

SCRIPT-ed

Volume 4, Issue 1, March 2007

Finding and Quantifying Australia's Online Commons

Ben Bildstein*

Abstract

Exploration of the online commons is relatively new. The original landscape of the World Wide Web was very spread out, and traversing it usually meant following links from page to page. At this stage, the only way to find works with public rights was to stumble across them. With the growth in search technology, the Internet became much more accessible, though it has surely grown to compensate. At this stage, it was possible to search for online commons, but only in the most rudimentary way – trying to guess which words were more likely to be found on pages with public rights, and then searching for them. Now, with the slow transition towards the Semantic Web, we are seeing an Internet that is even easier to traverse – where there are web pages that know something about themselves, something that can be communicated to search engines, and the landscape of the Internet can come to life.

* Ben Bildstein is a PhD candidate at the University of New South Wales. This research was conducted as part of the Unlocking IP project and is supported by an Australian Research Council grant. Significant contributions to this work were made by Professor Graham Greenleaf of the Law Faculty at UNSW and Catherine Bond, who is another PhD candidate in the Unlocking IP project.

This paper briefly describes some of the ways that online commons can express their public rights, followed by an exploration of the ways in which people can go about finding works that are part of Australia's online commons, using current tools. Then, using some of these techniques, data are gathered and analysed, to give an overview of the current state of some parts of the online commons in Australia. Lastly, consideration is given to what these finding may imply for the future of the online commons.

DOI: 10.2966/scrip.040107.8



© Ben Bildstein 2007. This work is licensed through a Creative Commons [Attribution-
No Derivative Works 2.5 Australia](https://creativecommons.org/licenses/by-nc/2.5/au/).

1. Introduction

The purpose of this paper is to explore the consumer perspective in Australian online commons, from the creation of public rights to finding documents based on such rights, to some simple methods for quantifying online commons, and brief consideration of some data obtained with these methods.

First, Section 0 considers some background, discussing concepts such as ‘public rights’ and ‘online commons’.

Section 0 reviews various ways that works can gain public rights in the context of Australia and the World Wide Web. Particular attention is given to AShareNet licences and Creative Commons licences; the former due to their Australian focus, the latter due to their prevalence as well as the interesting technical aspects.

Section 0 reviews various publicly available methods and tools for finding works with public rights, drawing specifically on the ways public rights can be created, as detailed in Section 0.

Section 0 makes use of some of the methods from Section 0 to gather some indicative data on how many relevant web documents can be found with them. This provides an opportunity to gain some insight into the current state of online commons in Australia.

Section 0 looks to the future, considering the impact of the Resource Description Framework (RDF) on finding commons, and where search engine technology might be heading in the medium term.

Section 0 discusses the results from the previous sections, while Section 0 considers future work.

1.1 Background

No discussion of online commons can make sense without a definition of *commons*, but it turns out that defining commons is not as simple as it might first appear.

Elkin-Koren gives a good coverage of the issue in *Exploring Creative Commons: A Sceptical View of a Worthy Pursuit*.¹ The paper uses the following definition from Heller in *The Tragedy of the Anticommons: Property in the Transition from Marx to Markets*: “multiple owners are each endowed with the privilege to use a given resource, and no one has the right to exclude another.”²

This definition needs a slight adjustment before it is appropriate for the World Wide Web context: the majority (if not all) of the documents that make up the World Wide Web are copyrighted and there are not multiple owners - let alone everyone being an owner. Rather than a commons of multiple owners, we can have a legal *licence* associated with a document, which grants *rights* (but not ownership) to the public (*public rights*), so that although the copyright owner continues to own the document,

¹ N Elkin-Koren, “Exploring Creative Commons: A Sceptical View of a Worthy Pursuit”, in P B Hugenholtz and L Guibault (eds), *The Future Of The Public Domain* (2006), 325 (hereafter referred to as Elkin-Koren, *Exploring Creative Commons*).

² M A Heller, “The Tragedy of the Anticommons: Property in the Transition from Marx to Markets” (1998) 111 *Harvard Law Review* 621, quoted in Elkin Koren, *Exploring Creative Commons*.

the public is allowed to make use of it, without needing to be considered “multiple owners.”

Elkin-Koren goes on to point out many more grey areas in the concept of commons in terms of intellectual property, including: “Does it cover unprotected aspects of copyrighted works or also any type of exploitation of works which falls outside the scope of copyright? ... Does it have to be free of any legal restraints? Is it enough that works would be widely disseminated? Could some restrictions apply and the work still be considered free?”³

These questions are very important from an ideological perspective, especially when considering specific licensing regimes such as Creative Commons⁴, and to what extent such regimes can and do achieve their stated aims. However, this paper does not aim to critique particular licensing regimes or other mechanisms for granting public rights in online documents or other copyrighted works; rather, it attempts to develop a consumer perspective, where certain classes of works that can clearly be seen to constitute a commons can be used to facilitate considerations of the finding of such works online and the possibilities for quantification of these classes of *online commons*.

1.2 A common-sense definition of commons

In *The Wealth of Networks*, Benkler says of commons:⁵

The salient characteristic of commons, as opposed to property, is that no single person has exclusive control over the use and disposition of any particular resource in the commons. 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.

This is a good description of why the commons can be important to consumers. The mechanism by which a resource becomes commons, or the reasoning by which a resource can be justified as classifying, may not be as important to consumers as what they can do with the resource.

If we simply consider *the commons*, for the purposes of this paper, to be simply those resources that are free for all to use, and are non-consumable. In terms of the World Wide Web (*the Web*), the *online commons* would then be those electronic documents, available without technical restriction on the Web, that may be used as described by Benkler.⁶

³ Elkin-Koren, *Exploring Creative Commons*.

⁴ <<http://www.creativecommons.org>>

⁵ Y Benkler, *The Wealth of Networks*, (2006) 61.

⁶ *Ibid.*

From this definition, one might ask the question: *are all online documents 'commons'*? From the common-sense perspective, the answer must be that they are not, if only because the converse would make the definition useless if not meaningless. From a practical perspective, the answer can still be no, because the copyright owner of a web document has the right to deny any member of the public the use of the document.

This paper uses this (somewhat informal) definition of the commons, to describe classes of Web documents that may be more interesting to consumers based on additional permissions or rights they may have or be entitled to in the documents. The term *public rights* is used to refer to the particular rights or permissions that the public (or in some cases particular classes of people) have, with respect to the documents.

2. Creating Public Rights

It is usually easy to tell whether a work is available online;⁷ it is far harder to decide if it has any public rights. There is no clear criterion, especially not one that could be readily implemented in software, and so it seems that the only way to map Australia's online commons will be to enumerate the various ways in which something can acquire public rights, and then map each type of material separately. As a start, what follows is a list of some of the many types of online commons relevant to Australia, based on the mechanisms by which they become free to reproduce.

2.1 AEShareNet Instant Licences

AEShareNet⁸ has developed four licences that can be used to grant public rights. These are called AEShareNet-U ('Unlocked Content' – free to use, reproduce and modify), AEShareNet-S ('Share and Return' – free to use, reproduce and modify, but copyright in the modifications pass to the licensor), AEShareNet-P ('Preserve Integrity' – free to use and reproduce, but may not be modified) and AEShareNet-FfE ('Free for Education' - free to use and reproduce, but only for educational purposes, and can not be modified).

There are two mechanisms for applying these licences to works.⁹ The first is to register the work with AEShareNet, who will enter it into their database for easy online discovery and retrieval. The second mechanism is to accompany the work with a hyperlink to the licence page, a logo designed by AEShareNet that symbolises the licence, and some standard text that describes the work as being licensed with the specific licence.

⁷ In fact it is very common for some web documents not to be indexed by search engines, which can make finding them very difficult. However, this would usually be achieved through use of the 'robots.txt' protocol, which requires affirmative action on the part of the Webmaster for each page or directory that is excluded (with the default being that documents are not excluded).

⁸ AEShareNet is a non-profit company which operates the AEShareNet website at <http://www.aesharenet.com.au>. For more information, see <http://www.aesharenet.com.au/whoAreWe/>.

⁹ <http://www.aesharenet.com.au/coreBusiness>.

2.2 Creative Commons Licences

Creative Commons have developed a suite of licences based on combinations of four licence elements, as well as a “a set of other licenses for more specialized applications,”¹⁰ such as the Sampling Licenses¹¹ to allow licensees to use parts of a work but not to reproduce the whole. This paper focuses on the mainstream licenses, which are made up of the following elements:¹²

- Attribution: Use of the work must be properly attributed to the author. All current licences include this element, but it was an optional element in the version 1.0 licences.¹³
- No Derivatives: The work is not allowed to be modified.
- Non Commercial: The work is not allowed to be used for commercial purposes.
- Share Alike: People who make modifications to the work must licence their derivative works under the same licence as the one that allowed them to use the original. This is similar to GNU’s ‘copyleft’ idea.¹⁴

These elements combine into six current licences,¹⁵ with five more licences that were only in version 1.0 of the licences,¹⁶ and the additional alternative of public domain dedication.¹⁷ These licences are designed to be used by doing the following three things to web pages that contain the content to be licensed:

- Put the Creative Commons ‘Some Rights Reserved’ logo on the web page, and make it a hyperlink to the relevant licence.
- Put some words nearby to the effect that the work is available under the relevant licence.
- Embed some RDF metadata in the HTML source of the web page. This RDF¹⁸ is capable of expressing the relationship between the content on the web page and the licence, and expressing the elements of the licence. It will be discussed later, but for now, the important thing is that this RDF is hidden content that is not displayed by web browsers, but can be seen and used by Internet search engines.

¹⁰ <<http://creativecommons.org/about/licenses/meet-the-licenses>>.

¹¹ <<http://creativecommons.org/about/sampling>>

¹² See <<http://creativecommons.org/about/licenses/>> for a fuller description.

¹³ <<http://creativecommons.org/weblog/entry/4216>>, last accessed 2007/03/01

¹⁴ For more information on copyleft, see *What is copyleft?*, <<http://www.gnu.org/copyleft/>>.

¹⁵ <<http://creativecommons.org/licenses/>>.

¹⁶ See note 13, above.

¹⁷ <<http://creativecommons.org/licenses/publicdomain/>>.

¹⁸ ‘Resource Description Framework.’ For a brief explanation, see <http://en.wikipedia.org/wiki/Resource_Description_Framework>. For the definitive reference, see <<http://www.w3.org/RDF/>>. We will talk about this in Section 0, 3.2.3 *RDF-Based Creative Commons Search*, below.

Any one of these three points would indicate licensing, but Creative Commons advocates using all three, which makes the statement of availability as clear as possible.

One of the interesting and complicating things about Creative Commons licences is that although there are not very many fundamentally different licences, they have been ‘ported’ into numerous legal jurisdictions, and there are multiple versions. In total, there are over 200 distinct jurisdiction/version/type licences.¹⁹

2.3 Software and Other Licences

Far and away the most common licence for making software publicly available is the GNU General Public License (GPL).^{20 21} This licence is generally applied by including a copy of the licence with the content that is to be licensed. For example, when licensing a compiled software package, one of the files that is installed might be a copy of the licence, and the software, when run, might output a message saying that it is free software and specifying how to find the licence. Similarly, when distributing source code, one of the files in the package is usually a copy of the licence, and the individual source code files usually make some reference to the licensing. As will be discussed in Section 0, below, this method of describing works as licensed makes finding and quantifying such works inherently difficult, as there is no single standard on how to achieve the licensing.

Along a similar vein to the GPL is the GNU Lesser General Public License (LGPL) used mostly for releasing software libraries.²² There are also many other Free and Open Source Software licences. The Free Software Foundation lists sixty-five licences as ‘Free Software’ licences.²³ Open Source Initiative lists fifty-eight licences as ‘OSI Certified Open Source Software.’²⁴ Many of these are based on, or modelled after, the GPL or other popular licences such as the BSD license. Other licences are designed for specific software projects, such as the Apache License, which is used to license (among other products) the Apache HTTP Server, the most commonly used web server on the Internet.^{25 26}

¹⁹ There are currently thirty seven (as of February 2007) jurisdictions available at <http://creativecommons.org/license/>. Each jurisdiction can contain any of: version 1.0 licences, version 2.0, version 2.1, version 2.5, and now version 3.0. Version 1.0 had 11 licences; the later versions have only six. See <http://creativecommons.org/worldwide/> for an up-to-date list.

²⁰ <http://www.gnu.org/copyleft/gpl.html>.

²¹ From SourceForge (<http://sourceforge.net/>), the open source software collaboration website, 59,534 (http://sourceforge.net/softwaremap/trove_list.php?form_cat=15) of 85,588 (http://sourceforge.net/softwaremap/trove_list.php?form_cat=18) software projects are GPL-licensed. This equates to 70% of these free software projects being GPL-licensed.

²² <http://www.gnu.org/licenses/lgpl.html>.

²³ <http://www.gnu.org/licenses/license-list.html>.

²⁴ <http://www.opensource.org/licenses/>.

²⁵ <http://www.apache.org/licenses/>.

²⁶ http://news.netcraft.com/archives/2006/02/02/february_2006_web_server_survey.html.

There are also many other licences that can be used to grant public rights in works other than software. For example the Free Art License applies the principles of free software to art, ensuring that the art remains free for public use.²⁷ The Open Publications License promotes open access to academic publications.²⁸ There is also a GNU Free Documentation Licence (FDL),²⁹ used mostly for releasing documentation for free software and other reference texts. The FDL was designed in conjunction with the GNU free software licences (the GPL and LGPL), and is similar in many respects, such as in the ways that works are represented as licensed. Wikipedia is an example of a work licensed under the FDL.³⁰

2.4 Public Domain

In this context, public domain works refer to works that have the broadest public rights. That is to say that either there is no effective copyright in such works, or for some reason the copyright is unenforceable.

There are a number of ways that works can enter the public domain. The most common way for works to enter the public domain is for copyright simply to expire through the effluxion of time. Yet this category of commons content is perhaps the hardest to discover online, for the fundamental reason that there need be no difference in presentation between the work as it existed during its copyright term and the work as it exists after expiration of copyright.

Works may also be able to enter the public domain through actions of copyright holders, where the copyright holder disclaims ownership of the copyright. This is the basis of the Creative Commons public domain dedication, which was written with the specific intent of allowing copyright owners to put their works in the public domain.³¹

Lastly, it is worth mentioning that in some cases, governments may legislate to make some classes of works, or some specific works, public domain works. In such cases, the issue of discovery will become one of how to find works that belong to the particular class, or how to find the particular documents that are specified by the statute.

3. Finding Licensing Information

There are a number of different techniques that can be used to find works with public rights. These include:

- Search for web pages that say that they are publicly licensed. Although this is perhaps the only definitive way to license something, it is also the hardest to search for, because there is no unique phrase that must be used to achieve it.

²⁷ <<http://artlibre.org/licence/lal/en/>>.

²⁸ <<http://www.opencontent.org/openpub/>>.

²⁹ <<http://www.gnu.org/copyleft/fdl.html>>.

³⁰ <http://en.wikipedia.org/wiki/Wikipedia:Text_of_the_GNU_Free_Documentation_License>.

³¹ <<http://creativecommons.org/licenses/publicdomain/>>.

- Search for copies of the licence. This can be effective for finding, for example, works covered by the GNU GPL, where every copy of the work must include a copy of the licence. Generally, where there is a copy of the GPL, there is also some software covered by it.
- Search for pages that have hyperlinks to the licence. In the case of licences that are applied in part by linking to a page that contains the text of the licence, any pages that link to the licence are likely to be covered by it.
- Search databases that contain information on the licensing of works. For example, <<http://commoncontent.org/>> is a repository for works licensed under Creative Commons licences; <<http://www.aesharenet.com.au/>> is a repository for works licensed under AEShareNet's licences.
- Search for metadata (such as RDF). For example, if a web page's HTML source contains metadata that says that the page is licensed under a Creative Commons licence, then the contents of that page are probably licensed under a Creative Commons licence.

And, of course, the easiest way to find works with public rights is to use a search engine that has inbuilt functionality for finding such works. Yahoo, Google and Nutch (with the Creative Commons plugin) are all examples of these, which will be explored later in this paper.

3.1 AEShareNet Database Search

The simplest example of a search facility that is designed to find works with public rights is AEShareNet's database search.³² It allows users to search all registered learning materials, and even specify the details of exactly what they want to find, in terms of usage rights or attributes of the learning materials themselves. As of November 2006, this search facility contains 2,636 records marked as 'available,' (excluding commercial licences).

The nature of the AEShareNet database search gives it something that regular web searches for content that is 'free to use and share' can't have; that is, the ability to search education-specific metadata. For example, if you are interested in learning materials for the food processing industry, you could search for learning materials with National Training Information Service (NTIS) code FDF03.³³ Regular search engines will allow you to search for pages that contain 'fdf03', but there is no guarantee that when it is found, it will be referring to an NTIS code. Even if you add 'ntis' to the search, so that the search query becomes 'ntis fdf03', there is no

³² AEShareNet's advanced search is at <<http://www.aesharenet.com.au/Members/search.asp?SearchAdvanced=yes&NewSearch=yes&BasicSearch=>>>.

³³ In theory at least. In practice, material *aes0010225427xghf* has NTIS code 'FDF03 [Food Processing Industry]', but searching for 'FDF03' or 'Food Processing Industry' or 'FDF03 [Food Processing Industry]' in the 'NTIS code' field yields no results. But searching with 'fdf03' in the search terms yields multiple results, because all the FDF03 learning materials seem to contain the term FDF03 in their description.

guarantee that any pages found are actually talking about education materials with NTIS code 'FDF03'.

The full list of metadata fields that AESShareNet advanced searches can specify is:³⁴

- Material type (reference material, student work, etc.)
- Format type (text, sound, software, etc.)
- Qualification level ('course in', 'diploma of', etc.)
- Material scope (course curriculum, training package, etc.)
- Only include materials with Material Pointers (URLs for further information)
- Status of material (available, under development, and/or 'other')
- Only include materials available for licensing by associates
- Licence regime (Commercial, End-user, Free for Education, Preserve Integrity, Share and Return, and/or Unlocked Content)
- NTIS codes
- Subject codes
- Learning material number
- Member (owner)
- Search for materials that match all criteria vs. search for materials that match any criteria.

Clearly, this list contains more options for customising searches for education materials than a regular web-search is ever likely to.³⁵

3.2. Comparing Three Search Engines – Google, Yahoo and Nutch

What follows is a comparison of three search engines: Google advanced search, Yahoo advanced search, and an open-source search engine called Nutch,³⁶ which has a plug-in that allows it to be restricted to indexing and searching only web pages that refer to Creative Commons licences (hereafter referred to as 'Nutch/CC').

Currently, there is no search engine that provides all the features that are useful for searching for works with public rights. Google, Yahoo and Nutch/CC all provide some unique functionality. What follows is a list of features, and the extent to which each search engine supports them.

³⁴ See note 32, above.

³⁵ In theory, it is possible that one day, all of these concepts will be able to be included in metadata, and a search engine, which doesn't necessarily know about them in advance, will be able to search them based on the user's knowledge of how they will be incorporated into RDF by web page creators.

³⁶ The Nutch homepage is <<http://lucene.apache.org/nutch/>>.

(The Nutch search engine was previously available for use on *creativecommons.org*, but as of August 2006, it has been replaced with links to other search engines.³⁷ Some of the data presented in this paper was collected from this implementation of Nutch while it was being hosted on *creativecommons.org*.)

3.2.1 Multiple Jurisdictions

Clearly, the most important attribute of any search engine feature, when it comes to searching for online commons, is that it covers all relevant licences. For example, a search for AEShareNet instant licences will not be of much use unless it can find works covered by any of the four instant licences.

This may seem obvious for AEShareNet, but for Creative Commons there are over 200 distinct licences (see Section 0, above). And in general, people who want to find works with public rights will be happy for such works to be licensed under any of the jurisdictions' licences. Therefore, it seems logical that any search engine that tries to search for Creative Commons works would search for works from all jurisdictions, at least as a default case.

From the search engine's point of view, this doesn't actually necessitate having knowledge of every jurisdiction. Instead, the search engine can give more weight to the RDF's *description* of the licence (see Section 0 below), and perhaps additionally check that the licence is hosted at *creativecommons.org*.

One way or another, Google and Nutch/CC certainly do return results for multiple jurisdictions. But it turns out that Yahoo, when it does a Creative Commons search, only returns results that link to United States licences. For example, Yahoo can not find <http://www.behindbigbrother.com/> in a search for Creative Commons works (as at 2006/11/10), even though it is properly licensed with a hyperlink to <http://creativecommons.org/licenses/by-nc-sa/2.1/au/>, proper RDF, and the 'some rights reserved' logo. In fact, Yahoo's Australian front end, <http://au.search.yahoo.com/>, doesn't even have a Creative Commons search option.

Given this, and that Yahoo ignores RDF metadata (see Section 0, below), and that Yahoo is capable of searching for links to a specific page, it appears that any Yahoo Creative Commons search could actually be achieved by an (albeit long) regular search that includes hyperlink-based search criteria.

3.2.2 Link-Specific Search

As explained earlier, one way to find works licensed under a specific licence is to search for web pages that link to that licence. This can be done using any search engine that supports the feature of restricting a search to web pages that link to a specific URL. This can by no means be taken for granted, as the fundamental task of a search engine is simply to find web pages containing the text the user specifies.

Yahoo fully supports this feature. For example, to search for web pages that mention Tasmania, licensed under AEShareNet's 'Free for Education' licence, an appropriate Yahoo query might be:

```
tasmania link:http://www.aesharenet.com.au/FfE2
```

³⁷ <http://creativecommons.org/weblog/entry/6002>.

Google only supports this feature in the most limited way. It is possible to do a search that returns *all* indexed web pages that link to a specific page (e.g. a licence), but it is not possible to restrict this search based on a conventional content query (as in the example above). Clearly, this is not a very useful way to find licensed works.

Nutch/CC does not support this feature. This means that, although Nutch/CC is useful for finding Creative Commons works, and, to some extent, finding works licensed under GNU GPL and LGPL and possibly others, it is not an all-purpose tool for finding publicly licensed works. This is because it relies on its developers' knowledge of licensing (in this case, Creative Commons licensing), rather than the user's.

3.2.3 *RDF-Based Creative Commons Search*

When someone licences his or her work under a Creative Commons licence, Creative Commons promotes putting embedded XML metadata, in the form of RDF, in the HTML source of the web page.³⁸ In their words, this is "a machine readable translation of the license that helps search engines and other applications identify your work by its terms of use".³⁹ Here is a hypothetical example of some RDF metadata:

```
<rdf:RDF>
  <Work rdf:about="">
    <license rdf:resource="http://creativecommons.org/licenses/by-nc/2.1/au/" />
    <dc:type rdf:resource="http://purl.org/dc/dcmitype/Text" />
  </Work>

  <License rdf:about="http://creativecommons.org/licenses/by-nc/2.1/au/">
    <permits
rdf:resource="http://web.resource.org/cc/Reproduction"/>
    <permits
rdf:resource="http://web.resource.org/cc/Distribution"/>
    <requires rdf:resource="http://web.resource.org/cc/Notice"/>
    <requires
rdf:resource="http://web.resource.org/cc/Attribution"/>
    <prohibits
rdf:resource="http://web.resource.org/cc/CommercialUse"/>
    <permits
rdf:resource="http://web.resource.org/cc/DerivativeWorks"/>
  </License>
</rdf:RDF>
```

A natural-language translation of this RDF would look like this:

³⁸ <<http://creativecommons.org/education/publish-website>>.

³⁹ <<http://creativecommons.org/about/licenses>>.

This work, which is of type text, is licensed under 'http://creativecommons.org/licenses/by-nc/2.1/au/'. The licence permits reproduction, distribution and derivative works, requires notice and attribution, and prohibits commercial use.

This, by its very design, makes it easy for a search engine that sees such a page to recognise that, for example, this page would be a good candidate for any search for pages that are 'free to use, share or modify' (because the licence 'permits reproduction' and 'permits derivative works'), but not pages that are 'free to use or share, even commercially' (because the licence 'prohibits commercial use'). In fact, from a search engine's point of view, this is a much clearer indication that a page is licensed under this licence than just a hyperlink to the licence URL would be, because there are other reasons that someone might want to link to the licence page than trying to licence their work (for example, for reference or demonstration).

Google and Nutch/CC are both capable of reading RDF, and any pages that use RDF to indicate that they are Creative Commons licensed will be available in a Creative Commons search through Google or Nutch/CC.⁴⁰ Note that without specialised functionality designed to read this RDF while indexing the web, there would be no way to search based on RDF, because search engines are in the business of searching rendered web pages, not the web pages' HTML source.

Yahoo seems to ignore RDF completely. Its knowledge of what is and isn't licensed under Creative Commons licences comes entirely from finding pages that link to Creative Commons licences.⁴¹

3.2.4 Link-Based Creative Commons Search

The other way that search engines can find Creative Commons licensed works is to search for hyperlinks to Creative Commons licences. Although it is possible to do this manually by specifying the various licences that are relevant (as per Section 0 above), it is more convenient if a search engine already has such knowledge.

Google's Creative Commons search seems to ignore hyperlinks to Creative Commons licences completely, relying solely on RDF. Yahoo and Nutch/CC will both return pages as Creative Commons licensed based on linking to Creative Commons licences.⁴²

⁴⁰ It turns out that there are some pages that Google can find that do not use RDF. These include pages that use the HTML <meta> tag to describe rights. For example, the page <<http://etext.library.adelaide.edu.au/c/conrad/joseph/>> can be found (as at 2006/11/10), probably by virtue of the following line of HTML: <meta name="dc:rights" content="http://creativecommons.org/licenses/by-nc-sa/2.1/au/" />.

⁴¹ <<http://www.lightningfield.com/>> can be found by Google and Nutch, but can't be found by Yahoo (as at 2006/11/10).

⁴² Surprisingly, <<http://www.lightningfield.com/about.html>>, which is a subpage of <<http://www.lightningfield.com/>>, can be found by Yahoo and Nutch, but not by Google (as at 2006/11/10). This is because it links to a licence, but does not contain RDF, whereas the parent page contains RDF but does not hyperlink to a licence.

3.2.5 Media-Specific Search

There is nothing stopping people licensing things other than text with Creative Commons licences. In fact, when someone licenses a page with a Creative Commons licence, they are probably licensing not only all the text in the page, but also the pictures, sounds, etc., that are part of the page. Therefore, it should be possible to restrict a search to particular types of media.

Google and Yahoo already have this kind of search,⁴³ but it cannot be combined with their Creative Commons search to provide a media-specific search for works with public rights.

Nutch/CC does support searching for a specific media type, based on the RDF component that specifies the 'type' of the work.⁴⁴

3.3 Showing Licence Elements In Search Results

It is not always easy to find out which licence has been used for a given page. Finding the page in the results of a search for Creative Commons works is a good indication that it is licensed, but generally the only way to find out which licence applies is to look at the source of the page. In some cases, you can search the content of the page for 'creative commons', but in others this yields nothing. Even searching the source of the page for 'rdf' does not necessarily yield anything, because the page may only be licensed through a hyperlink to the licence. In this case, searching the page source for 'creativecommons.org' is probably the only sure-fire way to find out which licence applies.

Therefore, it is convenient if the search engine being used displays licensing information in the search results. Google and Yahoo do not do this, but Nutch/CC does.

3.3.1 Creative Commons' Public Domain Stamp

In addition to the six current and five historical licences types that can be made by combining the attributes of 'by', 'sa', 'nd' and 'nc', Creative Commons also facilitates dedication to the public domain, and promotes a similar way of marking works as being part of the public domain as for marking works as being licensed under Creative Commons licences. Therefore, search engines should be able to use the same mechanism to find these public domain works as they can for regular Creative Commons licensed works.

Google, Yahoo and Nutch/CC can all find public domain works marked with the Creative Commons public domain mark. Yahoo finds works that link to the Creative

⁴³ Google's image search is at <<http://www.google.com/imghp>>. Yahoo's image search is at <<http://images.search.yahoo.com/>>.

⁴⁴ For example, <<http://ok-lah.blogspot.com/2005/08/friend-from-dubai.html>> (as at 2006/12/04) can be found by a Nutch search for 'interactive' materials. In fact, this page doesn't have interactive content, but Nutch thinks that it does because the RDF says that it is of type interactive.

Commons public domain page;⁴⁵ Google finds works that specify in their RDF that they are in the public domain;⁴⁶ Nutch/CC appears to find only RDF-based works, but this is very hard to demonstrate conclusively, because its indexes were never as comprehensive as Yahoo's and Google's.

3.3.2 Creative Commons' GPL and LGPL Stamps

There is a fundamental difficulty in using a web search engine to find GPL-licensed works. In almost all cases, the entire text of the GPL is placed in its own page and/or document, separate from the rest of the software release. While finding either the project page or the GPL page would be easy using conventional search techniques, searching for one based on its relationship to the other is not possible with current search technology. For example, the free Go-playing software Gnu Go is released under the GPL, and its homepage is <http://www.gnu.org/software/gnugo/>.⁴⁷ This is probably the page that someone would like to find if they were searching for free Go-playing software. The licensing information can be found at http://www.gnu.org/software/gnugo/gnugo_21.html, and any search for copies of the GPL on the Internet should be able to find this page. But unfortunately, the licensing page lacks context, and as such will not show up in any search for Go-playing software. For example, the Gnu Go homepage contains relevant terms like 'the game of Go', 'Go server', 'Go program', and 'Windows'. But the licensing page contains none of these terms. And the homepage doesn't even contain the terms 'gpl', 'license', or 'licensing'. So in the case of Gnu Go, there is a complete separation between the description of the software, and the licensing information. And this, in turn, means that it is nigh on impossible to find any of the official Gnu Go pages in a web search for GPL-licensed Go-playing software. It is, however, possible to find Gnu Go on Google by doing a search for 'gpl "go program"' – the first result is http://www.gnu.org/software/gnugo/free_go_software.html, which is a list of 'free go programs' hosted at *gnu.org*. But it is important to note both that this page still does not actually say that Gnu Go is GPL licensed (the references to GPL are all actually in relation to other free software), and nor would it even be definitive if it did (there are actually plenty of web pages that do say that Gnu Go is GPL licensed, but none of them actually create a licence offer, because they are all statements being made by third parties).

Creative Commons have attempted to solve this problem by coming up with a mechanism for marking works as GPL or LGPL licensed in a similar way to the Creative Commons licences and the public domain dedication. They have created graphics,⁴⁸ similar to the 'CC Some Rights Reserved' and 'PD Public Domain' graphics; they have a page on creativecommons.org that can be linked to to represent

⁴⁵ The page <http://www.oasis-open.org/cover/> is incorrectly found by Yahoo's Creative Commons search (as at 2006/11/10). This highlights the problem of link-based searches. They neither retrieve all relevant results, nor do they return only relevant results.

⁴⁶ For example, <http://www.openclipart.org/> (as at 2006/11/10).

⁴⁷ For more on Go, <see: [http://en.wikipedia.org/wiki/Go_\(board_game\)](http://en.wikipedia.org/wiki/Go_(board_game))>.

⁴⁸ The GPL graphic is located at <http://creativecommons.org/images/public/cc-GPL-a.png>; the LGPL graphic is <http://creativecommons.org/images/public/cc-LGPL-a.png>.

(L)GPL licensing;⁴⁹ and they have some RDF that can be used to state, in ‘machine readable code,’ that a work is GPL or LGPL licensed.

It is worth noting that in the case of GPL and LGPL, performing these actions to mark something as licensed under the GPL is not necessarily sufficient for derivative works to comply with the original licence requirements. This is because of the nature of the GPL, where there are strict restrictions on the way that licensing of derivative works must be achieved. As such, the most significant purpose of such ‘stamping’ of web pages is to allow search engines to find these works. At the moment, this is not being used much, but technically, it is a much easier way to search for (L)GPL-licensed works than searching for phrases that indicate licensing.

None of Google, Yahoo or Nutch/CC seem to support searching for this stamping.⁵⁰

3.3.3 Comparison

The following table summarises the various search-engine features that can be useful when searching for online commons, and which of the three search engines (Google, Yahoo and Nutch/CC) support them.

Feature	Google	Yahoo	Nutch/CC
Multiple CC jurisdictions	Yes	<i>No</i>	Yes
‘link:’ search term	<i>No</i>	Yes	<i>No</i>
RDF-based CC search	Yes	<i>No</i>	Yes
link-based CC search	<i>No</i>	Yes	Yes
Media-specific search	<i>No</i>	<i>No</i>	Yes
Shows licence elements	<i>No</i>	<i>No</i>	Yes
CC public domain stamp	Yes	Yes	Yes
CC-(L)GPL stamp	<i>No</i>	<i>No</i>	<i>No</i>

This comparison shows that the technical mechanisms to facilitate searching for commons are by no means standardised, and, not surprisingly, the dedicated commons-based search engine Nutch scores highest over all.

⁴⁹

<http://creativecommons.org/licenses/GPL/2.0/>
<http://creativecommons.org/licenses/LGPL/2.1/>, respectively.

and

⁵⁰ For example, <http://nickgravgaard.com/windowlab/> is correctly GPL stamped, but none of the three search engines can find it (as at 2006/11/10).

3.3.4 Google Code Search

On 5 October 2006, Google released *Google Code Search*, a new product that is still in Google Labs (essentially it is in beta).⁵¹ ⁵² It searches publicly available source code on the Internet, including web-accessible source code, and source code from repositories including CVS and Subversion repositories.⁵³ Its functionality for searching for publicly available source code far surpasses conventional search engines.

Google Code Search is not explicitly limited to publicly licensed source code, but it does recognise 18 public software licences. These include the GNU GPL, GNU LGPL, the Apache License, the BSD License and the MIT License. The search also returns results with “Unknown License[s]”, but in practice it appears that most results are licensed with one of the recognised licences. There are also examples of source code licensed under Creative Commons licences – a search for the following text returned “about 100” results:⁵⁴

this work is licensed under the creative commons

Additionally, because of the nature of source code – i.e. that it is more useful for re-using than for reading online (unlike web pages) – it seems likely that that online source code generally *is* intended for re-use.

Unfortunately, Google Code Search does not have an option to restrict search results based on the country of origin. However, it may be that country-specific search is not as relevant for source code. Source code is not a class of document that lends itself to being published on the web in HTML format, and in fact is usually released in a compressed format. Historically, this has probably led to web searches for source code not being very useful, with source code tending to end up in more useful databases such as SourceForge⁵⁵, where it can be searched based on metadata and descriptions. Because of the centralised nature of repositories such as SourceForge, country-of-origin information is likely to be lost. Additionally, multiple people often contribute to Free and Open Source software projects and, by the international nature of the Internet, they are likely to be from multiple countries, so country-of-origin information may be less meaningful.

4. Quantifying Australia’s Online Commons

The obvious first issue here is what is meant by ‘Australia’s commons’. There is no simple answer that makes deciding this easy. A reasonable approach might be to start from consideration of the *relevance* of ‘Australia’s commons’ compared with the

⁵¹ <<http://www.google.com/intl/en/press/annnc/codesearch.html>>. Google Code Search is at <<http://www.google.com/codesearch>>.

⁵² *Google Labs FAQ*, <<http://labs.google.com/faq.html>>.

⁵³ <http://www.google.com/intl/en/help/faq_codesearch.html>.

⁵⁴

<<http://www.google.com/codesearch?q=this%5C+work%5C+is%5C+licensed%5C+under%5C+the%5C+creative%5C+commons>>.

⁵⁵ <<http://sourceforge.net/>>

commons more generally. If we consider that the motivation of research into Australia's commons is that such works are either more relevant to, or more attractive for use by Australians, then we can define 'Australia's commons' simply as follows:

Commons content that is either created by Australians, hosted in Australia, administered by Australians or Australian organisations, or pertains particularly to Australia.

Informally, this includes all commons content that relates to Australia in any particular way and is hence more likely to be of relevance to Australians than to non-Australians.

For the purposes of this research, Australian content is construed narrowly to be content that either is created in Australia, or is hosted on Australian web sites. Specifically, this includes only the classes of content that: a) use Australian licences, suggesting that the licensors are Australian, or, b) are hosted either in the .au top level domain, or are considered Australian sites by the relevant search engine (where such search engines have an option to restrict to Australian sites). The issues relating to using proprietary search engines are discussed in the next subsection.

In this section, a few types of commons are quantified. This starts with AEShareNet's licences, followed by Creative Commons' licences, and lastly some Free Software licences. By no means is this analysis all-encompassing in mapping Australia's online commons. Rather, it aims to provide an overview of methods for use with a broad cross-section of public rights. AEShareNet is interesting because of its particular Australian focus; Creative Commons is both popular and technically interesting in its implementation; the GNU licences that are considered give a good example of a class of commons that is fundamentally difficult to find and quantify.

4.1 Using Search Engines for Quantification

Park and Thelwall give a good coverage of the issues of using traditional (proprietary) search engines to aid in hyperlink quantification, in *Hyperlink Analyses of the World Wide Web: A Review*.⁵⁶ The authors classify hyperlink analysis into two categories: hyperlink network analysis and webometrics. The former, hyperlink network analysis focuses on patterns in hyperlink *networks*.⁵⁷ However, in the current research on online commons, the main issues will be those of filtering sites based on being Australian in some way, and secondly, quantifying links to specific licence pages.

The other field of hyperlink analysis, webometrics, drew its original motivation from applying the aims and methodology of bibliometrics to the Web. Again, the focus is on the Web as a network of hyperlinked documents, from which point aspects such as the density of hyperlinks around a particular page or subset can be used to analyse the importance of the page or subset.⁵⁸ This is not what is being done in the current

⁵⁶ H W Park and M Thelwall, "Hyperlink Analyses of the World Wide Web: A Review" (2003) 8 *Journal of Computer-Mediated Communication* (hereafter referred to as *Hyperlink Analyses*).

⁵⁷ *Ibid.*

⁵⁸ *Ibid.*

research, but some of the problems that are presented in *Hyperlink Analyses* are also relevant here.

One of these problems is that proprietary search engines are opaque. From a research perspective, we would certainly like to be able to see inside the box and verify that things are working as they should. Another problem may be that the search engine index may not include all pages on the web, although this is a problem that has lessened in recent years.⁵⁹

There are, however, two good reasons to use proprietary search engines for this research. Firstly, it is significantly easier than creating a search engine especially, and proprietary search engines have the advantage that at least their public interface is available world-wide, on the Web.

The second reason to use a proprietary search engine is that it mirrors the way consumers can search for commons content on the Web. For example, if a document can't be found using a search engine, but *is* actually available online (if one knows where to look), it is simply not an *effective* part of the online commons for the simple reason that people are not going to be able to make use of it, even if they would want to.

The issue of restricting searches to Australian web pages is covered in the next subsection.

4.2 Using Yahoo to Search for Australian Content

Yahoo Advanced Search⁶⁰ has an option to restrict searches to “pages from Australia”. Although this is what is needed here, Yahoo has little documentation on exactly what this means. Yahoo documentation on the Australian search interface says of the ‘Country’ field only:⁶¹

Yahoo!7 [sic] normally searches through Web sites from all over the world. You can choose to see sites just from a particular country.

This gives little to go on, and perhaps suggests that the feature is as simple as making sure that, for sites to qualify as Australian, their domain names must end in ‘.au’. Another simple yet not entirely appropriate rule might be that the word ‘Australia’ or ‘Australian’ must be present.

Google gives slightly more documentation:⁶²

While all sites in our index return for searches restricted to "the web," we draw on a relevant subset of sites for each country restrict. When searching for pages from a specific country, keep in

⁵⁹ Ibid.

⁶⁰ <<http://search.yahoo.com/search/options>>, also available through Yahoo Australia at <<http://au.search.yahoo.com/search/options>>.

⁶¹ <<http://help.yahoo.com/help/au/ysearch/basics/basics-08.html>>.

⁶² <<http://www.google.com/support/bin/answer.py?answer=469>>.

mind that our crawlers identify the country that corresponds to a site by factors such as the physical location at which the site is hosted, the site's IP address, and its domain restrict.

Even though Google is more open about its internals in this respect, Yahoo will still be used because of Google's failings in the area of hyperlink specific searches, as discussed in section 0 above. Nevertheless, Google's description of its country restrict may shed some more light on Yahoo's search, given the likelihood that Yahoo's and Google's technology are similar. To test these ideas, various queries can be tested on Yahoo, and the results can be analysed to ascertain why they might qualify as "pages from Australia".

The page <www.evfit.com/food.htm> is returned as the first result of a search for "food -australia -australian -site:.au" with the "just pages from Australia" option enabled. The search query ensures that any results are not being classified as Australian simply by referring to Australia, nor by being part of the Australian top level domain.

This page appears to be correctly classified as relating in some way to Australia. Running the Linux program *traceroute* to trace the Internet Protocol route to the host, we can see that the server hosting this web page is connected directly to a server of an Australian Internet service provider (ISP), *Leading Edge Internet*.⁶³

```
10  bytecard-gw.cor1.lei.net.au (203.221.207.241)  29.138 ms
    32.120 ms 30.004 ms
```

```
11 evfit.com (210.11.144.75) 30.712 ms 30.939 ms 31.906 ms
```

The next test was a search for "computer", with otherwise the same parameters. The first result here was <www.computertechnicalcentre.com>. This page clearly pertains to a computer retailer in Cairns, in Queensland, so is correctly classified, but contextually there is no reference to Australia. A *traceroute* test shows that the web site is hosted in California, USA, through an ISP called *Multacom*. However, in this case, a WHOIS⁶⁴ search reveals that the domain name registration for *computertechnicalcentre.com* is associated with an address in West Wollongong, New South Wales. Again, the conclusion is that Yahoo is correct to classify this search result as an Australian-related page.

The final example comes from a search for "law", and the first result is <www.au.af.mil/au/awc/awcgate/awc-law.htm>. This is an American military university web site, and has nothing to do with Australia. The only plausible explanation for its inclusion in the results of this search is that the domain name contains the word 'au'. In this case, however, 'au' stands for 'Air University' and not 'Australia' as Yahoo is apparently assuming.

This final counterexample shows that, although Yahoo may be otherwise very good at determining which pages are Australian and which are not, it is not infallible.

⁶³ This *traceroute* was run from an Ubuntu Linux server connected to Exetel (<www.exetel.com.au>) as Internet service provider.

⁶⁴ WHOIS is a protocol for discovering information about the registration of a domain name. The definition is at <<http://www.rfc-editor.org/rfc/rfc3912.txt>>.

4.2.1 Usage of AEShareNet 'Instant' Licences

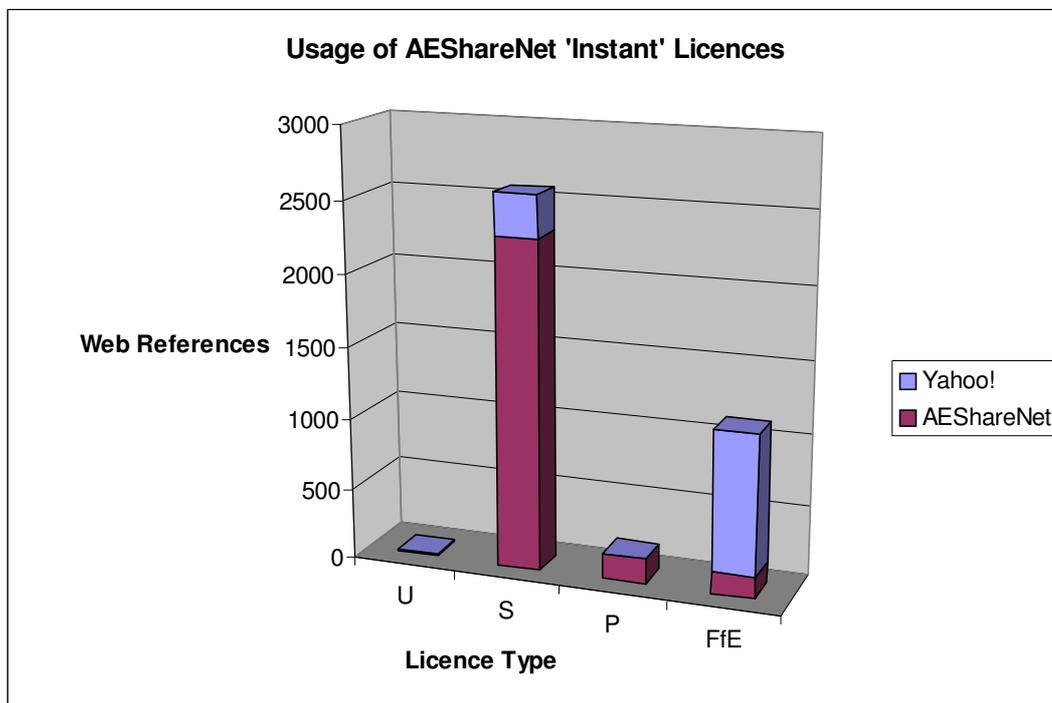
The following data was obtained by:

- Searching on <http://www.aesharenet.com.au/> for education materials that are marked as 'available';
- Searching the Internet using Yahoo for pages that link to AEShareNet 'instant' licences that are not hosted on <http://www.aesharenet.com.au/>.

Table 1. AEShareNet Data

Licence Type	Yahoo Hits	AEShareNet Hits
U	1	11
S	289	2,292
P	1	180
FfE	979	148

Image 1. AEShareNet Graph



This graph shows four interesting results:

- *aesharenet.com.au* has far more licence information than the Web (as indexed by Yahoo) (33% of results were from Yahoo, 67% were from *aesharenet.com.au*).

- On *aesharenet.com.au*, the ‘Share and Return’ licence type is by far the most used (of the instant licences).
- On the Web, as indexed by Yahoo, and excluding the AEShareNet website, the ‘Free for Education’ licence type is by far the most used.
- It is much more common to publish something that is Free for Education on the Web than it is to register it on *aesharenet.com.au*.

4.2.2 Australian Usage of Creative Commons Licences

The data in this subsection were obtained by using Yahoo to search for links to Creative Commons licences. When gathering any data on Australian usage of the Creative Commons licences, whether using hyperlink quantification or otherwise, it is important to address the issue of multiple jurisdictions. There is nothing stopping an Australian licensor from using a licence from another jurisdiction, even though this is clearly not how the Creative Commons system was intended to be used. In practice, there is no reason to think that a significant number of Australian would be using the licences of other jurisdiction, with the significant exception of United States licences. The reason for this is not that the United States licences would be considered any more pure or robust by Australian licensors – rather, it is simply that on the Creative Commons web site, the default jurisdiction was for a long time United States. The default is now the new ‘generic’ or ‘unported’ licence, but this has only been the case since 23 February 2007, when Creative Commons announced the version 3.0 licences, which make a distinction between United States jurisdiction licences and generic ‘unported’ licences.⁶⁵ As of 1 March 2007, there appear to be no Australian websites using these licences (as searched by Yahoo).

In gathering this data, a total of forty-two distinct searches were performed, comprising:

- A search for each of six licence types for each of five licence versions commonly used in Australia: US 1.0, US 2.0, US 2.5, AU 2.0, AU 2.1 and AU 2.5.
- A search for each of the five deprecated licences from the US 1.0 version. The data from these searches has been included in the data for the corresponding ‘by-’ licence in the US 1.0 set. The justification for this is that this is essentially what happens with later versions of the licences. For example, if somebody wants to licence their work under a current Australian Creative Commons licence and is only interested in making sure that the work is not used for commercial purposes, they will choose the by-nc licence, even though they are not particularly interested in requiring attribution.
- In version 1.0, the naming convention was slightly different, and instead of ‘by-nc-nd’, ‘by-nd-nc’ was used. Now, both names for the licence are valid, and so a search for either alias will return (different) results.
- On 23 February 2007, Creative Commons announced version 3.0 licences, which make a distinction between American and generic ‘Unported’

⁶⁵ < <http://creativecommons.org/weblog/entry/7249> >

licences.⁶⁶ As of 1 March 2007, there appear to be no Australian websites using these licences.

Results were restricted to Australian sites using Yahoo's country-specific search feature.

4.2.3 Australian Creative Commons Usage Data

The following table shows, for each combination of derivative-related attributes and commercial-related attributes (or lack thereof), approximately how many Australian web pages are linking to each licence version, according to the Yahoo search engine.⁶⁷

The 'US' columns represent Australian pages (according to Yahoo) that link to United States licences. The 'AU' column represents Australian pages (according to Yahoo) that link to Australian licences. This data does not include two classes of (otherwise relevant) Australian licensed work: first, there is a class of Australian content that uses United States licences and is not hosted on an Australian web site (as determined by Yahoo). Such works give so little indication of being Australian that these methods cannot identify them as such. Correspondingly, there is the class of works that link to Australian licences but are not hosted on Australian websites. The latter class can be quantified, but to do so here would skew the data towards Australian licence usage and invalidate any comparisons between Australian use of Australian licences vs. Australian use of United States licences.

This does not include the new version 3.0 US or 'Unported' licences, for which results would be negligible as described in Section 0 above, above.

Type	US 1.0	US 2.0	US 2.5	AU 2.0	AU 2.1	AU 2.5
by	574	4,850	5,640	31	1,210	468
by-sa	402	2,660	3,620	1,470	439	2,520
by-nd	1,870	1,040	411	268	162	1,980
by-nc	2,564	2,890	8,540	635	1,850	1,500
by-nc-sa	10,120	11,200	16,300	1,020	10,400	3,010
by-nc-nd	4,474	13,300	5,490	1,280	5,160	7,660

4.2.4 Australian Usage of Creative Commons by Licence Type

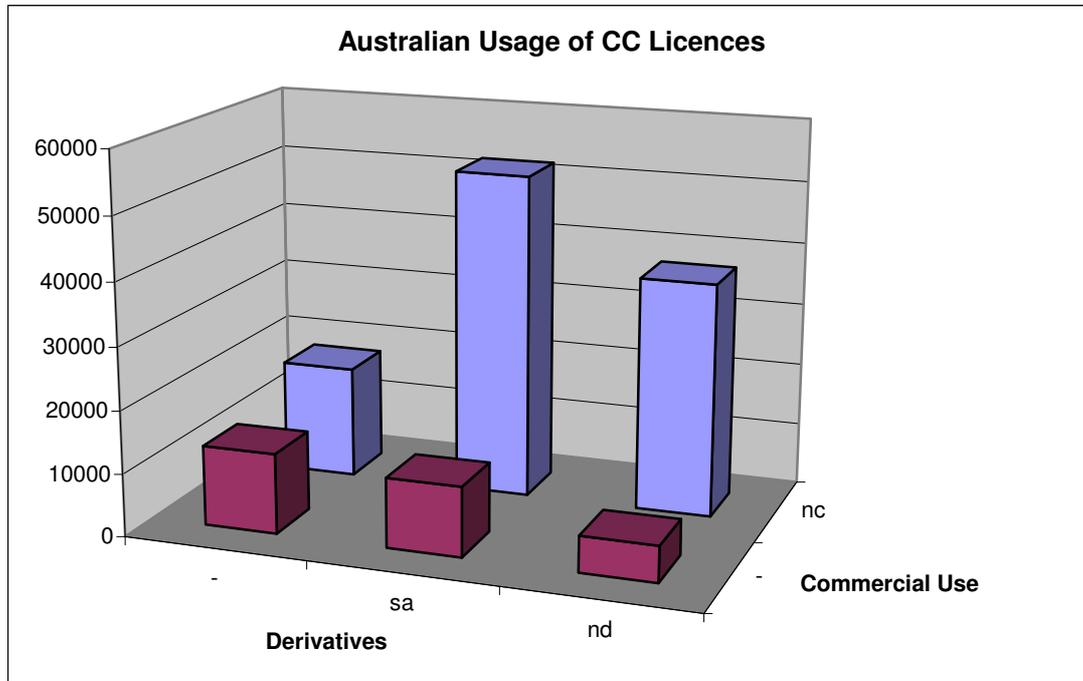
Totalling the rows, we can see the relative frequency of use of the various types of Creative Commons licences in Australia. The data is presented here organised by licence element, with one vertical bar for each licence type. The horizontal axis

⁶⁶ <<http://creativecommons.org/weblog/entry/7249>>

⁶⁷ This data was obtained on 7 March 2007.

represents restrictions on modification (no derivatives, share-alike, or neither). The depth axis represents restrictions on commercial use (non-commercial use only, or not). For example, the largest vertical bar, in the middle at the back, represents Australian usage of Attribution-NonCommercial-ShareAlike licences, without regard for the different jurisdictions.

Image 2: Australian Usage of CC Licences

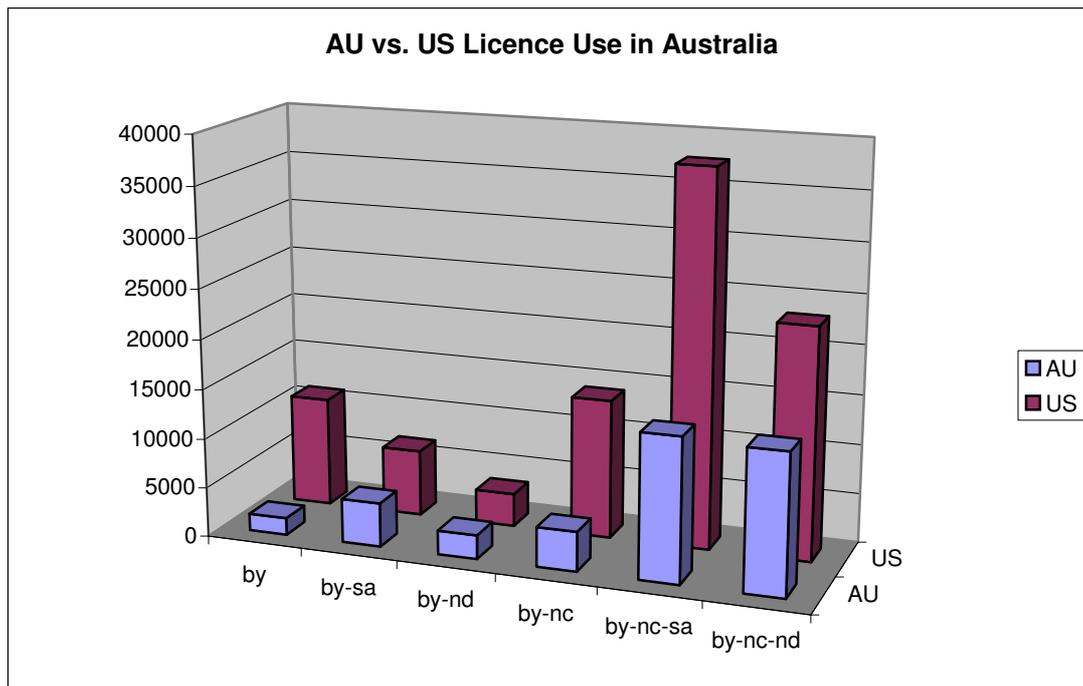


This graph shows a clear preference for restrictions on commercial use. Possibly, the reason that ‘by-nc’ is not as well used as ‘by-nc-sa’ and ‘by-nd-nc’ is that licensors are worried (whether justified or not) about licensees making derivative works, re-licensing them, and then having them used for commercial purposes.

4.2.5 Australian vs. US Creative Commons Licence Usage

It is interesting to compare, for each licence type, how many people are using United States’ jurisdiction licences and how many people are using Australian licences (on Australian sites). Note that this still refers to solely Australian usage, but differentiates between licensors who have chosen to use Australian licences, and those who have chosen the *de facto* default jurisdiction of the United States.

Image 3: AU vs. US Licence Use in Australia



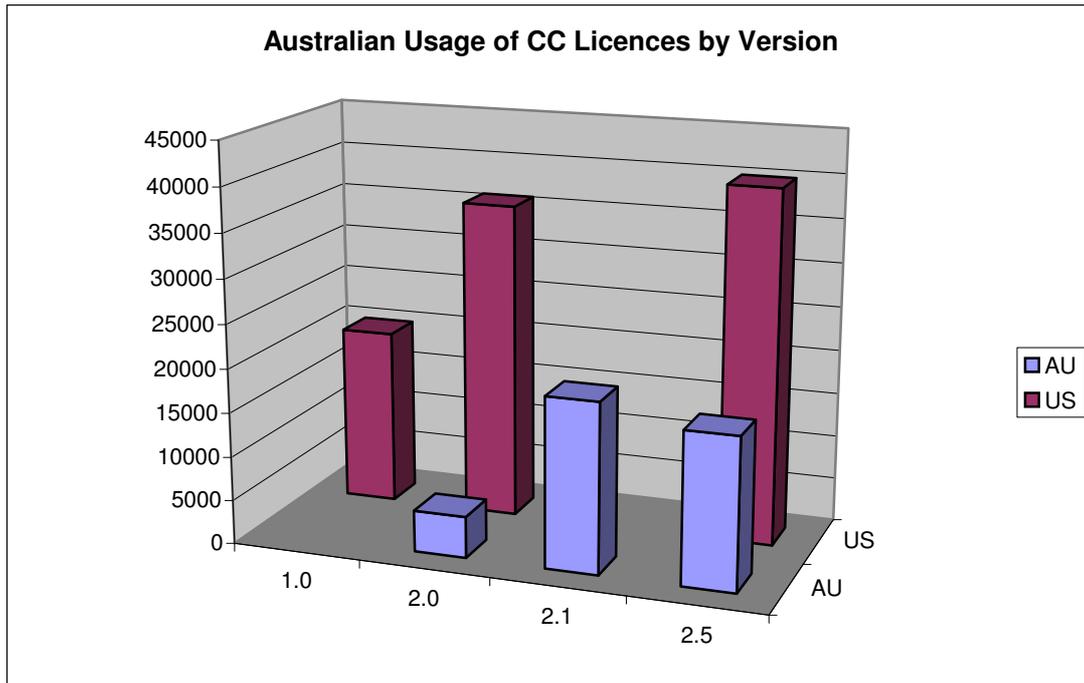
While use of ‘AU-by-nc-nd’ and ‘AU-by-nc-nd’ licences are approximately equal, significantly many more pages are licensed with ‘US-by-nc-sa’ than with ‘US-by-nc-nd’. This shows that, for one reason or another, the variation in usage of Australian licences is not following the variation in usage of American licences. This may be due to a slight difference in demographic between licensors choosing Australian licences and those choosing American licences.

4.2.6 Australian Usage of Creative Commons Licences by Version

To get a good idea of current motivation for publicly licensing works, what we really need to know is which licences are being used at present, for newly licensed works. That is, a distinction needs to be made between works that are currently licensed, but have been licensed with their current licence for a long time, and works that are being licensed now. This is very difficult to do, except by watching the increase in licence statistics over time. But if we take in to consideration the tendency of people to use the current version of the licences, we can consider usage statistics for older versions of the licences to be snapshots of licence usage from the point in time that they were superseded. On the other hand, we cannot consider the United States licence data to be unchanging simply because there are now Australian licences that are a better choice for licensors, because the process of licensing works still makes it very easy to use the United States licences, either through ignorance of the existence of the Australian versions, or through mistake (choosing a non-US licence is not mandatory and requires positive action by the licensor; the default is US).

The following graph shows the relative use of each version of the licences:

Image 4: Australian Usage of CC Licences by Version

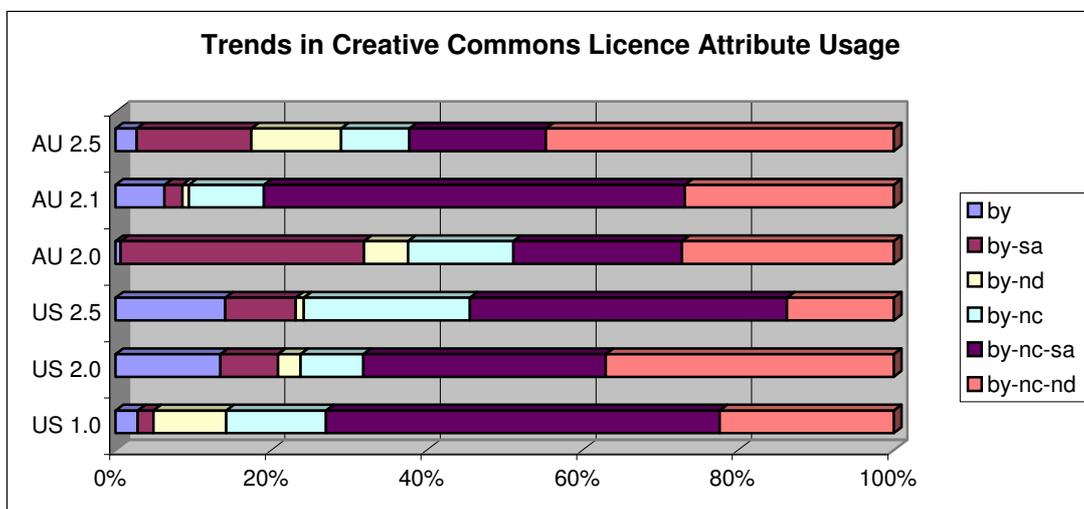


This graph clearly shows that, even though Australian versions of the licences are available, the tendency for people to use the American licences is still significant.

4.2.7 Trends in Creative Commons Licence Attribute Usage

From the data alone, it is very difficult to see the trends with respect to which licences are being used. The following graph shows, for each licence version, what proportion of those licences are made up of each licence element. If, as one might expect, the versions are uncorrelated with the licence types, we would expect the following graph to show an identical bar for each licence version, with the percentage contributed by a given licence type simply being the same as the percentage of total licences that use that type. The relative usage of the various licence types is still clear (e.g. 'by-nc-sa' has a much bigger presence than 'by-nc'), but the colours differ in size from row to row.

Image 5: Trends in Creative Commons Licence Attribute Usage



4.3 GNU-Licensed Works

4.3.1 GNU General Public Licence

Although there are over 100 ‘official’ Free Software and Open Source Software licences,⁶⁸ the GNU GPL is possibly the most common such licence in use today. This is probably attributable to its age, its being designed to be a general purpose Free Software licence, and that it is the flagship licence of the Free Software Foundation. Therefore, it is of great importance that GPL (and LGPL) based searches be possible.

As explained earlier,⁶⁹ search engines have fundamental problems finding GPL/LGPL/FDL licensed works. Still, it is possible to go some way towards quantifying the number of copies of the GPL on the Internet. Doing a Google search for pages that contain all of the following fragments of the GPL yields approximately 114,000 results globally, and approximately 134 from Australia:⁷⁰

"This General Public License applies to most"

"change it. By contrast, the GNU"

"this free software. If the software"

"distinguishing version number. If the Program"

4.3.2 GNU Lesser General Public Licence

The above method is also appropriate for the LGPL. A Google search for the following yields 28,900 results globally, and approximately seventeen from Australia:⁷¹

"This license, the Lesser General Public License"

"we copyright the library, and"

"verbatim copies of the Library's complete"

5. A Vision of the Future

Creative Commons’ contribution to the online commons movement is more than just the creation, promotion and spread of the various Creative Commons licences, though of course the significance of that should not be underestimated. But through the use of embedded RDF metadata in XML format, Creative Commons has started a process that has the potential to spread far beyond the original eleven licences. The use of RDF for marking pages as publicly licensed is already common enough for Google to be making use of the technology. Given that the first step has been made, it seems

⁶⁸ See Section 0, above.

⁶⁹ See Section 0, above.

⁷⁰ Google (Australia) ‘the Web’ search and ‘pages from Australia’ search, respectively.

⁷¹ See note 70, above.

likely that as the use of embedded usage rights information grows, Google and any other search engines that are using RDF will continue to develop to keep up.

Now the question is this: what changes can we expect with respect to embedded usage information? First, it is important to note that RDF is not specific to Creative Commons. RDF is designed to represent metadata, and the concept of metadata is infinitely broader than just public rights information.⁷² What this means is that Google, Nutch/CC, or any other search engine, need not restrict its indexing to only RDF that refers to Creative Commons licences. Consider the following hypothetical RDF from a web page that is licensed under AShareNet-P:

```
<rdf:RDF>
  <Work rdf:about="">
    <license rdf:resource="http://www.aesharenet.com.au/P4/" />
    <dc:type rdf:resource="http://purl.org/dc/dcmitype/Text" />
  </Work>
  <License rdf:about="http://www.aesharenet.com.au/P4/">
    <permits
rdf:resource="http://web.resource.org/cc/Reproduction"/>
    <permits
rdf:resource="http://web.resource.org/cc/Distribution"/>
    <requires rdf:resource="http://web.resource.org/cc/Notice"/>
    <requires
rdf:resource="http://web.resource.org/cc/Attribution"/>
    <prohibits
rdf:resource="http://web.resource.org/cc/CommercialUse"/>
    <prohibits
rdf:resource="http://web.resource.org/cc/DerivativeWorks"/>
  </License>
</rdf:RDF>
```

This may not be the best expression of the elements of the AShareNet-P licence, but the important thing is that a search engine, such as Google, if it is willing to look beyond Creative Commons licences, could see this RDF embedded in a web page and decide that the web page qualifies as ‘free to use and share’, and include this web page in searches for such pages. In fact, it is possible that Google is already capable of doing this, and that the only reason we are not seeing such results is that they are not yet present on the web. The point is that Google knows the *meaning* of the various licence elements (‘DerivativeWorks’, ‘Reproduction’, etc.), and if other web pages use the same terms, there is no reason for Google not to use them in the same way.

From this, we can start to get an idea of where online commons search could be heading. If RDF is really taken up by licensors (with the help of licensing facilitators like Creative Commons and AShareNet), it is possible that, one day, most works with public rights on the Internet will have their public rights described in terms of the

⁷² For more information, see *What is RDF on xml.com*: <http://www.xml.com/pub/a/2001/01/24/rdf.html>.

current rights and responsibilities (reproduction, notice, commercial use, etc.), and some new ones (share and return, educational use, etc.). In fact, it is possible to go beyond the requirement that the RDF be embedded in the licensed page, by using the ‘rdf:about’ element of the ‘Work’ node. For example, consider the following hypothetical example:

```
<rdf:RDF>
  <Work rdf:about="http://www2.tafe.sa.edu.au/lili/">
    <license rdf:resource="http://www.aesharenet.com.au/S4/" />
  </Work>
  <License rdf:about="http://www.aesharenet.com.au/S4/">
    <permits
rdf:resource="http://web.resource.org/cc/Reproduction"/>
    <permits
rdf:resource="http://web.resource.org/cc/Distribution"/>
    <requires rdf:resource="http://web.resource.org/cc/Notice"/>
    <requires
rdf:resource="http://web.resource.org/cc/Attribution"/>
    <prohibits
rdf:resource="http://web.resource.org/cc/CommercialUse"/>
    <permits
rdf:resource="http://web.resource.org/cc/DerivativeWorks"/>
  </License>
</rdf:RDF>
```

This RDF essentially makes the following statements:

The web page <http://www2.tafe.sa.edu.au/lili/> is licensed with licence <http://www.aesharenet.com.au/S4/>. The licence <http://www.aesharenet.com.au/S4/> permits reproduction, distribution and derivative works, requires notice and attribution, and prohibits commercial use.

This is not entirely accurate, because the current RDF usage rights language does not contain a term for the ‘Share-and-Return’ provision that means that the original licensor owns copyright in derivative works, but it is close. This RDF doesn’t have to be part of the web page at <http://www2.tafe.sa.edu.au/lili/> (although that way has the advantage of giving people confidence that the page is licensed). Instead, another web page, located anywhere on the web, could contain this RDF. For example, AShareNet could act as a repository for such RDF, making statements about various web sites and their licensing.

Clearly, there is the potential for licensing information to become an integral part of page source, and to expand beyond just the (already significant) variety of licences promoted by Creative Commons.

6. Conclusion

From the point of view of finding Australia's online commons, there are already a number of techniques that are quite useful, and the power of search engines to search for online commons is bound to grow. Yahoo is very useful for finding links to licences; Google is good from a technological purist's point of view, finding only those Creative Commons licensed materials that correctly embed RDF; Nutch/CC is showing potential to be a very useful specialised tool for searching for Creative Commons-related works, and clearly has a very well implemented knowledge of Creative Commons licensing.

At present, when doing a search some thought must be given to what kinds of results are sought, and accordingly what is the best way to search. A search for Australian education materials should start on *aesharenet.com.au*. A search for free software can only be done on a licence-by-licence basis, and needs to include some appropriate keywords or licence fragments. A search for source code can be done quite effectively using Google Code Search. Searches for photos would be best done using Nutch/CC. General searches for text, including weblogs, etc., can be done using Google or Yahoo.

From the point of view of quantifying Australia's online commons, the data clearly shows that the Creative Commons suite of licences has been enthusiastically taken up by Australians. Unfortunately, it is clear that the US-jurisdiction licences continue to be chosen for licensing, even in the presence of more appropriate Australian versions.

Lastly, looking to the future we can see that new technologies, as part of the 'Semantic Web' idea, have the potential to make all of Australia's online commons easily searchable and quantifiable. Licensors will still have to go to some modicum of effort to make sure their works are found, but that is both unavoidable, and exactly how it should be.

6.1 Future Work

The research presented here constitutes a baseline study. It gives an overview of the current state of the art, and gives some indicative figures on licence usage. From this starting point, statistics on licensing can be compared over time, using the same methods. If and when new methods are developed, dual quantification can be done to establish correction factors.

The field of exploration of the online commons, especially national commons, is quite new. Google and Yahoo, two of the biggest Internet search engines, have made the first steps on the long path to comprehensive coverage, and it seems reasonable to expect that these features will continue to be improved. The table in 0 shows that all three of the search engines considered have room for improvement, and Section 0 started to lay out some possible improvements beyond what any of these search engines are currently providing. However, some very significant advances in search engine technology, possibly involving artificial intelligence, may be required to cover the Free and Open Source Software licences and the way they are being used on the web, as seen in the Gnu Go example in Section 0.

In future work, more consideration will be given to the ideal features of a search engine for providing public rights-based search, as well as tracking and suggesting developments in the technical field, and analysing trends in the data.