Three Strikes for Copyright

Author : Jessica Silbey

Date : October 13, 2017

Abhishek Nagaraj, *Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia*, **Mgmt. Sci.** (forthcoming 2017), available at abhishekn.com.

How should copyright law change to take account of the internet? Should copyright expand to plug the internet's leakiness and protect content that the internet would otherwise make more freely available? Or, should copyright relax its strict liability regime given diverse and productive reuses in the internet age and the benefits networked diffusion provides users and second-generation creators? Answering these questions depends on what we think copyright is for and how it is used and confronted by creators and audiences. In a new article studying these questions in the very focused setting of Wikipedia articles about baseball and baseball players (there are more than you might imagine!), Professor Abhishek Nagaraj demonstrates that where production of new knowledge depends on pre-existing information, strong copyright law can reduce both the quality and quantity of new content.

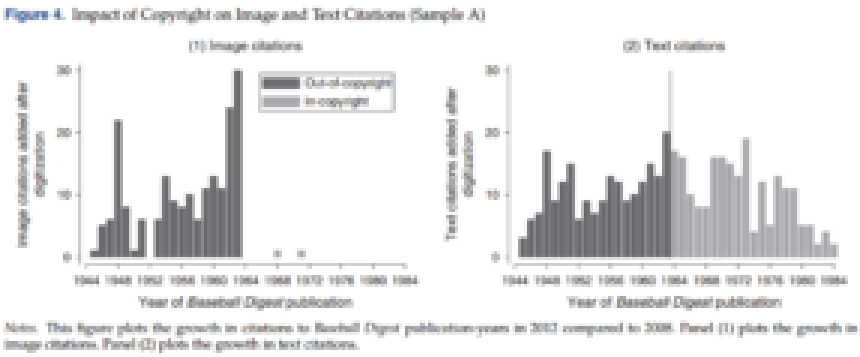
Professor Nagaraj studies the intersection of digital access and information diffusion. In the paper reviewed here, Nagaraj takes advantage of the lack of automatic renewal of copyrighted works published before 1964, rendering many to the public domain, to estimate the effect of access to public domain material on the quality of Wikipedia pages. His findings both confirm other studies in this area and raise new lines of inquiry.

This paper tells several stories. The first starts in 2008, when Google Books digitized all of the issues of *Baseball Digest* between 1940 and 2008. Of these, the pre-1964 issues are in the public domain; the rest remain under copyright. This first story explains how Wikipedia articles about baseball cite to the public domain *Baseball Digest* issues twice as much as the in-copyright sources. It is no surprise that digitization of an important source of information about baseball enables access and encourages the reuse of this resource on Wikipedia, the fifth-most visited website (with about 10 billion page views monthly). After digitization, citation to *Baseball Digest* increased 300% over pre-digitization levels. What surprises more is that public domain sources (which are also older) are cited more frequently than in-copyright sources, despite both being digitized fully by Google Books. As Nagaraj demonstrates, relying on quantitative analysis of citation frequency and open-ended survey questions with Wikipedians, copyright is a barrier to citation and reuse of the digitized material, and Wikipedians are paying attention to those barriers.

This paper tells another story about the consequence of the copyright barrier – i.e., that it diminishes the quality of the Wikipages about certain baseball players who played after 1964. For pages about baseball players who are neither famous nor obscure (e.g., the average player about whom a Wikipedia page would come in handy), those players who made their debut appearances before 1964 have higher quality pages than those who began playing after 1964. How does Nagaraj measure quality? By measuring what he argues is circumstantial evidence of higher quality content: citation to *Baseball Digest*, the number of images on the page, and the number of visitors to the page (as a measure of reader utility). Pre-1964 player pages for well-known (but not superstar) players have almost twice the number of citations to *Baseball Digest*, 1.78 images as compared to 0.92 for in-copyright player-pages, and they attract about forty-seven more visitors per month on average.

Nagaraj describes this effect in terms of a welfare impact, suggesting that pages negatively affected by copyright are unable to fully capture and deliver value to end users. In intellectual property debates, we often worry about quality over quantity, whether the "progress" to which the intellectual property clause of the Constitution aims is more stuff or better stuff, "better" being a tricky term. We also worry about the relevance of citation counts. A helpful and intriguing feature of Nagaraj's paper is his metric for quality that is both quantitatively measurable and qualitatively significant for the community the content serves (baseball fans).

The third story this paper tells is the most interesting of all. It concerns the differential impact of copyright restrictions on images versus text, which difference is driving the first story described above. Generally, digitization should lower the costs of reuse for both types of media, but Nagaraj shows that text is cited to at a significantly higher rate than images, leading to the reuse of and reliance on text at a much higher rate than images from the in-copyright *Baseball Digest* sources. This means that the digitization of content benefits textual content more than images (photographic or otherwise). Or, more precisely, the negative effects of copyright on citation and use disappear for text and are driven by a lack of reuse of images post-1964. Nagaraj hypothesizes and then demonstrates that because images require more "transformation" (under copyright fair use) than textual content to avoid infringement liability, the gains of access won by digitization are mitigated, for the in-copyright images, by copyright's imposition of greater transaction costs. It follows that the "reuse of out-of-copyright content is likely to be higher for images … than for text" (P. 16), suggesting that the public-domain status of images is more important for diffusion and reuse of this kind of information and content. In fact, Nagaraj shows that the "likelihood that an image will be reused from a post-1964 [in-copyright] issue of *Baseball Digest* is very close to zero, even after digitization [whereas] … the pattern for text citations is quite different. … In-copyright and out-of-copyright text citations track each other pretty closely, suggesting that copyright has very little impact on preventing the reuse of digitized textual material." (Pp. 17-18.) His chart reproduced below illustrates this point.



This last story raises fascinating questions. First, in a networked digital age in which photographs, logos, memes, and other graphic images have been shown to circulate more swiftly and broadly than text, in which images dominate the semantic web (including, importantly, journalism and social media), it is surprising that copyright may control the use of images as forcefully as Nagaraj documents about Wikipedia. This is a significant threat to veritable and free speech.

Second, the collection and management of photographs online is growing in concentration and expense. There are a lot of free images on the web, but they serve to entertain or illustrate, not to inform or contribute facts for debate. (There is a reason we see the same pictures of Mitch McConnell, Antelope Canyon, or the Lincoln Memorial on news sites). New, timely, and authenticated photos are harder to source; there are fewer photojournalists and even fewer aggregators and news agencies managing the collection and distribution of news. If, as Nagaraj's study proves, internet users rely on photographs to

anchor and explicate information sought, copyrighted photographs circulate less freely than text online (a statement many photographers would dispute, but that's another story), and because news outlets now less frequently pay for and distribute photojournalistic images, we have a significant information problem in the digital age.

Third, is it possible that Wikipedians are so closely hewing to the copyright fair use analysis that it is harder to claim transformative fair use of photographs than text? In my research, I found creative and innovative communities followed idiosyncratic norms of copying (or not copying) that did not align with intellectual property law. The story Nagaraj tells about the non-use of photographs versus the use of text under copyright on Wikipedia is a story of behavior arguably aligning with copyright doctrine. Reusing copyrighted photographs verbatim is harder under copyright fair use than quoting or paraphrasing parts of text. But if you surf the internet with its seemingly uncontrolled reproduction of photographs, you would be forgiven for thinking that copyright law doesn't act as a barrier to copying and distribution of photographs at all. So what explains the Wikipedian's careful non-use of photographs under copyright? Banners on Wikipedia pages indicate editors are indeed knowledgeable about complex copyright rules. But, perhaps more importantly for the welfare question, Nagaraj asks: is the Wikipedian's behavior that may be copyright compliant (although arguments exist on both sides) good for their encyclopedic project to produce and disseminate free high-quality and comprehensive information to world readers? These are big and important questions Nagaraj tackles admirably. For anyone interested in a model for robust quantitative experimentation in intellectual property with qualitative implications and analysis for further study, I highly recommend Nagaraj's newest paper.

Cite as: Jessica Silbey, *Three Strikes for Copyright*, JOTWELL (October 13, 2017) (reviewing Abhishek Nagaraj, *Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia*, **Mgmt. Sci.** (forthcoming 2017), available at abhishekn.com), https://ip.jotwell.com/three-strikes-for-copyright/.