Contact details:

Ben Bildstein
E-mail: ben.bildstein@student.unsw.edu.au
Blog: http://www.cyberlawcentre.org/hoc/
Phone: +61 408 134466
Post: GPO Box 2330 Canberra Australia 2601

Affiliations:

PhD candidate in the University of New South Wales Law Faculty;
Scholarship funded through the Australian Research Council-funded Unlocking IP project at the Law Faculty's Cyberspace Law and Policy Centre

Short biographical note:

Ben Bildstein is a PhD researcher at the Cyberspace Law and Policy Centre at the University of New South Wales. His undergraduate background is Computer Systems Engineering at the University of Tasmania where his honours thesis was a neural network / genetic algorithm approach to the board game of Go, and for a few years between degrees he worked as a professional software developer in the environmental data management industry. His current research is in the quantification of works with public rights online.

Title:

New methodologies for quantifying licence-based commons on the web

Abstract:

The movement towards a copyright commons has undeniable momentum, but actually measuring the current state of this commons, its rate of growth, or its composition is a significant challenge. With the wealth of licences available to use for licensing works, and the variety of media that can be licensed, there is much scope for interesting analysis of comparative size and trends. Questions that we would like answered include: Which Creative Commons licence attributes are most common? Which attributes are most popular in licences currently being chosen? How do various licences or licence regimes compare in popularity (e.g. GNU vs. Creative Commons)? How many copyright works are licensed with the various licences? To what extent are copyleft licences encouraging licensing of derivatives?

The current practice in quantifying licence-based commons on the web involves the use of web search engines such as Yahoo, Google, All The Web and Live Search. Searches can be performed that find links to licences such as Creative Commons licences (on a licence-by-licence basis), and then the search engines will provide an approximate number of search results. The statistics presented at <http://wiki.creativecommons.org/License_statistics> use this methodology.

The practice of using search engines for quantification has some disadvantages. Perhaps the most significant of these actually relates to the notion of using hyperlinks for quantification, rather than using any written statement of licence offer. That is, a link to a licence is not

definitively a licence offer, because the document that links to the licence may only be referring to the licence (i.e. making a statement about it). The other disadvantage of using proprietary search engines is that the accuracy of the results is largely unknown. Firstly, the results are not exact, but represent only indicative values, and second, the precise coverage of the search engine indexes is not known.

Another fundamental problem with using search engines is that the internal workings are hidden. One solution to this is crawling the web for the purpose of quantification. This allows for the collection of exact numbers, and it is the only way to have access to the raw HTML of the pages (conventional search engines do not have facilities for querying the underlying HTML). An example of where the latter is important is RDF metadata. Creative Commons advocates embedding metadata in RDF expressed as XML in comments inside the HTML of the page being licensed. But none of the major search engines index RDF in any way that is useful for quantification - it is as if the RDF is lost, ignored, or at least well hidden. (Actually, we know that Google does index the semantics of Creative Commons-oriented RDF, because it has been shown that Google is able to identify a page as Creative Commons licensed when the only known indication of licensing is the RDF in an XML comment, but this is only accessible indirectly through the 'usage rights' advanced search option.) For accurate and reliable data, there is a clear need for a new approach to quantifying and tracking the growth of online commons. This paper addresses the question of how viable a crawl-based approach is, and discusses the advantages and disadvantages over using proprietary search engines.

Also addressed is the question of what metric to use to measure the number of works licenced. That is, what is the most appropriate unit of measurement for the online commons? Current practice largely ignores this question, implicitly counting every web page that links to a licence as a separate creative work. However, results based on such a metric may be misleading. For example, a blog may reasonably be considered a single creative work, but it is very common for every post to a blog to be given its own 'permalink', making every post a separate web page. In this case, although the blog might be licensed as a single work, it may show up in statistics as dozens or even hundreds of 'licensed works'. Another example worth considering is that of an online encyclopedia, which may have hundreds of thousands of pages. If every page is considered a licensed work, the resulting statistics are incomparable to any statistics on paper-and-ink works, because a printed encyclopedia will count as only a single licensed work.

To investigate this issue, various metrics are used and compared. One such metric is that of licensed web sites, rather than individual pages. A site here can be delineated in one of two ways. The first is that a site extends precisely as far as the domain name it uses, meaning that all pages, and only those pages, that share a domain name in their URLs are considered to be from the same site. One problem with this is that multiple sites (in the traditional sense) do sometimes share the same domain name. For example, ISP-hosted personal web sites often share the domain name of the ISP. The second metric aims to overcome this problem by doing graph-theoretical analysis of the hyperlinks of licensed pages to find distinct groups of web pages that to not link to each other, and then consider such sub-webs to be distinct web sites.

Although this paper addresses many of the issues surrounding commons quantification, its scope is limited in many ways. First and foremost, it is only concerned with the online commons, and quantification of the offline commons is quite a separate area of research. Second, it does not address the issue of the deep web (that is, web document that are not linked to from other web documents, but are only accessible via direct entry of the URL or through submission of a form in another web page). The deep web has been shown to be of significantly greater size (in total number of documents) than the surface web, and when commons quantification research is eventually expanded to the deep web it may be that a fair proportion of the commons is hidden there. However, the hidden nature of the deep web combined with the open nature of the commons movement suggests that the majority of the commons might be in the more accessible surface web.

This paper is also primarily focused on link-based licences. Other licences, such as the GNU General Public Licence, do not require or even encourage licensed works to link to any copy of the licence. Although there are other techniques for quantifying such bodies of licensed works, and these are discussed, this is not the main focus here. Similarly, the traditional 'public domain' of un-copyrighted works is dealt with only lightly, as there is a great barrier to quantification in the lack of any metadata necessarily being associated with the licensed documents.

The ideas presented are tested in the context of a limited crawl of the web, restricted to the .gov.au sub-domain. The methodologies of proprietary search engine-based and crawl-based quantification are compared using this data set (using Creative Commons licences and AEShareNet licences). At the same time, the various metrics for measuring the number of works (web pages vs. web sites vs. distinct sub-webs) are compared using this crawl.