author of a work. In addition to differently licensing individual works, in the Internet world, the author may be in need of licensing all works in one website or a Web page under the same condition. For example, in Scribd, all works are licensed under the same CC license, as shown in Figure 3. Therefore, the second kind of people who need expressions of copyright authorization is the webmasters who operate the websites or the Web page owners who manage the Web page. In general, the site and all pages reside in this site may be owned by the same person; therefore, we use the term, webmaster, to represent the people who are in need of expressions for identically licensing all works.

This thesis is structured as follows: at the beginning of this thesis, we will review some primary concepts, such as digital libraries, software robots, the Internet copyright issues and related terminologies. In the next sections, concerning the above-mentioned two kinds of persons, who are in need of authorization expressions, we first try to pay our attention to the webmasters who authorize all works in the same page. The Robot.txt and the CC license as well as the Robots Meta tags can both be used to license works in the same Web page. Nevertheless, the Robot.txt and Robots Meta tags need a minor amendment to fully express the copyright authorization scope, whereas, the CC licensing scheme can be used to license works not only identically but also individually. Nevertheless, a new revision of the original CC license is proposed in the following section, which can reduce the disadvantages of the original CC license in terms of licensing each particular work. Next, we compare the foregoing revision and amendments before finally discussing some unsolved problems while suggesting additional issues that invite future research.

2. Literature Review and Terminologies Overview

2.1 Internet Library

The concept or definition of a digital library varies in respect of different perspectives. In respect of technology adopted in a digital library, a digital library may be defined as follows:

> *Digital Libraries basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net. (National Science Foundation, 1999)*

This definition was crtisized for putting it weight on merely technical aspects. (Seadle and etc., 2007). On the other hand, as regarding importance of the orginazation underlying the collections and computer systems where collections resided, a digital library could be:

> *Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. (Digital Library Federation, 1998)*

Based on this definition, a digital library and a digital archive are two different in terms of the nature of works collected and preservation functions. (Digital Library Federation, 1998)

On the other hand, Internet Archive is created as a repository of websites. Several aggregating projects, including Google, MSN, Yahoo, Internet Archive (Internet Archive, 2009) and several foreign national libraries have regularly taken snapshots subset of the Internet. In order to allow access when the original page temporarily inaccessible, or allow viewers to compare changes made to pages during a specific period, some commercial search engines, such as Google, MSN, and Yahoo, display search results always includes a link to its own cached copy, which is a temporary repository consisting all source codes of indexed websites (Field, 2006).

Although, from a librarian's professional perspectives, the features of an Internet library do not only embrace digital contents and access via the Internet; other facilities, such as online assistances and comprehensive online references are essential as well (Jones, 2001). On this ground, an Internet archive does not quailed as a digital library. However, the introducing of new technologies, such as search engines and new search algorithms blur the line between them in some degrees; that is why these two terms may appears in an totally equivalent form in some cases, such as "The Internet Archive was founded in 1996 to build an "Internet library" that will offer permanent access for researchers and scholars to historical collections that exist in digital format (Feldman, 2004)." In fact, in terms of the digitalized contents and easy accessibility via the Internet, an Internet and an Internet library could be generally considered as a same term. In addition, as shown in the following sections, both Internet libraries as well as Internet archives face the same legal menace: copyright infringement and, the resolution in terms of this threat are identical. Therefore, in this thesis, it is not necessary to make difference between the 'Internet library' and the 'Internet archive' and we use 'Internet library' to commonly represent them both and a further explanation will be render in some special circumstances.

2.2 Internet Copyright

2.2.1 The diversities and harmonization of international copyright law

In order to design a comprehensive copyright authorization scheme in respect of software robots, the most important and fundamental work is studying what are the essential components of the copyright in the internet context, especially with regard to the software robots accessing. In the context of internet, the accesses of software robots is boundless, that is to say, most of the accesses may cross the national boarders. From this standpoint, the authorization of copyright inevitably involves copyright legislations of more than one country. Therefore, the “copyright law” we have to study here is not limited in national legislations of any specific country, rather, is the international copyright laws.

The first critical fact that we have to notice is that copyright legislations in different nations are different, as many other fields of laws. The basic ideas, philosophy and principles of the same term “copyright” are quite different in different countries. In terms of the basic ideas behind the copyright, the worldwide copyright legislations can be generally classified into two separate systems: the author’s right and copyright right systems. The civil law countries, such as France and Germen, consider the author’s personality expressed in the work constitute the basic interest which should be respect and protect. On the other side, the common law countries, such as the UK, focus on the economic exploitation interest in the work, rather than the personality of author. Based on these separate basic ideas of copyright, there are several diversities between them which are important in the context of internet. For example, a work must be “original” is the same basic requirement in these two systems, however, the criteria of “original” is different, at least in theory. In author’s right countries, in respect the personality, a copyrighted work should represent the creation or intelligence of the author. In copyright countries, however, the traditional standard of “originality” is only sufficient “investment of money, time, and labor”, regardless of creation or intelligence (Sterling, J.A.L., 2003c).

With the increasing advert of interchange of the world, especially with the rapid growth of internet, the diversity in international copyright legislation is gradually deemed as some kinds of hurdles which may become an impede of the information society. As a result, many international treaties, such as Berne Convention, WIPO Copyright Treaty, WIPO Performance and Phonograms Treaty, appear to harmonize and reduce the differences of copyright legislations between different countries. The basic infrastructure constructed by those international treaties provides us a well basic scheme which we can use to analyze and discuss the substantial contents of copyright in relation to the software robots' access and authorization.

2.2.2 The rights within the term "copyrights"

As the "copyright" is not a single right; instead, it can be seen as a set of rights which, according to author, can be generally classified under the headings of "moral right" and "economic right" (Sterling, 2003d). In general, the moral rights are those which relate to the protection for the personality of the author as expressed in their creations (Cornish and etc., 2003b). Economic rights, on the other hand, are those concerning control over the commercial or industrial exploitation of works, and other means of use of the works which involve such acts as reproduction or representation, but do not of themselves necessarily involve prejudice to the reputation of the author or the integrity of the work (Sterling, 2003d). In the internet environment, the main rights which may be infringed are: moral rights and related economic rights, including rights concerned with reproduction and adaptation and rights concerned with communication to the public (Sterling, 2003e).

We have to notice that not only the economic right taking an important part in the copyright infringement on the internet, with the appearance of information aggregation service, but the moral rights gradually play more significant roles. The most recent noticeable-worthy case is a Belgium case: Copiepresse v Google (Copispresse, 2007). In this case, the plaintiff Copiepresse is the representation of some Belgium French newspapers, who assert that one of the services provide by Google, the Google news, infringes the copyright of the Belgium newspapers. The software robots of Google news, retrieved the titles of those papers and, revised the titles and published them on the website of Google news, without the writers' consents. Based on this fact, instead of alleging the infringement of economic rights, the plaintiff claimed that the paternity right and integrity right are be infringed as well. The court of the first instance agreed the allegation about moral rights and, however, this case is still in appeal (Copiepresse, 2009).

To sum up, while the copyright legislation in different countries quite diverse; however, in light of the international or regional conventions and treaties, we still can draw a basic scope of the economic and moral rights, which can be seen as the essentials of the copyright. Firstly, in respect of the adaptation/modification right and reproduction right arise little controversy, even in concern with the "transient copying". On the other hand, the legal meanings of distribution right are different in the US and other countries. However, we can generally use the term "distribution/communication right", which combing the "communication rights" defined in the WIPO treaties and the "distribution right" in the US, to represent the right of authors to control the dissemination of the works on the internet. Secondly, with regard to the moral rights, the paternity right and the integrity right are two commonly recognized moral rights and the other three moral right, the divulgation, the retraction right and the deconstruction right, are only partly recognized. However, according to the inalienability of the moral rights, the authorization scheme is not necessary to the moral rights.

2.3 The 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 Websites. 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 etc., 2003). The first is "Building collections": software robots have been extensively used to access and collect data of websites that are required to create an index for application programs, such as search engines. The second use is "Archiving": a few projects, like Citeseer (Citeseer, 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 websites 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 websites 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 the following diagram:

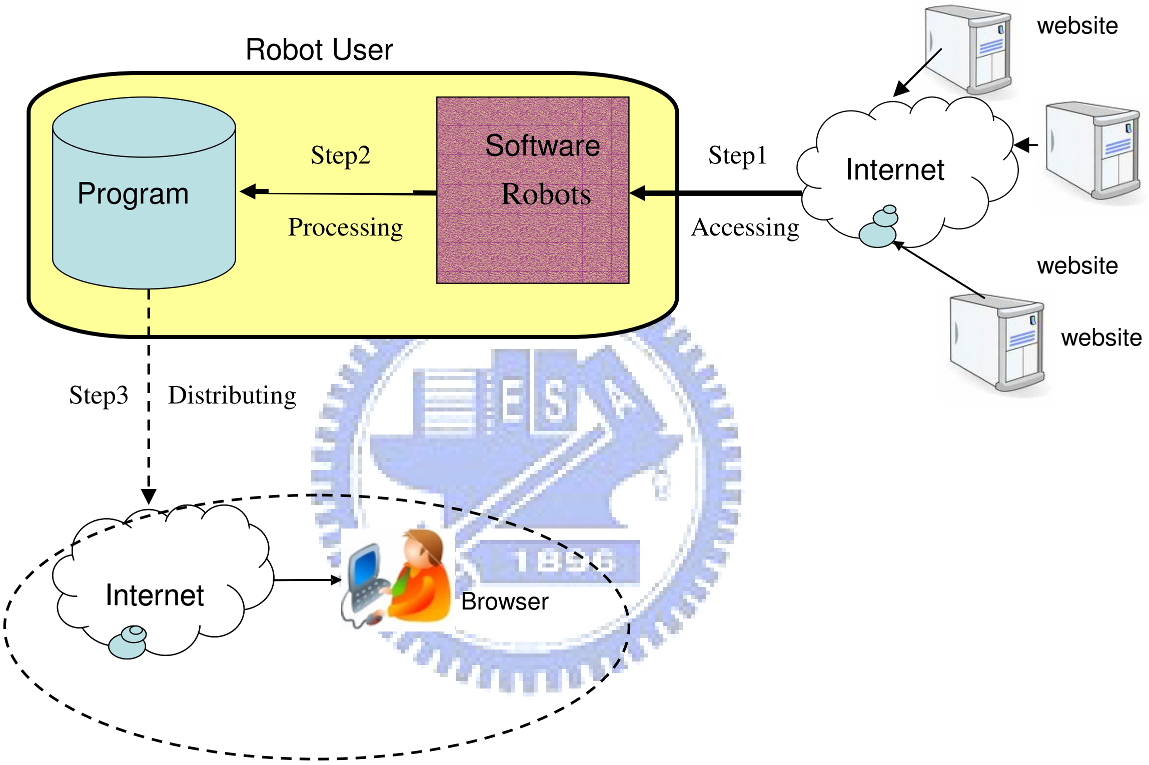


Figure 4: The process of how a software robot works

In this diagram, 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.

2.4 The DRM and Other Related Measures to Control Copyright

The digital right management (DRM hereafter) refers to technologies employed by right owners and devices manufactures to control, to restrict or manage the use of the works[4]. Although the right here does not limit to the copyright, the control and management of the copyright are the main parts involving in DRM. Some opponents allege that the use of the word "rights" is misleading and suggest that people should use the term Digital Restrictions Management to show its essential features (Free Software Foundation, 2006).

With regarding to the components of DRM, the authoring policy expression is one of the key components and, in fact, is a main challenge to implement DRM (LaMacchia, 2002). As a result, tools which can explicitly express the scope of rights granted to the users are very essential to implementation of DRM. In respect of the set of Rights Expression Language (REL), ODRL (Open Digital Rights Language) is an XML-based standard REL and can be adopted to describe the rights granted to the user (ODRL Initiative, 2009). However, although ODRL can used to express the CC license as well (ODRL Initiative, 2005), is still belongs to DRM family; the "open" here only refers to that it is an "open" standard or an "open source" project, not refers to the works licensed by it are open.

In addition to ODRL, a similar tool available to authors or publishers to control and manage the use of works is Digital Object Identifier (DOI hereafter). The International DOI Foundation (IDF) defines DOI as "a digital identifier for any object of intellectual property"; further, it explains that the DOI is used for "persistently identifying a piece of intellectual property on a digital network and associating it with related current data in a structured extensible way." (International DOI Foundation, 2008) Though DOI can be used to assist authors or publishers to implement his copyrights to their works as well (Rosenblatt, 1997), we have to notice that getting a new DOI is not "free"; an administrative fee is paid for each allocation by the agency to the IDF. As a result, it is not a proper tool for open works.

[4] http://en.wikipedia.org/wiki/Digital_rights_management#.22DRM-Free.22

2.5 Open Content and Free Culture

Rather than control and management, someone believes that free use and exchange of works can actually stimulate and encourage more culture developments. People believing in this idea are really opposed to DRM, because the control and management led by DRM totally contradict the basic principle of “open content” and “free culture”. However, the ideal of “open content” or “free culture” promoted by the groups of those people is only a vague concept; there are several practical varieties derived from this basic principle.

Free Software is one the earliest movement which not only influences the free culture but also the development of the software industry. Basically, free software shall grant users freedom to run, copy, distribute, study, change and improve the software (Free Software Foundation, 1999). To embody the concept of Free Software, several licenses are introduced. The most widely spread one may be the GNU General Public License (GPL hereafter), which allows users to run, copy and distribute the software, but users shall license their modification subject to the same conditions (Free Software Foundation, 2007). In addition to the GPL, the Berkeley Software Distribution License, which grants users almost every right, is another popular free software license (Open Source Initiative, 2006).

Even the free software movements evolves and grows rapidly, some difficulties still impede the further developments of it. The most obvious one is that all licenses are only fixed to the software; other kinds of works, such as images, audio works, are not embraced in the realm of any free software license. Moreover, the diversities of copyright laws lead the uncertainties of the real legal meaning of terms in copyright laws. For example, the “freedom of distribution” in the GPL, mainly based on the US copyright laws, needs some explanations when applying in other jurisdictions. Furthermore, the free software licenses basically ask the authors left their copyrights and such inflexibility actually affects its popularity to some extents. On these grounds, Lawrence Lessig, a law professor in Stanford University, designed and promoted a new licensing scheme, Creative Commons, which allows and encourages authors to grant their several baseline rights to others. The details of this license scheme will be explicitly rendered in the next section.

With regarding to the scholar works, Open Access is another branch based on the above “free and open” ideal. There are a variety of definitions of "open access;" in fact, this concept is still evolving with the development of Internet and free culture. However, the following definition, based on the "Budapest Open Access Initiative" (BOAI), is the most influential one to this day (Budapest Open Access Initiative, 2002):

> *The literature that should be freely accessible online is that which scholars give to the world without expectation of payment. Primarily, this category encompasses their peer-reviewed journal articles, but it also includes any unreviewed preprints that they might wish to put online for comment or to alert colleagues to important research findings. There are many degrees and kinds of wider and easier access to this literature. By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. . . .*

Several key points of Open Access can be derived from this popular definition: the first one is that the literatures should be freely available. The second one is that users should access to the works via Internet; that is to say, all works should be digitalized. The third essential element is all works should be only for academic uses. The last one is about the copyright laws: works subjects to Open Access are still under copyright law protections. Users are fully permitted to freely copy and distribute the works, apart from the requirements of proper attribution of the author and the assurance of the integrity of the work (Bailey, 2006). On the other hand, Open Content Alliance is a consortium of nonprofit organizations focus on digitizing several works without copyright protection and permitting users freely accessing to the digital contents via Internet. (Open Content Alliance, 2009; O'Leary, 2009)

3. Expressions for Licensing All Works in a Website

In this section, we will introduce two schemes, the CC license as well as the Robots.txt and Robots Meta tags, which can be used by webmasters to indicate the copyright authorization scope of the works in a website. Furthermore, some amendments will also be introduced to reduce the disadvantages of the Robots.txt and Robots Meta tags in terms of expressing copyright licensing.

3.1 Creative Commons License Framework

3.1.1 The basic of the CC license

The CC license is a license for the purpose of granting some or all of the authors' rights to the public. The CC license is not limited to software or documents. This license is designed for a broad range of contents, including but not limited to documents, animation files, and other types of information objects. The CC license is popular on the web now. The number of the documents licensed under the CC license and known as the CC licensed documents has been increasing in recent years. One significant boost to the CC licensing is Google's and Yahoo's inclusion of support to allow users to search only CC licensed documents (Google, 2007c; Yahoo, 2007). These two systems combined process nearly 80 percent of English language queries worldwide, these companies' support has been a positive step forward for the CC license (ClikZ Network, 2007).

The Creative Commons (CC) is an organization which designed the CC license[5]. It gives authors a way to grant some or all of their copyrights to the public. The first CC licenses appeared in December 2002. The guiding principle of the CC license is to complement copyright law rather than competing with it (Lessig, 2004).

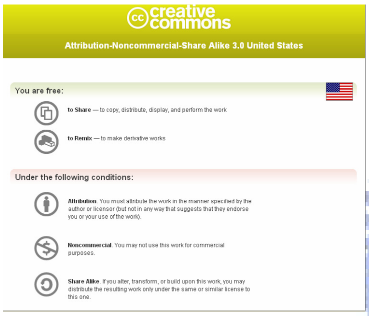
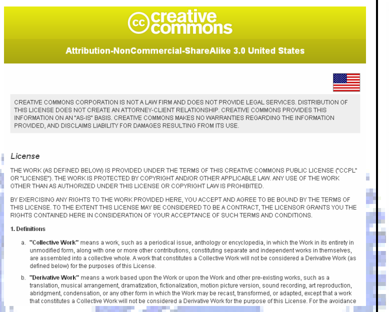
The present CC license can be used in a wide variety of works, including audio, video, images, and texts. There are three ways to express a CC license: the first way is called the "Commons Deed" which is a set of basic, human-readable, plain-language

[5] http://creativecommons.org/about/history (accessed July 3, 2007)

icons that states what a user may do with the content. The second way is called the "Legal Code", which is an authentication document with formal and explicit legal terms. The "Legal Code" always draws up the clear scope of licensing for the work. The third option is the "Digital Code, which consisting of lines of machine-readable metadata or a "digital signature" of the license. A software robot can process these metadata and tags the document as governed by the CC license. The key point is that an author may use one of these ways, or mix and match them to suit the author's needs. Table II shows an example of CC license of a document in all three ways.

Table II. An example of the three formats of CC license

Commons Deed[6]	Legal Code[7]	Digital Code[8]

 creative commons
Attribution-Noncommercial-Share Alike 3.0 United States
You are free:
to Share — to copy, distribute, display, and perform the work
to Remix — to make derivative works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Noncommercial. You may not use this work for commercial purposes.
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.
	 creative commons
	Attribution-NonCommercial-ShareAlike 3.0 United States
	CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.
	License
	THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.
	BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.
	1. Definitions
	a. **"Collective Work"** means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with one or more other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.
	b. **"Derivative Work"** means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance
		`<a rel="license"`
		`  href="http://creativecommons.org`
		`     /licenses/by-nc-sa/3.0/us/">`
		`  <img alt="Creative Commons License"`
		`     style="border-width:0"`
		`     src="http://i.creativecommons.org/`
		`          l/by-nc-sa/3.0/us/88x31.png"/>`
		`</a>`
		`<br/>`
		`  This work is licensed under a`
		`  <a rel="license"`
		`    href="http://creativecommons.org`
		`    /licenses/by-nc-sa/3.0/us/">`
		`  Creative Commons`
		`  Attribution-Noncommercial-Share Alike`
		`  3.0 United States License`
		`</a>.`

In respect of the scope of copyright authorization, to simply speaking, the CC license has four options: Attribution (by)[9], No Derivative Works (nd)[10], Share Alike (sa)[11] and No Commercial Use (nc)[12]. The characteristics and meanings of these four options are shown in the following table:

Table III: Four options in CC license[13]

Options	Abbreviation	Icons	Characteristics and Meanings

[6] http://creativecommons.org/licenses/by-sa/3.0/us/

[7] http://creativecommons.org/licenses/by-sa/3.0/us/legalcode

[8] http://creativecommons.org/license/work-html-popup?license_code=by-nc

[9] http://creativecommons.org/licenses/by/3.0/

[10] http://creativecommons.org/licenses/by-nd/3.0/

[11] http://creativecommons.org/licenses/by-sa/3.0/

[12] http://creativecommons.org/licenses/by-nc/3.0

[13] http://creativecommons.org/about/licenses

Attribution	By	BY:	The licensee must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
No Derivative Works	Nd	=	The licensee may not alter, transform, or build upon this work.
Share Alike	Sa	↻	If the licensee alters, transforms, or builds upon this work, he/she may distribute the resulting work only under the same, similar or a compatible license.
No Commercial Use	Nc	$	The licensee may not use this work for commercial purposes.

These four conditions can be combined to form six available different choices shown in the following table[14]:

Table IV: Six different choices of the CC license[15]

CC licenses	Abbreviation	Icons
Attribution	By	BY:
Attribution, No Derivative Works	by-nd	BY: =
Attribution, No Commercial Use, No Derivative Works	by-nc-nd	BY: $ =
Attribution, No Commercial Use	by-nc	BY: $
Attribution, Share Alike	by-sa	BY: ↻
Attribution, No Commercial Use, Share Alike	by-nc-sa	BY: $ ↻

[14] http://creativecommons.org/licenses/

[15] http://creativecommons.org/about/licenses

To explore the legal meanings of the four options, the first one, the Attribution (by) just emphasizes the importance and the inalienable characteristic of author's moral rights again. The second and the third options, No Derivative Works (nd) and Share Alike (sa) both connect to the modification right: the former barely prohibits any modification and the later permits further modification under some certain conditions. The last option, No Commercial Use (nc), only indicates one critical licensing condition, with no relation to any specific copyright rights. In light of the legal meanings of the four options, apart from the moral rights and the modification right, it can be seen that activities involving in the other two major economic rights, the reproduction and distribution/communication rights should be subject to the conditions set by the four basic options. For example, under the "Attribution, No Commercial Use" license, the licensee could not reproduce and disseminate the works for commercial purposes[16].

The CC license is considered much easier to use and understand than other licenses, like GPL (Lin et al, 2006). In addition, the CC license's official Web site provides an online license software "wizard" to help authors to choose the most appropriate license. The author answers three questions about the rights they want to grant[17].

3.1.2 How to use the CC license in different countries

Based on the above discussions and analysis about the copyright legislations in different countries, we may easily found out that the philosophies, structures and scopes are quite diverse, even more, the same term in two different legislations, such as "distribution right", represents varied meanings. In the context of Internet, such differences give arise difficulties with the exercise of copyrights, including both economic rights and moral rights, especially when some rights are recognized in some countries and not in others. The introduction of international conventions may lessen such inconsistence; however, the guidelines proposed in the international conventions

[16] http://creativecommons.org/licenses/by-nc/2.0/uk/legalcode

[17] http://creativecommons.org/license/?lang=en

are quite limited. From this perspective, the designer of a copyright authorization schemes with regard to software robots may have two options:

In light of the previous section, in different countries, the copyrights with respect to licensing, like modification, distribution and reproduction rights, almost have the same legal meanings. On this ground, the first one option is only dealing with the minimum copyright rights which provided in the international conventions. This approach may meet the basic requirement of a copyright authorization scheme, but cannot satisfy the needs in some complicated situations. Another main drawback of this approach is that legal interpretation is still inevitable while cross board conflicts appear.

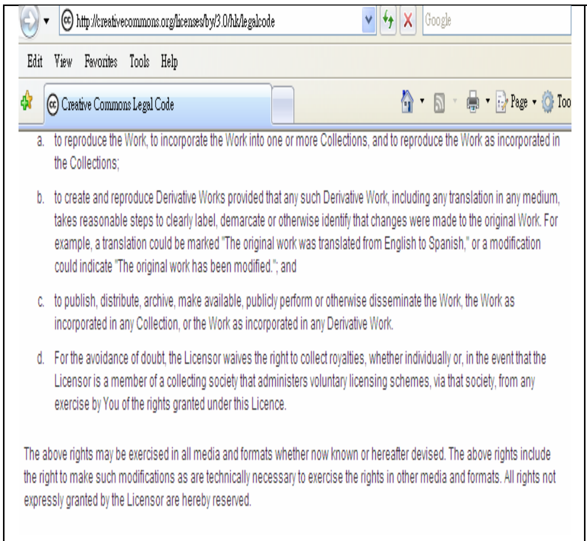
Another approach, on the other hand, is giving up providing a solid tool, rather, trying to provide a distinct license in respect of different jurisdictions. The CC license adopts the second approach. In fact, it tries to use different licenses or legal terms in different countries to port 6 basic licenses the various licenses to accommodate local copyright and private law. For example, the legal codes of the same Attribution license are quite different in Hong Kong and England, as shown in Table V. To sum up, through different legal codes to substantially explain the real licensing scope, it can generally be said that the CC license framework provide a set of relatively good tools with regard to fully expressing diverse copyright authorization scopes.

Table V: Different legal code of the same Attribution license in two jurisdictions

Legal Code in Hong Kong[18]	Legal Code in England[19]

[18] http://creativecommons.org/licenses/by/3.0/hk/legalcode

[19] http://creativecommons.org/licenses/by/2.0/uk/legalcode


http://creativecommons.org/licenses/by/3.0/hk/legalcode
Google
Edit View Favorites Tools Help
Creative Commons Legal Code
Page
Too

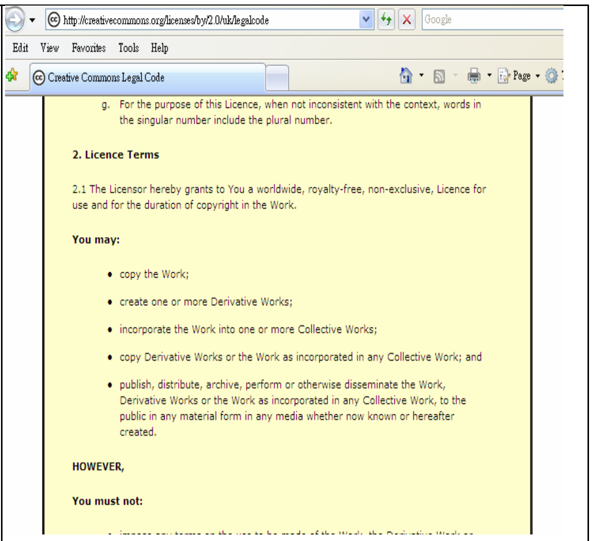
a. to reproduce the Work, to incorporate the Work into one or more Collections, and to reproduce the Work as incorporated in the Collections;

b. to create and reproduce Derivative Works provided that any such Derivative Work, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified."; and

c. to publish, distribute, archive, make available, publicly perform or otherwise disseminate the Work, the Work as incorporated in any Collection, or the Work as incorporated in any Derivative Work.

d. For the avoidance of doubt, the Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this Licence.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by the Licensor are hereby reserved.


http://creativecommons.org/licenses/by/2.0/uk/legalcode
Google
Edit View Favorites Tools Help
Creative Commons Legal Code
Page

g. For the purpose of this Licence, when not inconsistent with the context, words in the singular number include the plural number.

2. Licence Terms

2.1 The Licensor hereby grants to You a worldwide, royalty-free, non-exclusive, Licence for use and for the duration of copyright in the Work.

You may:

- copy the Work;
- create one or more Derivative Works;
- incorporate the Work into one or more Collective Works;
- copy Derivative Works or the Work as incorporated in any Collective Work; and
- publish, distribute, archive, perform or otherwise disseminate the Work, Derivative Works or the Work as incorporated in any Collective Work, to the public in any material form in any media whether now known or hereafter created.

HOWEVER,

You must not:

3.1.3 How to license and mark works with CC license

After choosing one of the six different CC licenses, the next, and most technical step is adopting appropriate ways to mark the work to let others understand which license has been chosen and, what the scope of authorization is. The methods of makers are various in respects of the types of works.

The most common way is that a CC marker, a line or graphic stating CC license, should be on the work or papering somewhere near the work, such as embedded in a Website to indicate that all works in this website are under CC licensed[20]. An ideal CC marker should contain the Commons Deed and a full URL[21]. A full URL is necessary as the Deed can not show the specific jurisdiction of the license. This general method are almost suitable for any type of works, including text, image, audio, video files and, even physical medias[22]. An example of a CC maker is as shown in the following figure:

[20] http://creativecommons.org.tw/static/technology/webpage

[21] http://wiki.creativecommons.org/Marking#Crediting_in_Images

[22] http://wiki.creativecommons.org/Marking_Audio

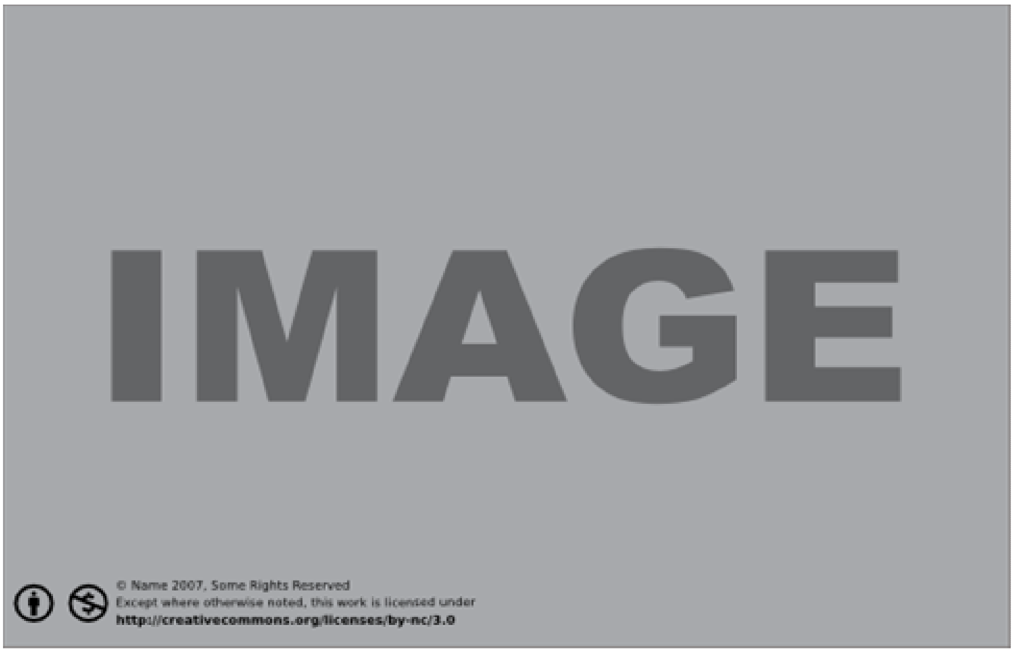


IMAGE
© Name 2007, Some Rights Reserved
Except where otherwise noted, this work is licensed under
http://creativecommons.org/licenses/by-nc/3.0

Figure 5: A sample image with the CC license marker[23]

On the other hand, there are several other various ways for different types of works. For example, for audio works, a brief sound clip, or an “audio bumper”, at the beginning or end of the work, consisting of the name of the license, the full URL link to licesne and a copyright notice stating the author’s name, date and copyright information is also an effective way[24]. A video bumper is a visual notice, which often is embedded at the beginning or end of the video work which includes the similar information of the audio bumper[25]. Moreover, as to longer plain text works, a full segment of legal code embedded within the work can replace the combination of the common deed and a full URL[26].

But we have to keep in minds that there are three ways to state a same CC license: the deed, the legal code and the digital-code. In general, the digital-code of the CC license takes the form of HTML tags embedded in the body of a CC licensed

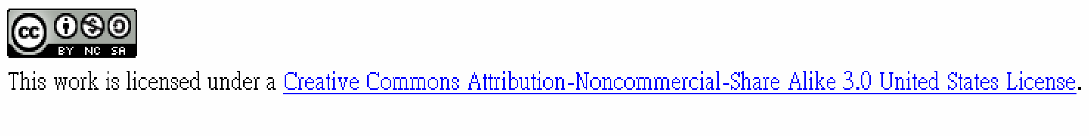
[23] http://wiki.creativecommons.org/Marking_Image
[24] http://wiki.creativecommons.org/Marking_Audio
[25] http://wiki.creativecommons.org/Marking_Vedio
[26] http://wiki.creativecommons.org/Marking_Text

document[27]. The following example shows a section of the digital-code for the “Attribution-Noncommercial-Share Alike” CC license:

```
<a rel="license"
href="http://creativecommons.org/licenses/by-nc-sa/3.0/us/">
<img alt="Creative Commons License" style="border-width:0"
src="http://i.creativecommons.org/l/by-nc-sa/3.0/us/88x31.png" />
</a>
<br />This work is licensed under a
<a rel="license"
href="http://creativecommons.org/licenses/by-nc-sa/3.0/us/">Creative
Commons Attribution-Noncommercial-Share Alike 3.0 United States
License</a>.
```

In the upcoming codes, “by-nc-sa” in the link http://creativecommons.org/licenses/by-nc-sa/3.0/us/ expresses that this webpage is authorized under the CC license “Attribution, No Commercial Use and Share Alike” condition. A webmaster can directly embed this segment of HTML codes in his page or website to illustrate the authorization scope of the works in this site.

The following figure demonstrates a part of a Web page where the segment of above code is embedded:



This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License.

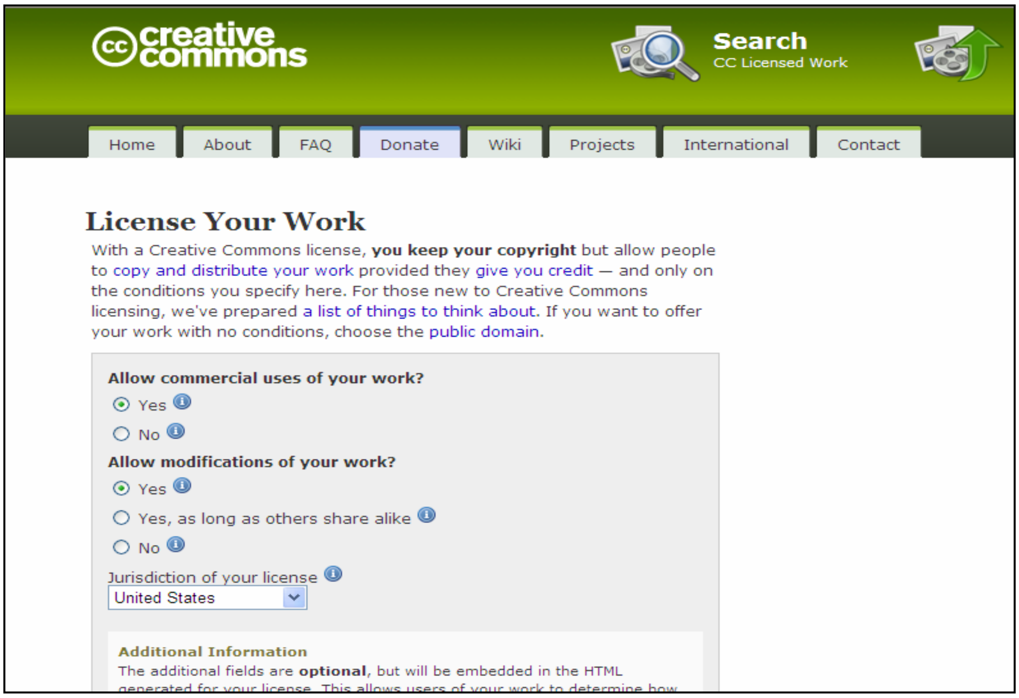
Figure 6: A part of a Web page containing identical licensing information

Explicitly speaking, the above part of a page can be divided into two parts: the first one is the deeds and the second one is an URL which links to the Web page containing all essential information of the specific license, including the meanings of the deeds and a link pointing to the legal code. For a robot, the second part is more important, because it contains all necessary licensing information. We will explain this point in the section 4.4.2.

[27] http://wiki.creativecommons.org/Frequently_Asked_Questions#What_is_the_Commons_Deed.3F_What_is_the_legal_code.3F_What_does_the_html.2Fmetadata_do.3F

However, the syntax of digital-code in CC is too complex for people to write. Based on the sample of digital-code above, non-programmers will be baffled by the syntax in the code snippet. In fact, CC's designers are aware of this issue. The CC license's Web site provides a user-friendly tool that can generate the needed digital-code. Once the code has been produced, an author needs to cut and paste the generated digital-code into their files. The syntax of the CC license code is meant for an indexing subsystem, not a human. Some humans may be uncomfortable with the extra step the CC license system requires to place the needed instructions in a document file.

In order to overcome the difficulty above; in fact, the CC website offers users a simple tool to generate the digital-code:



License Your Work

With a Creative Commons license, **you keep your copyright** but allow people to copy and distribute your work provided they give you credit — and only on the conditions you specify here. For those new to Creative Commons licensing, we've prepared a list of things to think about. If you want to offer your work with no conditions, choose the public domain.

Allow commercial uses of your work?

● Yes

○ No

Allow modifications of your work?

● Yes

○ Yes, as long as others share alike

○ No

Jurisdiction of your license

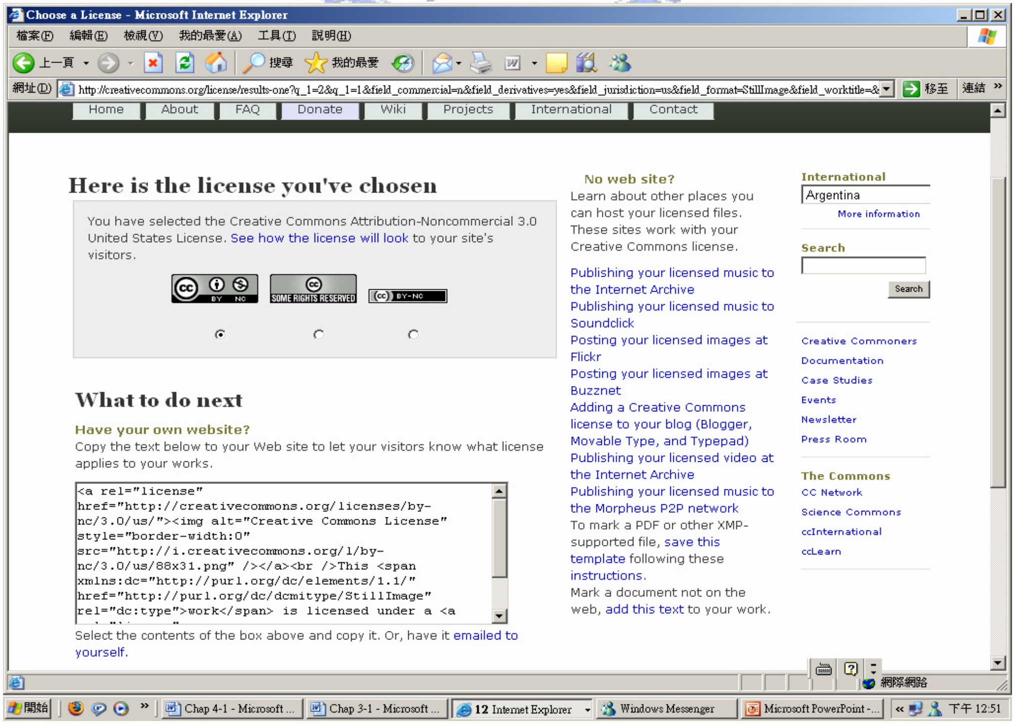
United States

Additional Information

The additional fields are **optional**, but will be embedded in the HTML generated for your license. This allows users of your work to determine how

Figure 7: A tool offered by the CC website to generate digital-code[28]

As shown in the above figure, a page owner has to answer three necessary questions: the former two are about the authorization conditions and the third one is about the jurisdiction. The following figure is the resulting page where the page owner can find out the generated digital-code and further instructions which teach the owner how to embed the code in his page. .



http://creativecommons.org/license/results-one?q_1=2&q_1=1&field_commercial=n&field_derivatives=yes&field_jurisdiction=us&field_format=StillImage&field_worktitle=&

Here is the license you've chosen

You have selected the Creative Commons Attribution-Noncommercial 3.0 United States License. See how the license will look to your site's visitors.

● ○ ○

What to do next

Have your own website?

Copy the text below to your Web site to let your visitors know what license applies to your works.

```
<a rel="license"
href="http://creativecommons.org/licenses/by-
nc/3.0/us/"><img alt="Creative Commons License"
style="border-width:0"
src="http://i.creativecommons.org/l/by-
nc/3.0/us/88x31.png" /></a><br />This <span
xmlns:dc="http://purl.org/dc/elements/1.1/"
href="http://purl.org/dc/dcmitype/StillImage"
rel="dc:type">work</span> is licensed under a <a
```

Select the contents of the box above and copy it. Or, have it emailed to yourself.

No web site?

Learn about other places you can host your licensed files. These sites work with your Creative Commons license.

Publishing your licensed music to the Internet Archive
Publishing your licensed music to Soundclick
Posting your licensed images at Flickr
Posting your licensed images at Buzznet
Adding a Creative Commons license to your blog (Blogger, Movable Type, and Typepad)
Publishing your licensed video at the Internet Archive
Publishing your licensed music to the Morpheus P2P network

To mark a PDF or other XMP-supported file, save this template following these instructions.

Mark a document not on the web, add this text to your work.

International

Argentina

More information

Search

Creative Commoners
Documentation
Case Studies
Events
Newsletter
Press Room

The Commons

CC Network
Science Commons
ccInternational
ccLearn

[28] http://creativecommons.org/license/?lang=en_US

Figure 8: The results of CC digital-code generator[29]

3.2 The Robots.txt and Robots Meta tags with regard to Copyright Authorization Expression

The Robots.txt and Robots Meta tags were both proposed in 1990s. The Robots.txt is also called the “Robots Exclusion Protocol” (Snyder and etc. 1998), “Robot Exclusion Standard” (Koster, 1995) 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 the 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 websites’ are concerned, research indicates that, in 2001, around 40% of the websites owned by the global high-rank companies adopted the Robots.txt and Robots Meta tags (Drott, 2002).

3.2.1 Introduction of Robots.txt

The Robots.txt is a file which 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 following examples illustrate several common uses of the Robots.txt:

Table VI. Examples about Robots.txt

	Examples	Meaning
1	User-agent: * Disallow:	Allow all robots complete access

[29] http://creativecommons.org/license/work-html-popup?license_code=by-nc

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.

3.2.2 The introduction of Robots Meta tags

Sometimes, the page creators do not administer their own websites. For example, a staff member in a university creates his personal webpage on the website 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 improved 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 follows:

Table VII. 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

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.

3.2.3 Two functions of Robots.txt and Robots Meta tags

3.2.3.1 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 websites 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.

3.2.3.2 The new function: expressing online copyright authorization

Apart from mere advice, based on a recent noticeable US federal case, Field v. Google, Inc. (Field, 2006), 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 websites (Field, 2006). The plaintiff, Mr. Field, who posted 51 copyright works on his website 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” (Field, 2006) 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 (Field, 2006). As a predictable result, Google routinely used its software robot, GoogleBot (Google, 2008a), to retrieve the plaintiff’s website, 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" (Field, 2006). 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 (Field, 2006).

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 websites 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 (Copispresse, 2007), 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, 2007).

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 wish 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 website 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 website 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 wish about what kind of robots are allowed, what are excluded and what kind of links should not be followed.

However, 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 etc., 2003) and, as a result, sporadic amendment proposals are based on personal experience rather than general principles (Conner, 1996; Koster, 1994).

3.3 Uniqueness of Robots.txt and Robots Meta Tags

The Robots.txt and Robots Meta tags are in connection to the Web pages, rather than to a specific work; that is to say, the connection between the copyright authorization information and the work is not “portable” and, as a result, the authorization information may easily be lost during the course of dissemination and transformation.

Nevertheless, on the other hand, based on the following two grounds, the use of the Robots.txt and Robots Meta tags to express authorization scope should not be overlooked: Firstly, the Robots and Robots Meta tags are much simpler than the CC license, in terms of the fact that one line of the Robots and Robots Meta tags can be used to express the authorization information of all works in one website. Secondly, in light of the Field case above (Field, 2006) and the eBay, Inc. v. Bidder's Edge, Inc case (eBay, 2000), under some specific circumstances, the absence of the Robots.txt and Robots Meta tags can also confer some kind of facts, for instance, allowing the access from the software robots. That is to say, at least, the Robots.txt and Robots Meta tags can be a supplementary tool to adopt by the webmasters or authors to express their implicit will. Thirdly, based on the eBay, Inc. v. Bidder's Edge, Inc case (eBay, 2000), the Robot.txt and Robot Meta tags are not only for expressing copyright authorization scope, but can be used as a instrument by webmasters to avoid trespass as well. For instance, Yahoo and MSN both support “crawler-delay” directive to delay the robots (Yahoo, 2008d; MSN, 2008b).

Based on the above reasons, even the Robots.txt and Robots Meta tags is somewhat “weak” with regard to copyright authorization, it can still be seen that a more comprehensive set of the Robots.txt and Robots Meta tags can help webmasters to manage the access from software robots and as a result, reduce the possible conflicts. In the next sections, the disadvantages of the current Robots.txt and Robots Meta tags are be discussed and a new amendment is proposed.

3.4 Few Deficiencies of Robots.txt and Robots Meta tags in Respect of Copyright Authorization Expression

3.4.1 Some uncertainties with respect to 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.

3.4.1.1 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 the 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 the wish 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.

3.4.1.2 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 copies 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.

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

In the above section, 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 4.3, 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 3.3.1.2.

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 the following table:

Table VIII. Possible Copyright Infringement caused by robots and the tags

	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

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 website, 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 website 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, 2003b). 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 the wish of the authors in respect of reproduction and nothing of adaptation. In addition, as for distribution of the work in the third “distributing” step, the wish expressed in this tag is ineffectual in resolving the potential infringements resulted from ambiguous authorization scope.

3.5 Adding Tags to Fully Express Copyright Authorization Scope and Dismiss Ambiguous Old 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, 2003a). The two new tags are shown as follows:

Table IX. 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

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

Table X. Examples about 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

4. Expressions for Licensing Individual Works

Other than identically licensing by the webmasters, all works resides in a page can be separately authorized by its authors. In other words, under some circumstances, authors may need to individually authorize their works subject to totally different conditions and scopes, rather than authorizing under the same conditions. For example, in Flickr (Flickr, 2004), every uploader or author has an opportunity to license the works under one of the six CC license choices. We will introduce expressions to license individual works in this section.

4.1 Introduction

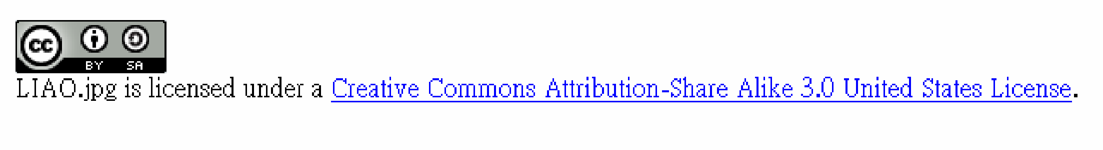
In the foregoing sections, we have discussed the reasons why the CC licensing scheme is a proper choice to demonstrate the copyright authorization scope of identical work. In this section, we will explicitly examine three possible ways of adopting CC license in terms of licensing a work: showing the part of licensing information in the Web page, embedding the information in the body of the file and the CCFE, a method of storing licensing information in the name of a file to reduce the disadvantages of the former two.

4.2 Showing CC Licensing Information of Work in Part of Website

In addition to expressing the identical CC licensing information of all works, the CC license scheme can also be used to individually indicate the copyright authorization scope of each work: the author embeds the digital-code into a host HTML or XML file (Flickr, 2004). A software robot “reads” the host file as well as accesses to the binary file pointed to in the HTML or XML “wrapper”. In fact, the CC website offer authors a simple tool to generate the license codes. To specify the licensing information to a specific work, other than the three questions mentioned in section 3.1.1, the next important step is pointing out the name of the authorized work and the website will automatically generate a segment of HTML code which can be embedded in the Web page. For example, the codes about a file, “LIAO.jpg”, licensed subject to the CC ”by-sa” license is:

```
<a rel="license"
href="http://creativecommons.org/licenses/by-sa/3.0/us/">
<img alt="Creative Commons License" style="border-width:0"
src="http://i.creativecommons.org/l/by-sa/3.0/us/88x31.png" /></a>
<br /><span xmlns:dc="http://purl.org/dc/elements/1.1/"
property="dc:title">LIAO.jpg</span> is licensed under a <a rel="license"
href="http://creativecommons.org/licenses/by-sa/3.0/us/">Creative
Commons Attribution-Share Alike 3.0 United States License</a>.
```

The page containing the above segment of HTML code is shown as follows:



LIAO.jpg is licensed under a Creative Commons Attribution-Share Alike 3.0 United States License.

Figure 9: A part of a Web page containing licensing information of “LIAO.jpg”

Any viewer as well as software robot can easily understand the “LIAO.jpg” is licensed under the “by-sa” CC license. Even this licensing method is simple and direct, however, we have to notice that the licensing information is separated from the file itself; as a result, how to mapping the right licensing information to the specific work is a real problem. Moreover, the licensing information any be lost during the course of transformation and. anyone who downloads the “LIAO.jpg” can distribute the file without any licensing information or, even worse, with faked licensing information.

4.3 Embedding CC Licensing Information in Body of a File

In order to avoid the unwelcome condition arose by the separate CC licensing information; the next option is directly embedding the CC licensing information in the body of the file. In fact, the CC license framework supports authors to embed license information in digital files[30] (Creative Commons, 2009p); nevertheless, as we have seen, CC has not supported “any digital format”. Generally, anchored in hypertext

[30] http://wiki.creativecommons.org/images/6/61/Creativecommons-licensing-and-marking-your-content_eng.pdf

markup language (HTML) and its variants, the digital-code of the CC license is text and embedding text in a binary file is problematic. The file may be corrupted, or a conversion process is needed which adds another manual step to the process (Creative Commons, 2009p; Wikipedia, 2007a). Therefore, the CC license does not address the issue of placing the instructions in an audio, video or Microsoft PowerPoint files. The method to resolve this problem is an author may use a software program such as CcPublisher or XMP, to embed license information within MP3, PDF or some binary files are developed (CcPublisher, 2007; Creative Commons, 2009q) to embed CC license declaration segment into the body of a binary file. The obvious problem with this method is that the author must handle the additional processes manually.

Even as problematic as the hurdles put in the path of the author wanting to use the CC license for binary files, accordingly, indexing robots launched by Google, Microsoft, and Yahoo, among others, are not programmed to recognize the CC license embed in a binary file. Today, these robots are "blind" to embedded CC license information. Of course, it is possible to design a search engine which can process the license description in the host HTML files. Commercial search engines have certain priorities, and, at this time, adding support for embedded CC license data in binary files is not an urgent matter.

4.4 Storing CC Licensing Information in Name of a File and CC File Extension Protocol—CCFE

4.4.1 Which attached part to a file is proper to store CC licensing information?

Due to the lack of a general method of embedding the CC licensing information in the body of the file, some technical problems occur, especially in respect of indexing and searching of CC licensed works. To solve this problem, finding out another part of a file to store the licensing information is a reasonable approach. More importantly, to keep the most obvious advantage of embedding the licensing information in the body of the file, the information stored in this part should be "portable" as well.

A reasonable starting point is examining general features of several popular file systems (Wikipedia, 2007a) and trying to find out their common parts. After some general overviews, we may easily make a conclusion that the file name and some metadata, such as the file length, the last modified time and etc., are the most common parts in the directories of common file systems. However, from the author's points of view, file name is only the part which a user full authority to modify. On the contrary, in general, users have no rights to change the file length, the last modified date and other similar metadata of a file. As a result, the file name is a reasonable suggestion to store the CC licensing information of a file.

Considering whether the licensing information will be lost in the course of distribution, in addition to the file system, the transportation protocol should be the next concern. To simply verifying the above suggestion, we may examine Hypertext Transformation Protocol (HTTP hereafter) (Fielding and etc., 1999), the most popular transportation protocol on the Web. Generally speaking, a file name, as a part of the requested URL, is an essential field in the typical GET commend. That is to say, the requester must firstly access to the real name of the file which he wants to download. In fact, in respect of downloading a file, getting the full name of the file is one of the most necessary steps, since the file name indicates the final and specific position of the downloading target. For example, a simple GET message requesting a "LIAO.jpg" in the host "ccc.kmit.edu.tw" as shown as follows:

```
GET /LIAO.jpg HTTP/1.0
Accept: image/gif, image/jpeg, application/msword, */*
Accept-Language:zh-tw
User-Agent: Mozilla/4.0
Content-Length:
Host: ccc.kmit.edu.tw
Cache-Control: max-age=259200
Connection: keep-alive
```

On the above simple verifications, storing the licensing information in the file name is a proper suggestion, because the author has a general authorization to modify it and the information contained in it will not be lost in the course of distribution on the Web.

4.4.2 The essential elements of CC licensing and how to express them

After deciding where to store the licensing information, the next things we have to consider are what kinds of licensing information should be stored and how to express them. In the subsection 4.2, we have seen that, in order to license a work under the CC license, there are four queries to be answer. In respect of their goals, these four queries can be categorized into three groups: the answers of the first two queries determine which one of the six licensing options is chosen and the third query, about the jurisdiction, determine the laws of which nation should be followed. The last one, about the title of the work, indicates which specific work is licensed. A set of answers of the four queries of these three groups decide the final copyright authorization scope under the CC licensing. However, while examining the generated codes, as shown in section 4.1, another element, the version of the license, which indicates the exact version of the various licenses in the same name, is necessary as well as other essential elements. In fact, the version number eliminates the possible confusion involving in the CC licenses of the same license name. For example, the present US “by-sa” CC license is version 2.5. The content of every version of the US “by-sa” is slightly different, especially from the legal professions’ perspectives. As a result, without the version number “2.5”, a licensee may make a mistake in deciding what the permissible actions are as he may wrongfully believe that the work is licensed under the version 1.0 of the US “by-sa” license.

To sum up, the licensing option, the jurisdiction, version of the license and the name of the work consist of the four essential elements of expressing the specific CC licensing authorization scope of a specific work. Therefore, in respect of storing the CC licensing information in the file name, excepting the file name itself, the other three elements, the licensing option, the jurisdiction, and the version should be definitely contained in the file name in a suitable form.

After deciding the essential elements of the licensing information, the next concern is how to express them. Considering the simplicity and the limit of the length of the file name in some file systems, a short but explicit expressing form is a reasonable choice. On this ground, we adopt a kind of abbreviations to express each element. The first element, the licensing option is much simpler to deal with than the others, since a set of official abbreviations is available, as shown in Table III. With regard to the jurisdiction, the same abbreviations as the name of the path in the full URL, For example, in the " http://creativecommons.org/licenses/by-sa/3.0/us/", the path name "us" represents the USA; "tw" represents Taiwan. The last element, the version is much easier; we directly use the version number to indicate what the exact version is.

4.4.3 CC File Extension Protocol—CCFE

After figuring out what the essential elements are and how to express them in respect of storing the CC licensing in the file name. We propose a new protocol, the CC File Extension Protocol (CCFE hereafter) to express the copyright authorization scope of the specific work. In the CCFE method, the CC license is embedded into the file name, not the body of the file. The method applies to any file type, not just binary files. Both people and programs can read the file extension and access the CC license information. The CC license information is therefore preserved in the process of duplication and transmission. The method addresses the problem of license portability.

The CCFE extends each file name with a CC license part. The CC license part is a file extension with the ".CC.<conditions_list>" syntax. The <condition list> of a CCFE file contains all essential information a CC license needs: the version, jurisdiction and one of the six possible licensing choices.

The <condition_list> part in CCFE is a list of conditions shown in Table VI. Each condition is separated with an underline "_".For example, the string ClintonDebate.cc.tw3.by_nc_sa.mpg makes explicit the author's intentions regarding "subject to the Taiwan CC licenses, version 3, the author allows altering, transforming, or building upon this work, only distributing the resulting work under the same or similar license as well as attributing the work to the author; but commercial use is disallowed". Table IV shows additional examples of CCFE file names.

Table XI. Examples of CCFE file names and their meanings

Original File Names	CC license represented	CCFE File Names
Article1.htm	Attribution	Article1.cc.tw3.by.htm
Article2.pdf	Attribution Share Alike	Article2.cc.tw3.by_sa.pdf
Article3.jpg	Attribution No Commercial Use	Article3.cc.tw3.by_nc.jpg
Article4.mpg	Attribution No Derivative Works	Article4.cc.tw3.by_nd.mpg
Article5.mp3	Attribution No Commercial Use Share Alike	Article5.cc.tw3.by_nc_sa.mp3
Article6.ppt	Attribution No Commercial Use No Derivative Works	Article6.cc.tw3.by_nc_nd.ppt

The only limitation of our method is that the CCFE only works on file systems which support a long file name. Since there are only two special symbol used in CCFE: the period symbol "." and the underline symbol "_", our method is compatible with most widely spread operating systems, such as UNIX, Linux, Apple OS X, and Microsoft Windows. Furthermore, CCFE may also be inserted into a URL without an encoding step.

Table XII. The syntax of the popular file naming systems, URL and CCFE

Operating System	File System	Max Filename Length	Allowable characters in file name

Linux	ext2,ext3, ext4	255	Any byte except NUL
Windows 95b	FAT32	255	Any Unicode except NUL
Windows NT	NTFS	256	Any Unicode except NUL, " / \ * ? < > | :
FreeBSD	UFS1/UFS2	255	Any Unicode except NUL
Mac OS X	HFS+	255	Any valid Unicode
URL	URL	N/A	Any ASCII except <>#"{ }|\^~[]` and nonprintable character (0-1F and 80-FF)
CCFE	CCFE	30	cc.{A-Za-z\._}[0-27]

Table VII provides several examples of the method in the syntax of popular file systems. First, the maximum length of the CC condition element in CCFE is only 30 letters (27 for the combination of all conditions and three for the heading “cc.”). There are some variations in the maximum length of file names in each operating system (Wikipedia, 2007a). For most modern operating systems, the maximum file length is longer than 255. We believe that the CCFE extensions require a modest part of the filename of these operating systems. URLs have no file name length restriction in the RFC1738 standard. Most browsers and web servers have some restrictions in practice, but the maximum filename length is always longer than 255 (RFC1738, 1994). So the CCFE filename length is not material in the Internet environment. Second, the last column in table V shows the allowable characters in directory entries. Most modern file systems allow dot and underline symbol and most operating system allow dot and underline symbol in the interface of file system. There is no prohibited letter in CCFE. CCFE is, therefore, compatible with Linux, Microsoft Windows, BSD, and Internet file and path naming conventions.

5 Conclusions

The main objective of this study is to investigate the expressions of copyright authorization used for automatically acquiring free Internet materials. The two new expressions, CCFE and amended Robot.txt and Robots Meta tags, may contribute to this area. Based on these improvements and characteristics rendered by these two schemes, some interesting comparisons and positions are exhibited as follows.

5.1 Comparisons

5.1.1 The comparison of Robot.txt and Robots Meta tags and CC licensing scheme in respect of identically licensing all works

As we have shown in the above sections, the Robots.txt and Robot Meta tags as well as the CC licensing scheme can both adopted by the webmasters to assign the authorization scopes of all works in the Web page or website. In other words, through these two methods, webmasters can easily use a short segment of code or a few tags to license all works in the same site under the same condition, rather than awkwardly repeating the licensing condition of each work. Nevertheless, these two schemes are different in the following aspects:

At first, the CC licensing scheme can express more specific rights within one term "copyright": including the modification, distribution and reproduction rights. On the other hand, the Robots.txt and Robots Meta tags are restricted to each step in the process of software robots, consequently, the expansion edition of the Robots.txt and Robots Meta tags merely represent the copyright authorization related to each step of software robots, rather than the specific rights embraced in copyrights. Secondly, the users of the CC licensing scheme can choose the jurisdictions; on the contrary, the users of the Robots.txt and Robots Meta tags have no free to choose the laws they want to follow. As a result, it can be said that the later one, the Robots.txt and Robots Meta tags, is easier to use but may arise a few of uncertainties in some situations.

In addition to the differences above, the other obvious difference between these two licensing methods is that, based on their basic features, the Robots.txt and Robot Meta tags is only visible to the software robots, but invisible to the viewers. On the other hand, the CC licensing information in HTML form can be seen by viewers and, this part of page illustrates all essential elements to specify the copyright authorization scope, including the deeds and the necessary URLs. In fact, such invisibility of the Robots.txt and Robot Meta tags actually offers the viewers who confront copyright infringement allegations a reasonable defense as they can argue that they have not see any copyright notice in the page under a general circumstance.

The following table summarizes the differences between the Robots.txt and Robot Meta tags as well as the CC licensing scheme in terms of expressing authorization scope of all works in one page or site.

Table XIII. The differences between the two approaches in respect of identically licensing all works

	Authorization Scope	Visible	Jurisdiction
The Robots.txt and Robot Meta tags	Bound to robots' accessing steps	Invisible to human viewers	Unspecific
The CC licensing Scheme	Specific copyright authorization scope	Visible to human viewers	Specific

5.1.2 The comparison of showing licensing information in page, embedding information in body and storing information in filename (CCFE)

As we mentioned above, the present CC licensing scheme offers users a method to license their works host in a Web page. A software robot "reads" the host file and then accesses to the binary file pointed to in the HTML or XML "wrapper". This method is quite easy and no further tool is needed. However, the method around creates a problem with what we call "portability". When a user copies the binary file, the wrapper or host file can be detached, thus the CC license is disconnected from the binary component. In other words, the licensing information showing the page is apart from the file itself; that is to say, the right licensing information is hard to locate and, even worse, this information may be lost in the course of transformation.

To resolve the foregoing problem, two approaches can be adopted: the first idea is embedding the CC licensing information in the body of the file itself. This method is practical for text files; however, there are no general tools to embed the licensing information into binary files. The authors must handle the additional processes manually. The second one is trying to store the CC licensing information in other attached parts of a file. After analyzing several common parts of a file in the existing popular file systems, we finally propose CCFE, which suggests users to store the information in the name of a file, because the file name is not only a part which general users have authorities to modify, but a necessary part of the HTML messages as well.

Even we propose the CCFE to store the CC licensing information in the name of a binary file, we can not ignore that this innovative approach may still confront several challenges, in respect of technology and legal aspects separately:

From the technology aspect, compared to the XMP and other similar tools, a CCFE file exposure the authorization information to all users on the Internet. In other words, any one who can access to this file has an opportunity to modify the authorization information contained in the filename and, then distributes the file with faked authorization information. On the contrary, the information embedded in the body of a binary file has a chance to implement higher security.

From the legal aspect, the first challenge is that, comparing to the original digital-code, the licensing information containing in the condition list of CCFE is more unspecific. The most obvious point is that the condition list does not provide a URL linking to the Web page containing the specific CC license. That is to say, a user unfamiliar to the formats of CCFE may argue that he or she has no chance to access to the specific legal contents of the CC license; as a result, the user may have a possible defense to against the copyright infringement allegations under this circumstance. The easiest technical method to avoid this unwelcome scenario is offering a tool, such as a browser, to help users to understand the authorization scope contained in the filename. Moreover, the best and most useful way is improving the file system or the OS. When a user opens or deals with a CCFE file, the OS can simultaneously alert the user and provide a clear explanation of the licensing information to the user. Accordingly, a user who uses an OS with this function has no chance to argue that he or she does not comprehend the authorization information contained in the filename.

5.2 Implications of the two methods

5.2.1 The implication to Internet library creator

Through these two methods, a digital library creator can automate the acquisition of licensed files, confident that no copyright violations will be inadvertently made. Librarians can effectively and correctly categorize these collected documents by the rights explicitly granted by the authors. Furthermore, CCFE as well as the amended Robots.txt and Robots Meta tags allow online multimedia digital libraries to be assembled with text, images, and other information objects, as these two approaches can applied to all file regardless of the file types. .

Moreover, the methods proposed in this thesis can help people to easily construct a personal Internet library, because through the proposed methods, users can use general purpose search engines to search and collect works only fulfilled copyright conditions. Actually, in the past, the designers of search engines pay most of their attention on returning relevant results quickly (Brin et al, 1998; Yang et al 2007). General purpose search engines have not been designed to meet the needs of professional researchers. Sometimes specialists need to search for specific topics and may have to locate documents suitable for inclusion in a digital archive, inclusion in a collection, distribution to students, or modification. For example, architects working for a construction company may need images of other structures to include in a brochure about a new building. Ideally, an architect could use a search engine like

Microsoft Live or Google to locate images. However, to locate a CC licensed pictures which permit commercial use, Web search engines are almost useless. Nevertheless, if CC license conditions are embedded in the CCFE file name as we propose, the architect can use existing search engines' advanced search functions to limit the query to images and other binary files with a CC license. As a result, an Internet library constructor who has no special knowledge with software robots can easily create an Internet library.

To sum up, these two new mechanisms, the CCFE and the amendment of the Robots.txt and Robots Meta tags, can be quite useful to the ones who want to build up an Internet library, especially who has no great capital. Because these two tools can help, with the spreading trends of open source and free culture (Lessig, 2004), library constructors to collect plenty works on the Internet which fulfill the copyright authorization conditions; otherwise expensive negations and agreements with authors are the only effective ways toward huge amounts of collections.

5.2.2. The implication to copyright law—the scope of fair use

Like Professor Lessig points out, the CC will encourage further uses of the works and effectively extend the scope of "fair use" in the Internet world (Lessig, 2005a). On the same ground, the promotion of CCFE and the new Robots.txt and Robots Meta tags will enlarge the realm of fair use as well, because these two approaches just keeping all advantages of the original versions but clarifying them. Moreover, the wide spreads of the CCFE and the Robot.txt and Robot Meta tags will reduce the needs of DRM to some extents, since authors can adopt these two schemes to license their works on the Internet rather than adopting rigid tools of DRM. Accordingly, less need for harmful DRM will really increase the possibility of use of the "fair use" defense.

5.2.3 The implication to free culture and open content movement

One of the goals of the methods proposed in this thesis is to encourage the open movement. In fact, due to the clearer copyright authorization scope, the CCFE and the

Robots.txt and Robot Meta tags can really help communications and distributions of works on the Internet. Moreover, authors who have freely shared works from others may tend to license his works under a kind of free licenses to contribute to the open movement.

5.2.4 The implication to the computer and information science researchers

As we mentioned earlier, code is one of factors which can affect the real world to some extents. In other words, not only law itself, but also code can resolve a legal problem. The two approaches, CCFE and the amended Robots.txt and Robots Meta tags, are good examples, because although they are designed on the knowledge of information science, both could be used to avoid possible copyright infringement allegations. Furthermore, the evolvement of the Robot.txt and Robot Meta tags, from pure advices to a legal mechanism to express the owner's will, may demonstrate a underlying fact that how courts' decisions may actually affect the developments of the IT technology. As a result, any IT researcher should pay more attention to the advents of IT laws.

5.3 Future Works and Further Suggestions

With the rapid growth of the Internet, an Internet library consisting of works on the Internet will play a more and more important role in knowledge dissemination and culture exchange and, software robots will inevitably become one of the most necessary and useful tools in organizing a successful Internet library. Consequently, identifying copyright authorization scope is quite essential, because this factor will actually decide the total amount of collections and, more importantly, the risk of copyright infringement litigations. 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 thesis 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.

Also, a librarian experienced with acquiring online digital materials will understand the scope and implications of our method of implementing a CC license. As we mentioned above, the implementation method of the current CC licensing framework is too complex to be widely used. CCFE attaches the license data via the file name itself. Most search engines can allow users to limit their queries to CC license files without any changes to their existing software or systems. Finally, CCFE works on text and binary files, a feature simply not supported by the present CC license method.

However, as we have seen in the previous sections, one of the most obvious features of the CC license scheme is that the user has a chance to choose the jurisdiction. In fact, this feature of the CC license may give arise to some ambiguities in the context of Internet as most of the distributions on the Internet is cross borders. For example, one works licensed under the US "by" CC license may arise some doubts about further modifications in the civil law countries which do more respects of the author's personality. There are two possible approaches to resolve this problem. The first one is organizing a group, like iCommons, an international voluntary organization consisting of legal experts (Lessig, 2005b), to ensure each CC license in various jurisdictions following the same basic rules. However, there is no systematic study to examine the equivalence of the present licenses in different jurisdictions. Another approach is to design a standard CC license, only including minimum rights embraced in the WIPO Copyright Treaty and other related international copyright conventions.

Moreover, 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 the CC license 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 author of copyrighted works within a website, such as the authors of video files on Youtube (Youtube, 2008) to express their complicated authorization scopes may form a new research direction. Moreover, apart from the copyright related subjects mentioned in this thesis, the great power of software robots now also arise some concerns about revealing personal privacy and data protection on the Internet (Thelwall and etc., 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.

References

1. ACM Digital Library (2009) Term of Usage: Digital Library, available at: http://www.acm.org/publications/policies/use (accessed July 3, 2007).
2. Bailey, C. W. (2006), “What is open access?”, available at: “http://www.digital-scholarship.org/cwb/WhatIsOA.pdf (accessed March 3, 2009).
3. BBC (2006), “BBC Creative Archive pilot has ended”, available at: http://creativearchive.bbc.co.uk/news/archives/2006/09/hurry_while_sto.html (accessed July 3, 2007).
4. Barker, P. (2001), “Creating the Digital Library. A Special Report from the Primary Research Group: Book Review” The Electronic Library, 19(3), 186-187.
5. Bolin, R. (2006), “Locking down the library: How copyright, contract, and cybertrapass block Internet archiving”, 33 Pepp. L. Rev. (2006) P761
6. Budapest Open Access Initiative (2002), "Budapest Open Access Initiative, 14 February 2002, available at: http://www.soros.org/openaccess/read.shtml (accessed March 3, 2009).
7. CcPublisher (2007), available at http://wiki.creativecommons.org/ CcPublisher (accessed Dec. 3, 2007)
8. ClikZ Network (2007), “U.S. Search Engine Rankings, September 2007”, available at http://www.clickz.com/3627655 (accessed Dec. 3, 2007)
9. Chau, M. and Chen, H. (2003), “Personalized and focused Web spiders”, in: Zhong, N., Liu, J., Yao, Y. (Ed.), Web intelligence, Springer-Verlag, pp.197-217.
10. Cheong, F.C., (1996), “Internet Agents: Spiders, Wanderers, Brokers, and Bots”, New Riders Publishing, Indiana, USA.
11. Citeseer.ist (1997) available at: http://citeseer.ist.psu.edu/ (accessed July 3, 2007).
12. CiteseerX.ist (2009) “Submit Documents to CiteseerX” available at: http://citeseerx.ist.psu.edu/submit (accessed Jan. 23rd, 2009).
13. Copiepresse (2007), “Copiepresse v. Google, Inc”., Copiepresse v. Google, Inc., No. 06/10.928/C (Feb. 2, 2007).
14. Copiepresse (2009), available at: http://www.copiepresse.be (accessed Feb 1st, 2009)
15. Conner, S. (1996), “An Extended Standard for Robot Exclusion”, available at: http://www.conman.org/people/spc/robots2.html (accessed March 11, 2008)

16. Cornish, W. and Llewelyn, D. (2003a), “Intellectual Property: Patents, Copyright, Trade Marks and Allied Rights”, Sweet & Maxwell, London, PP475
17. Cornish, W. and Llewelyn, D. (2003a), “Intellectual Property: Patents, Copyright, Trade Marks and Allied Rights”, Sweet & Maxwell, London, PP485
18. Creative Commons (2009q), “XMP”, available at http://wiki.creativecommons.org/XMP (accessed Feb. 3, 2009)
19. Digital Library Federation (1998), “A working definition of digital library”, available at: www.diglib.org/about/dldefinition.htm (accessed July 20, 2008)
20. Drott, M. C. (2002), “Indexing aids at corporate websites: the use of robots.txt and META tags”, Information Processing and Management, Vol 38 No2, pp209-219.
21. eBay (2000), “eBay Inc. v. Bidder’s Edge, Inc.”, 100 F. Supp. 2d 1058 (N.D. Cal. 2000).
22. Feigin, Eric J. (2004), “Architecture of Consent: Internet Protocols and Their Legal Implications”, Stanford Law Review, Feb 2004, pp901-942.
23. Feldman, S. (2004), “Interview: Brewster Kahle”, June 2004 Queue, Volume 2 Issue 4 P24-33
24. Field (2006), “Field v. Goolgle, 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).
25. Fielding, R. and Gettys, J. and Mogul, J. C. and Frystyk, H. and Masinter, L. and Leach, P.; Berners-Lee, T. (1999). "Hypertext Transfer Protocol - HTTP/1.1”, available at http://www.w3.org/Protocols/rfc2616/rfc2616.html (accessed March 11, 2009).
26. Flickr (2004), available at http://www.flickr.com/creativecommons/ (accessed July 3, 2009)
27. Free Software Foundation (1999), “The Free Software Definition”, available at http://www.gnu.org/philosophy/free-sw.html (accessed July 3, 2008)
28. Free Software Foundation (2006), “Digital Restrictions Management and Treacherous Computing”, available at http://www.fsf.org/campaigns/drm.html (accessed July 3, 2008)
29. Free Software Foundation (2007), “GNU General Public License Version 3”, available at http://www.gnu.org/licenses/gpl-3.0.txt (accessed July 3, 2008)
30. Google (2008a), “How Google crawls my site?”, available at: http://www.google.com/support/webmasters/bin/topic.py?topic=8843 (accessed March 11, 2008)
31. Google (2008b), “Preventing content from appearing in Google search results”, available at:

http://www.google.com/support/webmasters/bin/topic.py?topic=8459 (accessed March 11, 2008)

32. Google (2007c), "Google Advanced Search", available at http://www.google.com/advanced_search?hl=en (accessed July 3, 2007)
33. Gorman, G. E. (2006), "Giving way to Google", Online Information Review, Vol 30 Iss 2, pp97-99.
34. Hart, Michael S. (2004) "Gutenberg Mission Statement by Michael Hart", available at: http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Mission_Statement_by_Michael_Hart (accessed Jan. 3rd, 2009)
35. Hirtle, P. B. (2003), "Digital Preservation and Copyright", available at: http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hirtle.html (accessed Jan. 3rd, 2009)
36. Hundie, K. (2003), "Library operations and Internet resources", The Electronic Library, Vol 21 No 6 pg. 555-564
37. International DOI Foundation (2008), "Frequently Asked Questions about the DOI System: 1. What is a DOI® name?", available at http://www.doi.org/faq.html#1 (accessed March. 3rd, 2009)
38. Internet Archive (2009), available at: http://www.archive.org/index.php (accessed Jan. 3rd, 2009).
39. Issuu (2009a), available at: http://issuu.com (accessed Jan. 3rd, 2009).
40. Issuu (2009b), "Copyright FAQ", available at:http://issuu.com/about/copyright (accessed Jan. 3rd, 2009).
41. Jones, P. (2001), "Open (source)ing the doors for contributor-run digital libraries", Communications of the ACM Vol 44 Iss 5 pp 45-46.
42. Koster, M. (1993), "Guidelines for Robot Writers", available at: http://www.robotstxt.org/guidelines.html , (accessed March 11, 2008)
43. Koster, M. (1994)," A Standard for Robot Exclusion", available at: http://www.robotstxt.org/orig.html, (accessed March 11, 2008)
44. Koster, M. (1995), "Robot in the Web: threat or treat?", available at: http://www.robotstxt.org/wc/threat-or-treat.html, (accessed March 11, 2008)
45. Koster, M. (1996), "Evaluation of the standard for robots exclusion", available at: http://www.robotstxt.org/wc/eval.html, (accessed March 11, 2008)
46. Koster, M. (1997), "HTML Author's Guide to the Robots Meta tags", available at: http://www.robotstxt.org/wc/meta-user.html (accessed March 11, 2008)
47. LaMacchia, B. (2002), "Key Challenges in DRM: An Industry Perspective", Proceedings of the 2002 ACM Workshop on Digital Rights Management, Wellington. pp. 51-60

48. Lessig, L.(2004), "Free culture: how big media uses technology and the law to lock down culture and control creativity", Penguin Press, New York.
49. Lessig, L.(2005a), "CC in Review: Lawrence Lessig on CC & Fair Use", available at: http://creativecommons.org/weblog/entry/5681 (accessed March 11, 2009)
50. Lessig, L.(2005b), "CC in Review: Lawrence Lessig on iCommons", available at: http://creativecommons.org/weblog/entry/5700 (accessed March 11, 2009)
51. Lessig, L.(2006a), "Code: And Other Laws of Cyberspace, Version 2.0", Basic Books, New York. pp123
52. Lessig, L.(2006b), "Code: And Other Laws of Cyberspace, Version 2.0", Basic Books, New York. pp124
53. Lin, Y-H, Ko, T-M, Chuang, T-R, Lin, K-J (2006) "Open Source Licenses and the Creative Commons Framework: License Selection and Comparison", Journal of Information Science and Engineering, Vol 22 No 2 pp1-17.
54. Lopatin, L. (2006), "Library digitization projects, issues and guidelines :A survey of the literature", Library Hi Tech, Vol 24 No 2, pp273
55. McCray, A. T. and Gallagher, M. E. (2001), "Principles for digital library development", Communication of ACM, Vol 44 Iss 5 , pp48-54
56. Millard, C. (2007), "Copyright In Information Technology and Data", in: Reed, C. and Angel , J. (Ed.), Computer Law, Oxford University Press, pp.337-396.
57. 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, 2009)
58. MSN (2008b), "Limit crawl frequency", available at: http://search.msn.com/docs/siteowner.aspx?t=SEARCH_WEBMASTER_REF_RestrictAccessToSite.htm (accessed March 11, 2009)
59. National Science Foundation (1999), "Digital Libraries Initiative: Available Research, US Federal Government", available at: http://dli2.nsf.gov/dlione/. (accessed March 11, 2008)
60. Netcraft (2008), "Netcraft Web Server Survey", available at: http://news.netcraft.com/archives/web_server_survey.html (accessed July 9, 2008)
61. ODRL Initiative (2005), "ODRL Creative Commons Profile", available at: http://odrl.net/Profiles/CC/SPEC.html (accessed March 9, 2009)
62. ODRL Initiative (2009), "The international effort to develop and promote ODRL", available at: http://odrl.net/ (accessed March 9, 2009)
63. Open Content Alliance (2009) "Open Content Alliance: FAQ", available at: http://www.opencontentalliance.org/faq/ (accessed March 13, 2009)

64. Open Source Initiative (2006), “The BSD License”, available at: http://www.opensource.org/licenses/bsd-license.php (accessed March 9, 2009)

65. O'Leary, Mick (2009), “Open Content Alliance Embodies Open Source Movement”, Information Today, Vol 26 Iss 1, pp 37-38

66. Rao, S.S. (2003), “Copyright: its implications for electronic information”, Online Information Review, Vol 27 Iss 4, pp. 264-275.

67. Raghavan, S. and Garcia-Molina, H. (2001), “Crawling the Hidden Web”, Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Sep. 11-14, 2001, Rome, Italy, available at: http://dbpubs.stanford.edu:8090/pub/2000-36 (accessed July 3, 2007).

68. Rao, S.S. (2003), “Copyright: its implications for electronic information”, Online Information Review, Vol 27 Iss 4, pp. 264-275.

69. Reed, C. (2004), “Internet law: Text and Materials”, Cambridge University Press, Cambridge, pp71.

70. RFC1738 (1994), available at ftp://ftp.rfc-editor.org/in-notes/rfc1738.txt (accessed July 3, 2007)

71. Rosenblatt, B. (1997). "The Digital Object Identifier: Solving the Dilemma of Copyright Protection Online", Journal of Electronic Publishing, Vol 3, Iss. 2, pp135-156.

72. Samuelson, P. (2003), “Unsolicited Communications as Trespass?”, Communication of ACM, Vol 46 No 10, pp15-20.

73. Scribd (2009) available at: http://www.scribd.com/ (accessed Jan. 3rd, 2009).

74. Seadle, Michael. (2006), “Copyright in the networked world: using facts”, Library Hi Tech, Vol 24 No 3, pp. 463-468

75. Seadle, Mi and Greifeneder, E.(2007), ”Defining a digital library”, Library Hi Tech Vol. 25 No. 2, 2007 pp. 169-173

76. Sieman, J. S. (2007), “Using the implied license to inject common sense into digital copyright”, North Carolina Law Review, Vol 85, pp885-930.

77. Smith, G (2007), “Copiepresse v Google - the Belgian judgment dissected”, available at: http://www.birdbird.com/english/publications/articles/Copiepresse-v-Google.cfm ?RenderForPrint=1 (accessed July 2, 2008)

78. 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 453-462.

79. Spinello, Richard A. (2007), “Intellectual property rights”, Library Hi Tech, Vol 25 No 1, pp. 12-22

80. Sterling, J.A.L. (2003a), “World Copyright Law”, Sweet & Maxwell, London,

pp530

81. Sterling, J.A.L. (2003b), "World Copyright Law", Sweet & Maxwell, London, pp533
82. Sterling, J.A.L. (2003c), "World Copyright Law", Sweet & Maxwell, London, pp17
83. Sterling, J.A.L. (2003d), "World Copyright Law", Sweet & Maxwell, London, pp337
84. Sterling, J.A.L. (2003e), "World Copyright Law", Sweet & Maxwell, London, pp531
85. Sutter, G. (2007), "Online Intermediaries", in: Reed, C. and Angel , J. (Ed.), Computer Law, Oxford University Press, pp.233-282.
86. 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, pp9-35.
87. 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.1771-1779.
88. Wikipedia (2001), "GNU Free Documentation License", available at http://en.wikipedia.org/wiki/GNU_Free_Documentation_License (accessed July 3, 2007)
89. Wikipedia (2007a), "Comparison of file systems", available at http://en.wikipedia.org/wiki/Comparison_of_file_systems (accessed Dec. 3, 2007)
90. Wikipedia (2007b), "Resource Description Framework", available at http://en.wikipedia.org/wiki/Resource_Description_Framework (accessed Dec. 3, 2007)
91. Yahoo (2007), "Creative Commons Search", available at http://search.yahoo.com/cc (accessed July 3, 2007)
92. Yahoo (2008a), "Yahoo! Slurp—Yahoo!'s Web Crawler", available at: http://help.yahoo.com/l/us/yahoo/search/webcrawler/ (accessed March 11, 2008)
93. 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)
94. 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)
95. 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, 2009)