SCRIPT-ed

Volume 4, Issue 1, March 2007

Creating commons by friendly appropriation

Graham Greenleaf [*]

Abstract

The relationships of the world-wide-web and its search engines to the ways in which 'intellectual commons' are created, has received little consideration. I argue that the operation of Internet-wide search engines constitutes the creation of an intellectual commons. The history and features of the Google search engine are the principal example. They illustrate what is probably a very unusual method by which commons are created, which I call 'friendly appropriation'. I identify eight conditions which are conducive to the creation of commons by friendly appropriation. Some examples are given of other situations which may constitute friendly appropriation, and of some which do not.

Instances of commons arising by this means may be rare, but a fully-developed theory of intellectual commons needs to recognise when they occur.

DOI: 10.2966/scrip.040107.117

© Graham Greenleaf 2007. This work is licensed through SCRIPT-ed Open Licence (SOL).

---

[*] Cyberspace Law and Policy Centre. University of New South Wales. This paper is part of the 'Unlocking IP' research project supported by an Australian Research Council Linkage grant. Details are on the project home page at <http://www.cyberlawcentre.org/unlocking-ip/>.

1. How do we create intellectual commons?

An intellectual commons[1] exists wherever the public, or some class of the public, can exercise some part of what were previously the exclusive rights of copyright owners, and the copyright owners are not able to rescind the continuation of this right[2]. Where copyright has expired in works because of effluxion of time, all previous rights in those works are part of the intellectual commons. A statutory licence to educational institutions to conduct photocopying under certain circumstances creates a much more limited intellectual commons. When a software author publishes a program under the General Public Licence (GPL), a different but very limited intellectual commons is commons is created.

We can describe and analyse the intellectual commons in many different ways. Drahos, for example, distinguishes between negative and positive commons depending on whether a resource is initially owned by no-one, so individual use depends on no-one's consent (negative), or whether ownership is shared by a community and individual use depends on the consent of all (positive).[3] He also distinguishes between commons which are inclusive (where all individuals hold rights) and exclusive (where use of a resource is confined to a particular group). Benkler identifies two very similar parameters which divide commons into four types: (i) 'whether they are open to anyone or only to a defined group'; and (ii) whether the system is regulated (most commons, for example sidewalks) or unregulated ('open access commons'). Benkler says the most important open access commons (which Drahos would call 'negative inclusive') is 'all of pre-twentieth century knowledge and culture, most scientific knowledge of the first half of the twentieth century, and much of contemporary science and academic learning'.[4]

These distinctions are very useful, and I will return to them later, but I would like to examine a different aspect, the question of the different ways in which intellectual commons may be created.

1.1 What relocates works on the property – commons continuum?

The distinction between private rights and public rights in works is not an 'all or nothing' matter. There is a continuum of where works can be located, from what we could call 'copyright-free' at one extreme to 'maximalist' or 'über-copyright' at the

---

[1] I use the same terminology as Drahos (2006) but not necessarily exactly the same meaning of 'intellectual commons'.

[2] f the public's rights can be rescinded without breach of the licence conditions, or the licence conditions can be changed unilaterally, then it may still be sensible to recognise a commons, but it is one which is contingent or revocable. The NSW government's licence to the public to re-use legislation and cases is expressly contingent in this sense.

[3] P Drahos, "Freedom and diversity – A defence of the intellectual commons", 1 *AIPLREs* 1 (2006).

[4] Y Benkler, "The Political Economy of Commons", (2003) 4:3 *Upgrade* @ <http://www.benkler.org/Upgrade-Novatica%20Commons.pdf>, at p7.

other[5]. In fact there are numerous continua, representing each potential exclusive right (and how they are affected by exceptions, compulsory licences, fair dealing etc), and other factors such as duration of copyright. Most works are located at intermediate points along these continua.

The state, through legislative changes, can relocate works at various points on these continua. Extending the copyright term to life of the author plus 70 years relocates most works further toward the 'proprietary' end of the 'duration' dimension of the continuum. Widening the scope of what constitutes fair dealing shifts those exclusive rights affected more toward the 'public domain' end. Creation of a new compulsory licence does so in an even more dramatic fashion.

Copyright owners, through voluntary licensing, can also create some public rights in works where other proprietary rights are still held by the copyright owner ('some rights reserved'). Creative commons licences and the GPL are the best known. AEShareNet licences are among the most widely used in Australia for educational materials.

Wherever the state, through legislation, expands the extent to which the public, or some class of the public, can exercise some part of what were previously the exclusive rights of copyright owners, new commons are created or previous commons expanded. Similarly copyright owners themselves may create new or expanded commons by voluntary licensing.

1.2 Commons-based production

Yochai Benkler tries to identify the types of circumstances in which commons in relation to the use of information are likely to arise, and to be effective. He defines 'commons' in terms of institutional arrangements:

> *'Commons are a particular type of institutional arrangement for governing the use and disposition of resources. Their salient characteristic, which defines them in contradistinction to property, is that no single person has exclusive control over the use and disposition of any particular resource. Instead, resources governed by commons may be used or disposed of by anyone among some (more or less well defined) number of persons, under rules that may range from 'anything goes' to quite crisply articulated formal rules that are effectively enforced'.*[6]

Benkler distinguishes 'commons-based production' of value from production managed through property (including intellectual property), contract and managerial commands, characterizing it as follows:

---

[5] The 'copyright-free' extreme exists, and is just another way of saying 'the public domain' in the sense of all rights extinguished by effluxion of time. However, 'über-copyright' does not exist in any copyright system, such as ours, which includes some rights created by statutory licences, or exceptions for fair dealing, and does not allow opting-out from those rights, or their revocation by contract .

[6] Supra note 4, at p6.

> *'production is "commons-based" when no one uses exclusive rights to organize effort or capture its value, and when cooperation is achieved through social mechanisms other than price signals or managerial directions. Large-scale instances of such cooperation are "peer production".'*[7]

In addition to the paradigm of open source software, 'the Internet abounds with commons-based peer production', he says, noting Wikipedia, the Open Directory Project and SETI@home as leading examples.[8] His book, *The Wealth of Networks*[9] opens with a description of six ideal types of information production strategies which do not depend on rights-based exclusion (ie intellectual property or contracts). One of the unique features of the production of information is that it is also an input into its own production. Where information producers can obtain the information inputs they need for their own production at no cost, they can increase the efficiency of their own production.

1.3 Where do the web and search engines fit in commons theory?

One thing that surprises me is that the world-wide-web itself, including the operation of search engines that are now central to our effective use of it, do not seem to feature in Benkler's examples of commons-based success stories, or fit easily into his ideal types of information production strategies. If the web is some form of commons, as common sense suggests, how is that commons created? And what do search engines have to do with its role *as a commons*?

Obvious sources don't provide much information on these questions. Neither the Wikepedia entry for 'world-wide-web' nor the entry for 'Google' make any mention of 'commons'[10]. More popular studies of the 'search engine industry' such as John Battelle's *The Search*[11] and David Vise's *The Google Story*[12] also have nothing to say on the subject. Perhaps these are questions with answers so obvious that no-one bothers answering them. Or is this an instance of something that may be obvious once pointed out, but that no-one has paid much attention to.

After looking at the question, my tentative suggestion is that we need to recognise an additional method by which commons are created, one which does not involve either legislation or voluntary licensing. In the right circumstances, what I call 'friendly expropriation', can create a commons without a statutory licence. Something

---

[7] Y Benkler, "Commons-Based Strategies and the Problems of Patents", 305 *Science* 1110 (2004), p1110.

[8] Ibid.

[9] Y Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom,* (2006).

[10] Wikipedia entries 'World Wide Web' at <http://en.wikipedia.org/wiki/World_wide_web> and 'Google' at <http://en.wikipedia.org/wiki/Google>

[11] J Battelle, The Search – How Google and its Rivals Rewrote the Rules of Business and Transformed our Culture, (2005).

[12] D Vise, *The Google Story*, (2005).

resembling a *de facto* equivalent to a compulsory licences can result from general acquiescence in uses of works which are, or might be, copyright breaches.

To illustrate what is meant by 'friendly appropriation' I will discuss how Internet-wide search engines such as Alta Vista and Google have created a 'searchable commons' of the world-wide-web. To illustrate where these conditions do not seem to apply, even though they might appear similar at first glance, I will discuss Google Book Search. From these examples, I suggest conditions that seem to be required before commons can be created by these methods.

2. The web as an intellectual commons

2.1 The web as a browseable commons

Up to the mid-1990s, the world-wide-web grew to quite a few million pages of content, most of it available for free access. So it was already a commons of quite an unprecedented size and scope, made possible by new technological developments. It was a commons not merely because it allowed the public free access to a wide variety of content. Public libraries had done that, on a very large scale since the mid-nineteenth century though their origins are much earlier.[13] However, neither the right to read a work nor the right to loan a copy to another person were ever within the exclusive rights of the copyright owner, so libraries created a commons based on such functions as borrowing systems, reading rooms and inter-library loans. It was a commons, created by a new intermediary, the public library, but one that existed outside the scope of authors' rights, and therefore required no diminution of them.

From its inception the web was a new type of commons in copyright terms. Most content available via the web consisted of copyright works, but for users to enjoy that content even by reading it on screen, it was necessary for their browser software to download a copy of the work into the web cache of their PC and to retain it for a period the length of which depended on their computer and browser settings. Web browsing therefore involved users exercising what may have been part of the copyright owner's exclusive rights, the right to reproduce the work. This was obviously what the owner intended by putting the work on the web for free access.[14] Depending on the jurisdiction concerned, this was either not a breach of copyright because such 'temporary' reproductions were not part of the reproduction right (eg s43A *Copyright Act 1968*, Australia), or because the owner's actions constituted an implied licence to users to reproduce the file to the extent necessary for web browsing. In practice, owners were also giving users an unpreventable ability to make a permanent copy of the file and to print it out for their private purposes. None of these results were particularly contentious, but they combined to create a more valuable body of public rights to use content than any development since public libraries.

But at one level of organisation the web was still a shambles. How could you ever find the content you wanted? I heard it described in 1995 by analogy to 'The Library of Congress after a tornado removed all the spines from the books and upended all the

---

[13] Wikipedia entry 'Public Libraries' at <http://en.wikipedia.org/wiki/Public_library>

[14] By placing a file in the 'public_html' directory of a web server, or equivalent thereto.

pages'. Catalogues[15] of web sites, of which the most famous was Yahoo! were simply extensive lists of links to other web sites, classified into categories and accompanied by brief descriptions of each site. They provided the most systematic means of content discovery then available.[16] They worked well enough in the early years of the web, but were difficult to scale up as the web expanded. So at this stage we could say that the world-wide-web constituted a 'browseable commons'.

2.2 What search engines add

From 1996 search engines have made it possible to search an increasing proportion of all content available via the web. Digital's AltaVista search engine was the first large scale provision of a public facility to search the full text of web pages located all over the world-wide-web.[17] At its release on 15 December 1995 it made 16 million web pages searchable, which may have been most of the content of the web at that time, and immediately eclipsed some earlier web search engines using similar web crawler (or 'spider' or 'robot') technology.[18] By the time Google was formed as a company (September 1998-), search engines had fundamentally changed the value of the web as a commons by creating the capacity to search the full text of web pages, and making that central to how the web operated for users. 'Search' became central to the operation of the web, reducing both the importance and the value of Internet catalogs and of human-memorable domain names as methods of discovery of new web content and navigation to known web content.

What advantages do search engines add to the pre-existing 'browsable commons'? And what functional features allow those advantages to be created? Without becoming too technical[19] we can include at least the following elements:

- *Searching for web content is usually far faster and more effective* than browsing catalogs, for many reasons,[20] though catalogues continue to perform a supplementary role in web discovery and navigation. Since about 1995 the web has expanded at a rate far exceeding the capacity of any human-constructed catalogs, so they are always out of date. The search engine technology that makes the difference is the web robot (also known as 'spiders' or 'crawlers', but I will stick to 'robot' from here on. Essentially a web robot

---

[15] Catalogues are also called 'intellectual indexes', and in some contexts 'menu hierarchies'. I will use 'catalogue'.

[16] There are of course other means: people learn of domain names from advertisements, from emails by colleagues, by following links on other pages (the basis of the web) and by guessing what a domain name might be (eg www.cocacola.com).

[17] The Archie application searched file names using the FTP protocol, and the names of documents on gopher servers could also be searched by Veronica (Battelle 2005, p39). WAIS servers also existed prior to AltaVista. But none of them searched web pages because they predated the web.

[18] The WWW Wanderer and WebCrawler were two of the search engines pre-dating AltaVista by a year or so (Battelle 2005, Chapter 3).

[19] For a straight-forward explanation of search engine functions in the context of Google, see 'What's a Search Engine?' in Clarke 2006.

[20] Reasons include the costs of intellectual indexing (cataloging) in comparison with automated indexing, which ensures that it is always very shallow in comparison, and the difficulty of keeping catalogs up-to-date in relation to a rapidly and vastly expanding web.

is a program that automatically follows links from one web page to another, downloading a copy of all web pages it encounters to the servers of the search engine operator. These pages are then used to create a word-occurrence index, sometimes called 'concordances' or 'inverted files', of the locations of every word on every page downloaded, and for other purposes such as analysis of links between pages.[21] The search engine's text retrieval software then searches this word-occurrence index in order to find web pages matching the user's search request.

- *pages most relevant to a search request can usually be found*, listed at the top of a set of search results, despite the vast and continuing growth in web content and the consequent huge sizes of sets of search results. This is achieved by various relevance ranking algorithms (also called precedence algorithms) used by search engines. Relevance ranking does not involve making additional copies of the web pages being ranked, so does not itself involve any additional use of exclusive rights. However, Benkler points out that Google's PageRank, the most successful of the ranking methods to date, is also one of the best examples of the peer-based production because its 'core innovation' 'was to introduce peer-based judgments of relevance' .[22]
- *of search results helps the user assess the relevance of a page.* The normal means of displaying search results involves the reproduction of the title of the document, plus a few lines of text from the original web page, either from the start of the document or showing the user's search terms in context. Normally, the user will then browse the original document because the title of the document is linked to the web location of the original document. Hence no further copying of that document by the search engine operator is involved, though of course there will be copying by the user in the normal course of browsing. Where image search facilities are involved, reproduction of at least 'thumbnail' copies of the original images are likely.
- *that are temporarily inaccessible from the web can still be found.* In addition to using the copies of all web pages made by its robots to create the word occurrence index, Google allow users to display the copy of the page held by them, in the 'Google cache', by clicking on the 'Cached' link in the Google search results display rather than on the title of the document. The other major search engines including Yahoo! And MSN do similarly.[23] This is very valuable when the server on which the original document is located is

---

[21] See Wikipedia entry 'Web crawler' <http://en.wikipedia.org/wiki/Web_crawler> for a simple description.

[22] 'More fundamentally, the core innovation of Google, widely recognized as the most efficient general search engine during the first half of the 2000s, was to introduce peer-based judgments of relevance. Like other search engines at the time, Google used a text-based algorithm to retrieve a given universe of Web pages initially. Its major innovation was its PageRank algorithm, which harnesses peer production of ranking in the following way. The engine treats links from other Web sites pointing to a given Web site as votes of confidence. Whenever someone who authors a Web site links to someone else's page, that person has stated quite explicitly that the linked page is worth a visit. Google's search engine counts these links as distributed votes of confidence in the quality of the page pointed to.' See Benkler, supra note 9, p76.

[23] *Field v Google*, (2006) 412 F. Supp. 2d 1106, FN 2.

inaccessible to the user for any reason.[24] This involves the search engine operator making additional reproductions of the original web page and making it accessible in such as way that it is likely to constitute 'making available' or 'communicating' the work.

- *may be displayed in formats different from the original*. The Google cache, for example, also presents documents with valuable formatting features not found in the original. The user's search terms are highlit in colour wherever they occur in the cached version of the page. Documents that may exist in their original location on the web as PDF documents or Word documents are converted intoHTML format and may therefore be easier to access and use for some users. Such transformations of the original pages may constitute adaptations of the original work or equivalent concepts in other copyright laws. A further variant on this is that search engines often also provide, at the user's request, automated translations of the original document into another selected language. Whether good or bad, these translations are adaptations of the original work.

2.3 Search engines, commons and copyright

From the above description, it should be apparent that the revolution that search engines have brought to the web has only been possible because the operators of search engines have been able to exercise, as a matter of fact, part of what would normally be the exclusive rights of copyright owners all over the world, particularly the rights of reproduction and adaptation, but also the newer rights such as 'making available' and 'communication', in those countries where they apply. Although less significant, users are also able to reproduce content, when browsing search results, re-formatted documents or translations. Arguably this also involves exercising rights of copyright owners, at least in some jurisdictions.

If this is correct, what search engines have done is create a new form of commons by turning part of the exclusive rights of copyright owners into public rights in the sense that they can be exercised by intermediaries, the search engine operators, for the benefit of all web users. They have transformed the previous 'browseable commons of the web' into a new 'searchable commons of the web'.

Most of these practices are now a decade old, and the companies based around these technologies and practices are thriving. Legal challenges to the practices seem to be only intermittent and marginal – although they are increasing in 2006. In most countries it seems there have been no serious legal challenges at all, but this may be explained in part by the fact that the major search engines are based in the USA.

In the USA, the first legal challenge to Google's caching practices was decided in its favour only in January 2006. A US District Court judge in Nevada[25] upheld five different Google defences against the claim that the caching practices breached copyright. There was no direct infringement by Google when users downloaded pages from its cache, because this did not involve a volitional act by Google (Field had not

---

[24] For example, it may be down for maintenance, or some intermediate Internet link may be broken, or access may be so slow that the user's computer 'times out'.

[25] *Field v Google.*

claimed for indirect infringements that may have resulted from mere creation of the cache).[26] The defences were upheld on the hypothetical assumption that direct infringement had been shown. Google was found to have an implied licence, which the Court held 'may be inferred based on silence where the copyright owner knows of the use and encourages it'. Despite knowing about that he could use the 'no archive' meta-tag on his web pages to prevent caching by Google, Field 'chose not to include' it, knowing that 'Google would interpret the absence of the meta-tag as permission' .[27] For similar reasons, a US doctrine of estoppel also prevented Google from pursuing its copyright claim because he 'intended for Google to rely on his silence' and because 'Google was not aware that Field did not with to have Google provide 'Cached' links to his works', it 'detrimentally relied on Field's silence'.[28]

The caching practices were also found in *Field v Google* to be 'fair use', satisfying the four factors required in such an analysis.[29] Caching was held to be a 'transformative' use, adding something new to the original works and not merely superseding them. Google's cached pages also had links back to the original, indicating they were not intended to supersede them. The Court accepted that 'there is no evidence that Google profited in any way' from the caching of Field's work, and did not display advertising in relation to it (p16).[30] Although Field's work was creative, which would normally be a factor against fair use, the fact that it had been made available for free access on the Internet by Field, and with the intention that search engines index it, mitigated against this. The fact that the whole work was used did not matter in this context, since caching did not serve its purpose unless that occurred (p18).[31] There was no evidence of a market for Field's work, so caching could not affect that (p19).[32] Finally, in Google's favour was its 'good faith in operating its system cache' because it adopted industry-standard protocols to allow site owners to avoid pages being cached (p20).[33] Not surprisingly, these 'fair use' factors are similar to many of the factors I have identified as indicating where 'friendly appopriation' may be successful.

Field v Google also found that Google's practices were entitled to protection under one of the 'safe harbor' provisions of the *Digital Millenium Copyright Act* (17 USC s512(b)(1)). It satisfied the requirement of 'intermediate and temporary storage' because material was held in the cache for only 14-20 days (presumably the current range of time for re-spidering a site). It was by an 'automated technical process' and one of its principal purposes was to enable users to access the pages when they could not do so from the originating site for whatever reason (p23). [34]

---

[26] Ibid, p10.

[27] Ibid, p11.

[28] Ibid, p12.

[29] Ibid, p13.

[30] Ibid, p16.

[31] Ibid, p18.

[32] Ibid, p19.

[33] Ibid, p20.

[34] Ibid, p23.

Google has had similar success in a US District Court concerning its cache in *Parker v Google*.[35] However, in *Perfect 10 v Google*,[36] another District Court held that 'P10', an adult content website, was likely to succeed in its claim of direct infringement of copyright by Google's practice of reproducing 'thumbnail' images from its website in Google's image search. An interim injunction was issued against Google continuing to do so. Google's arguments that the thumbnails constituted 'fair use' were rejected. This Court took quite a different approach to some of the issues decided in Google's favour in *Field v Google*. For example, Google's use of the thumbnails was held to be commercial in nature. Despite the Court recognising 'the enormous public benefit that search engines such as Google provide', and that only some of the fair use factors weighed in Perfect 10's favour, it held there was no fair use. Agence France Presse (AFP) has initiated a similar action, claiming Google News infringes 5 copyright by displaying headlines and thumbnails.

So after ten years of the operation of search engines, there are only a small number of US Court decisions that go directly to the question of whether the key aspects of the operation of search engines are within copyright law. Most of these are decisions at the lowest level of the US judicial hierarchy, and although they certainly give Google and other search engines some reasons for optimism, they are not unequivocal. Even in the USA, the legal status of many aspects of the operation of search engines must still be regarded as uncertain.

The USA gives more liberal protection to innovative uses of works, through its 'fair use' doctrines than most other jurisdictions. Outside the USA, the overall operations of search engines are even less likely to be clearly protected by copyright law. In some jurisdictions, the meaning of 'reproduction' may exclude certain types of temporary reproductions broadly enough to protect some of these practices, or there may be explicit statutory exceptions that protect the core elements of searching, or the meaning of 'fair use' may provide protection. Although a US District Court has held Google's cache does not infringe copyright,[37] it is difficult to see this occurring in jurisdictions such as Australia. These are all ways by which the law assists the creation of commons, but they are not sufficient to explain the whole phenomenon of the searchable commons: they are not uniform between jurisdictions; and they are insufficient to explain some elements such as user access to cached copies and the various transformations involved in cached copies.

For example, there is as yet little judicial interpretation of the Australian exception for a 'a temporary reproduction of the work or adaptation as part of the technical process of making or receiving a communication' (s43A *Copyright Act 1968*), and it is unlikely or at least uncertain that this would provide any protection for the caching practices of search engines, though it may well be sufficient to protect the creation of the concordance necessary for a search engine to operate. The current 'fair dealing' provisions in the Australian legislation would seem unlikely to provide any assistance to search engines. There is also little prospect that the concept of implied licences will be interpreted broadly enough, at least in jurisdictions like Australia,[38] to give the

---

[35] *Parker v Google* (2006) 422 F. Supp. 2d 492.

[36] *Perfect 10 v Google* (2006) 416 F. Supp. 2d 828.

[37] *Field v Google.*

[38] Trumpet Software Pty Ltd Anor v OzEmail Pty Ltd Ors 560 FCA 1.

operators of search engines the breadth of licence they would need. Nor would the mere fact that these practices have persisted for the best part of a decade, in itself, do so.

So, while it is difficult to say how common the Australian experience would be without further investigation, it does seem likely that the copyright laws of many countries outside the USA would leave it questionable whether at least some of the key operations of search engines constitute breaches of copyright law.

Over the last decade, the operations of search engine have obviously not been treated on a global basis as if they were the largest scale and most systematic set of breaches of copyright ever to occur. Whether they have been or not, we do not know - at least some aspects of the law are still too uncertain in all jurisdictions, though not necessarily the same aspects in each.

How then do we explain the 'searchable commons of the web'? On the basis of the argument sketched above, we cannot simply do so by saying that the law provides enough exceptions to or limitations on copyright protection to ensure that search engines could function within the law. Nor can we explain it in terms of consent: explicit consent to index or to cache is not obtained from owners of web content by search engines, and it would be impossible for them to do so.

So it seems that the global 'searchable commons of the web' cannot be explained by the various forms of compulsory limitations on or appropriation of proprietary rights that originate from copyright statutes. Nor can they be explained by voluntary licensing whether express or implied. To find an answer, I think we need to also look at the question of why search engines have worked in practice, to the general satisfaction of both website operators and users of the web.

2.4 Why search engines were able to create a commons

We can identify many of the factors contributing to the creation of this searchable commons by the operation of search engines. I will generally refer only to 'Google searches', but the same comments will be applicable to many other search engines.

(i) Search engines contribute ***a significant innovation*** to the pre-existing Internet. Internet-wide search engines organise the content of the Internet better than catalogs. What search engines provide is something that copyright owners could not achieve for themselves: individual owners cannot individually ensure that their works will be found by those looking for them.

(ii) They contribute something of ***general public benefit***. Search engines are available for gratis use by Internet users, who obviously obtain great benefit from being able to find content more effectively. There is thus an incentive to copyright owners not to disrupt the operation of search engines on a large scale. To do so could result in adverse publicity and possibly legislative intervention to protect the operation of search engines. This is discussed further below. There is also an expectation from users that a website's content will be searchable via Google, and a risk to any website that if users cannot find it or its content via that route, they are now less likely to persist in using other methods of trying to find it such as

consulting catalogues, or using the site's own search engine when they do find it.

(iii) Search engines contribute something of ***benefit to most copyright owners***. Search engines are only making more available, content which the owner has already made available for free access. The amount of content on the web is so vast now that search engines are almost the only way that most people know how to find content where they have not been previously informed of the URL. Guessing URLs and looking up catalogues are still possible but of diminishing importance. Hence copyright owners have become dependent on their operation. Increasing access to many commercially-oriented websites is now so valuable that improving it supports an industry of 'search engine optimisation'. There is thus a very strong incentive to website owners to both (a) keep their content being accessible via search engines; and (b) acquiesce in any copyright infringements necessary for them to do so.

(iv) Search engines ***cause harm to few copyright owners.*** There are quite a few ways that search engines could potentially cause harm to the individuals or organisations whose websites they make searchable, but at least the most successful search engines have taken steps to minimise such harm. A few examples will have to suffice.

a. If a website depends on 'hits' recorded on its site, or on particular parts of its site, in order to obtain its own revenue ,from advertising or otherwise, Google searches will generally be beneficial as more users will come to the site. Sites can make their whole content searchable but re-direct users so that they only enter a site via a front page or wherever else advertising is found. Large-scale access to the search engine's cached pages could undermine this, so opt-out mechanisms are important to avoid this. See further below.

b. The viability of many Internet-based businesses may depend on how visible they are when users search for terms relevant to their line of business. By keeping paid placements, ('sponsored sites') separate from its normal listings, Google has avoided mass dissatisfaction with its rankings. All search engines have to contend with attempts by some search engine optimizers and their clients to 'game' their rankings. Benkler's observation is that Google has fought them and that its 'strategic choice is to render the distributed judgments of relevance on the Web more or less faithfully'.[39]

c. Conversely, some websites, while available for public access and providing their own search engines, do not want their content searchable via general search engines. AustLII and other legal information institutes, for example, do not want case law about individual people to be found by searchers looking for old classmates for school reunions. Websites where the content of pages change very frequently and old copies could be very misleading will not want their pages in search engine caches. They may also be reluctant to have old

[39] Supra note 9, p291.

data proving the basis of which pages are found (and which snippets are displayed), even though the new page is displayed when the user goes from the search results to the current page.

(v) The ***copyright position is not clear-cut*** with the operation of search engines, in a variety of ways. First, the content is of very mixed proprietary value. Many owners of material provided via the Internet either don't care what its position is under copyright law, or have very low awareness of the extent of protection normally provided by copyright law. Other providers of free access content via the web have a general awareness of copyright law, but for good reason are quite unsure about their position once they have made the content available for free Internet access. They may have little awareness of how search engines work at a technical level, and therefore no clear idea of what potential infringements may be occurring. In most and perhaps all jurisdictions the law on these issues is not clear, as sketched above.

(vi) It is ***not practical for search engines to obtain prior consent*** of all those websites they make searchable. The transaction costs in attempting to identify and contact the owners of billions of web pages, given that the whole purpose of an Internet-wide search engine is to make searchable as large a percentage of web content as is possible, would be prohibitive.. Nor is it feasible to encourage a sufficiently high percentage of owners of web pages to place some uniform, machine-readable 'consent to index' on their websites. 'Opt-in' has never been a viable option for Internet-wide search engines. As Benkler might put it 'searchability' is an aspect of this resource 'that make its clearance through a market particularly clunky, expensive, and inefficient'.[40]

(vii) Search engines have always provided the ***ability to opt out from searching*** for those who don't want their content to be searchable even though it is available for free access. Website owners may exclude parts of or their whole sites from robots indexing.They may exclude individual pages on a site from being searched. There are also after-the-fact methods of getting pages taken out of word occurrence indexes and having the cached copies deleted, even though they have been made searchable initially. Whether these opt-out mechanisms are effective or only notional is discussed further below.

(viii) But ***the rate of opting-out has not been enough to threaten the viability*** of search engines, despite the high costs of running them. There also does not seem to be large-scale opting out from one search engine but not others, so search engines do not seem to be operating (in this sense) by a market mechanism in which copyright owners choose which search engine they prefer their content to be searched through and using their exclusive rights or opt-out mechanisms to enforce this. Instead, copyright owners as a whole act as if a commons exists, allowing all search engines to search their content rather than picking and choosing between them.

---

[40] Supra note 4, p7.

2.5 How effective are search engine opt-out mechanisms?

Taking Google as an example, it is worth looking at the extent to which search engines provide an *effective* means of allowing copyright owners to opt out of their works being searchable or being included in the Google cache. Google does publish its removal policies,[41] but to what extent does it adhere to them? It says it will 'stop indexing pages of a site only at the request of the webmaster who's responsible for those pages', and that 'removals will take effect the next time Google crawls your site'. Nine different forms of prior blocking or subsequent removal are specified, of which a few deserve note here.

Google says it observes the Robot Exclusion Protocol[42] which enables website owners to exclude whole sites or specified directories therein from crawling by robots, by the placement of an instruction file ('robots.txt') at the root directory of the server. Use of the protocol in relation to a new website should stop Google from ever indexing it, and use on an existing website should cause removal of pages from the index, and cache, next time the site is re-indexed by a Google robot. AustLII's experience is that, while Google's robots usually observe the 'robots.txt' file, sometimes a robot will 'go feral' and ignore the protocol. It usually takes a considerable effort to contact a real human at Google in order to have the error rectified, but our experience is that it will be rectified eventually. Of course AustLII is a large website with a professional technical staff, so it may be unsafe to generalize this experience.

Google also operates an 'Automatic URL removal service'[43] which it says is for when removal requests are urgent. Use of this is supposed to cause a Google robot to go back to the site specified quickly and re-index the pages specified[44]. Provided that a correct robot exclusion file has been placed on the site, this should cause the pages to be removed from both the searchable index and the browsable cache. One problem with this is that the removal system requires you to access a non-standard port[45] on the server on which the pages are located, which may be very difficult for some webpage owners, particularly those in large organizations or using ISPs for hosting.

---

[41] 'Removing my content from the Google index', on Google at <http://www.google.com/remove.html>; see also Calishain and Dornfest (2006) for an independent assessment.

[42] Robot Exclusion Protocol at <http://www.robotstxt.org/wc/robots.html>; see introduction by Martijn Koster 'A Standard for Robot Exclusion' (1994-) <http://www.robotstxt.org/wc/norobots.html> where it is described as 'a common facility the majority of robot authors offer the WWW community to protect WWW server against unwanted accesses by their robots'. 'It is not enforced by anybody, and there no guarantee that all current and future robots will use it.'; for future extensions see Paul Ford 'Robot Exclusion Protocol' at <http://www.ftrain.com/robot_exclusion_protocol.html>

[43] The 'Automatic URL removal service' login page is at <http://services.google.com:8882/urlconsole/controller?cmd=reload&lastcmd=login>

[44] For this purpose, a 'robots.txt'. file can be placed at the same directory level as the files that are not to be crawled. This will cause those pages to be excluded from the index for 180 days, but the 'Automatic URL removal service' would then have to be used again. This in effect gives website owners who do not have access to the root of a server 180 days to get the controller of the server to change the 'robots.txt' file at the server root. Otherwise, they must exclude their site from Google at least twice a year.

[45] Port 8882 rather than the normal port 80 is required.

Google claims that it removes pages that are now 'dead links' from its index, next time the website is crawled, provided the page address displays a HTML 404 error message.[46] But it admits this may not work if other sites still link to that page. It also does not mention removal from the Google cache. AustLII's experience is that removed pages can remain both searchable and in the Google cache for as long as two years after the originals are removed from AustLII. This means that both the title of the page, and the 'snippets' in the search results can remain visible (even apart from the cache) long after the live page itself is removed.

Google also observes other protocols which reduce the likelihood of conflicts with copyright owners. Owners of web pages who do not control the servers on which the pages sit can exclude those pages from being indexed by placing a meta-tag on each page.[47] Similarly, page owners can use a meta-tag to prevent 'snippets', a small text extract from a page, from appearing with the title of a page in the list of search results.[48] Removing snippets also removes cached pages. The same meta-tag approach can be used to remove cached pages without preventing the page from being searchable.[49]

These opt-out mechanisms are not perfect, but they are not notional either. It seems that Google and other search engines have sufficiently effective opt-out mechanisms to handle a 'manageable' level of dissatisfaction with the many aspects of their operation. Such dissatisfaction as there is does not lead to a degree of opting out which threatens the operation of any one search engine or search engines as a whole. This 'manageability' is now partly a reflection of the dominant position that most search engines including Google, now play in discovery and navigation of the web: it is very difficult for most website operators to opt out if they wish their content to be discovered by anyone who has not seen it before. What about independent distribution of addresses and passwords? Eg to student materials The fact that it is so inconvenient to opt out strengthens the extent to which 'searchability' t has become a commons, but reduces the extent to which it is 'friendly'.

3. Friendly appropriation

My hypothesis is that, where third parties make use of Internet content on a large scale which involves exercising some of the exclusive rights in a work, a set of factors something like (i) – (vii) above may encourage the copyright owners concerned not to

---

[46] 'How do I remove a dead link from the search results?' on Google at <http://www.google.com/support/webmasters/bin/answer.py?answer=34440&ctx=sibling>.

[47] On each page, a meta tag is placed in the <HEAD> section of the page stating <META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW"> to block all robots (part of the Robot Exclusion Protocol). Robots from specific search engines can also be blocked. See <http://www.google.com/support/webmasters/bin/answer.py?answer=35303>.

[48] 'Removing snippets' on Google at <http://www.google.com/support/webmasters/bin/answer.py?answer=35304>

[49] To prevent all search engines from showing a cached version of a page, the tag <META NAME="ROBOTS" CONTENT="NOARCHIVE"> can be placed in the <HEAD> section of the page; see 'Remove cached pages' on Google at <http://www.google.com/support/webmasters/bin/answer.py?answer=35306>

exercise their exclusive rights, and thus the creation of a *de facto* commons. The third parties may be either intermediaries or end-users.

Google and other search engines have 'appropriated' parts of the reproduction rights, and various other rights, of the owners of works placed on the Internet for free access. But they have done so under conditions which are make their actions 'friendly' and which have therefore resulted in acquiescence. In fact they have created a new vast commons which we could call 'searchability' of Internet content. So I describe this situation as the creation of a commons by friendly appropriation.

3.1 Conditions for friendly appropriation

On the basis of the rather limited but important example of search engines, we can now make some rather tentative suggestions for when a commons is likely to arise in relation to some part of the exclusive rights of a class of copyright owners.

1. The commons organises the content better through some technical or organizational innovation;
2. Obtaining prior consent from all copyright owners is not practical;
3. The public benefits from the use being made of exclusive rights;
4. The copyright position of the works covered is unclear in relation to the use being made of exclusive rights, at least in relation to some works or under some common circumstances;
5. At least some copyright owners will benefit from the use being made;
6. Few if any will suffer significantly from the use being made;
7. An opt-out mechanism is provided, and is reasonably effective;
8. Opting out is unattractive so only a minority do so.

Where these conditions apply, a commons may result, created by friendly appropriation.

3.2 Are these conditions applicable elsewhere?

This hypothesis obviously needs to be tested against other possible examples and no doubt refined. I don't think these conditions are likely to be widely applicable. The most likely candidates I can think of would be some types of content created by the public sector, where similar expectations to those affecting legal materials may apply, and the public takes the attitude 'we have already paid for this as taxpayers' or 'access to this information is necessary for a functioning democracy', and governments consider it wise not to object. Do such examples exist?

An example from personal experience of the latter is the initial creation of the British and Irish Legal Information (BAILII) website <http://www.bailii.org> in early 2000. AustLII created the original BAILII site by taking all legislation and case law available on any UK government sites and creating a set of databases from them. The UK content was protected by Crown copyright. Explicit permission was not obtained from the agencies concerned, nor from the UK committee then investigating the creation of a UK equivalent of AustLII. The site was launched by AustLII without anyone else's permission, but also without objections. Once available, it satisfied the needs of the UK parties concerned, and before long Courts and agencies were supplying updates of the databases. Perhaps we could say that for a brief time the

character of this information as a commons was based on friendly appropriation, but that rather rapidly changed to one based on licences implied by the supply of data. Much of BAILII's provision of data probably continues to be on the basis of such implied licences, as is the case with AustLII. I would not be surprised if there were more examples around of intellectual commons where original friendly appropriations are succeeded by commons based on implied or express licences.

Whatever other examples do occur will probably be less important than the position of the web as a searchable commons. Though it may be obvious, it is such an important case that the precise conditions of its existence and maintenance should be of concern to anyone interested in the expansion and protection of commons.

3.3 Friendly or unfriendly appropriations: What happens when the line is crossed?

Perhaps the best place to finish this initial exploration is a few comments about appropriations of intellectual property that are not regarded as friendly, the dividing line between the two, and possible scenarios depending on whether practices are perceived as 'friendly' or 'unfriendly'.

In *Free Culture*, Lessig makes the bold claim that '*every* industry affected by copyright today is the product and beneficiary of a certain kind of piracy'.[50] He supports this claim with four US examples of what were clearly regarded at the time by intellectual property owners as 'unfriendly' appropriations: the way in which the early film industry ignored Thomas Edison's patents over filmmaking technologies; the assertion by the early developers of the recorded music industry that they should be able to make and sell recordings of performances of musical works despite the copyright of the composer, in a situation where the scope of the composer's right to control 'public performances' was unclear in the face of this new technology; the rise of radio in a situation where the law did not give the owner of a sound recording control over whether it could be played on radio; and the practice of cable TV operators of providing cable access to TV broadcasts without payment to the stations concerned. In the first case Edison's patents expired to resolve the issue, but in the last three Congressional intervention provided the same solution, one that did not give a complete victory to either party. A compulsory statutory licence legitimated the practice previously regarded as piracy, so that existing content owners could neither prevent their works being used not unilaterally decide the price for that use, but the new users were required to pay a licence fee determined by some independent means .[51] In these cases of 'unfriendly appropriation' a commons was created by a compulsory licence, resolving a previously uncertain or unsatisfactory legal situation.

Such legislative interventions are clean-cut resolutions of a previously messy situation. So are clear victories in the courts to one side or other – though of course they can sometimes trigger legislative interventions to the opposite effect. But life, and the development of commons, is not always so clean-cut, and untidy and uncertain situations can sometimes persist for a long time. Nevertheless, what happens during this uncertainty may be of considerable significance, as the short

---

[50] L Lessig, *Free Culture* (2004), p61.

[51] Ibid, Ch. 4.

history of Internet-wide search engines illustrates. Our theories of commons need to accommodate and recognise these developments.

3.4 Where does Google Book Search fit?

A parting question: where does Google Book Search[52] fit in this discussion? In relation to out-of-copyright works the basic concept of converting the world's printed commons into a searchable commons is clearly a wonderful augmentation of the intellectual commons by technological advances. However, considerable questions remain about the extent to which this particular example may become an undesirable appropriation of commons for proprietary ends. Google is not allowing whole books to be downloaded by users, even if they are part of the commons:

> *"If the book isn't under copyright at all, you can browse the entire book in the Full Book View, but the aim of Google Book Search is to help you discover books and learn where to buy or borrow them, not read them from start to finish."*[53]

But it is in relation to books which are still within their copyright term that most dispute arises. With most such books Google is allowing the full text of the book to be searched, but only allowing a few snippets to be viewed from a page that contains the search terms. Some publishers are providing their publications to Google for this purpose. From the perspective of other publishers, Google Book Search is very definitely 'unfriendly' appropriation, and a number of legal actions have been commenced in the USA. Whether these aspects of Google's practices are within the 'fair dealing' exceptions in US copyright law, or are protected by other aspects of that law, is uncertain and remains to be seen.

Many authors may take the same views as their publishers, but others may not. If an author's book is still in print, there are many factors an author might want to consider before deciding that it is a good idea that consumers should be able to search ever word of the book before deciding to buy it. But if an author's book is out of print the calculus may be very different: with no financial return in issue, authors may be delighted to have readers discovering their otherwise lost works via Google, particularly if they are then directed to libraries from which they can borrow them. From the publisher's perspective the position is very likely to be the reverse: their financial interests are only served by consumers discovering books that are in print, and can well be harmed by readers discovering that there are a lot of good books on a subject that cannot be found on the shelves of either a local bookshop or online at Amazon, but perhaps can still be found in a library. Better still if the author has authorised a public repository to make the book available for free access.

Both searchability of books, and the ability to download books which are either out of copyright or out of print (but where the copyright owner consents), are likely to be the next big extensions to the intellectual commons. However, unlike the essentially friendly history of the search engine to date, most books are unlikely to become part of the commons by friendly appropriation. This question of the boundary between the

---

[52] Google Book Search at <http://books.google.com.au/?hl=en>

[53] Google Book Search FAQ <http://books.google.com.au/intl/en/googlebooks/help.html#pagelimit>

proprietary and commons is more likely to be resolved in the courts and the legislatures.