Stanford University • December 2008

SIEPR policy brief

Stanford Institute for Economic Policy Research

on the web: http://siepr.stanford.edu

How Many Scanned Books on the Web?

By Paul A. David and Jared Rubin

The 21st Century Dream of a "Universal E-Library" Meets the Realities of the 20th Century's Copyright Legislation

Much fanfare accompanied the announcement in December 2004 that Google, the operator of the world's most popular Internet search service, had concluded an agreement with Stanford, Harvard, the University of Michigan, the New York Public Library, and the Bodleian Library at Oxford to begin converting their holdings into digital files that would be "freely searchable over the Web."[1] Ten months later, the Open Content Alliance (OCA), a rival consortium hosted by Yahoo and initially joined by Microsoft, was formed and announced plans to digitize hundreds of thousands of out-of-copyright works. The purpose of the many institutions that formed the OCA – including the Library of Congress, the National Archive in England, and the University of California, Berkeley – was to make scanned books accessible to any search engine, not only Google.

Although Google announced (on August 12, 2005) that it would "pause" its scanning activities to give publishers and other copyright holders the chance to "opt out," that didn't forestall the filing of a lawsuit (on September 20, 2005) in

continued on page 2 ...

1 Markoff, John, and Edward Wyatt (2004), "Google Is Adding Major Libraries to Its Database," New York Times, December 14, 2004. Available at *http://donswaim.com/nytimes.google.html*. Librarian at Stanford University, Michael A. Keller, was quoted in this story as seeing the future in the initially rosy view that greeted this project: "Within two decades, most of the world's knowledge will be digitized and available, one hopes for free reading on the Internet, just as there is free reading in libraries today."

About The Authors

Paul A. David is Professor of Economics (Emeritus) and SIEPR Senior Fellow at Stanford. He is also Professorial Fellow at the UN University-MERIT (Netherlands) and Emeritus Fellow of All Souls College, Oxford. David has published more than 160 articles in journals and edited books, and several books, including *The Economic Future in Historical Perspective* (Oxford UP). He is known internationally for his historical studies of economic growth and research on "path dependence" in modern economies — particularly their long-term demographic, institutional, and technological dynamics. David's work recently has focused on public policies affecting investment in science and technology, and he serves on an Experts Group advising the EC Commissioner for Research. An elected Fellow of the American Academy of Arts and Sciences, the American Philosophical Society and the British Academy, in 2009-2010 David will be president-elect of the Western Economics Association International.

Jared Rubin is an assistant professor of economics at Cal State Fullerton. He received his PhD in economics at Stanford University in 2007 and held a SIEPR dissertation fellowship during 2006/07. His research concentrates on both the economic effects of religious and political institutions in Islam and Christianity as well as U.S. copyright history.

which three authors – including the Authors Guild, a group representing more than 8,000 published authors – claimed that Google had engaged in "massive copyright infringement." Google then resumed its library scanning program (in November 2005), focusing mainly on the older works that were unambiguously out of copyright and works that were always in the public domain. But within a month the project was again halted, when five major commercial publishers sued Google requesting damages and injunctive relief, asserting that the "massive, wholesale and systematic copying of entire books still protected by copyright" infringes on the publishers' rights.

The settlement of these lawsuits was announced recently (on October 29, 2008): Google agreed to pay book publishers and authors $125 million, and to a 37-63 split of the revenues from the project. As a result, for works whose publishers have not "opted out" of Google Print book-scanning program, the entire text will be available on the Web for a fee, whereas "snippets" not exceeding 20 percent of the text will be displayed without charge. Universities and libraries cooperating will be offered subscriptions to the entire collection of digitized works.[2]

Some commentators lauded this settlement – among them Stanford law professor Lawrence Lessig, who noted that it would allow greater access to copyrighted works in general and particularly to books that currently are out of print. Others, however, expressed concerns that the settlement was placing in Google's hands too much control over access to digitized works that would remain out of the public domain – in some cases for a long time to come. Perhaps for that reason Harvard University decided to not allow Google to scan its holdings of "in-copyright" works.

Most of the world's books that remain "in copyright" – popularly estimated today at about 90 percent of some 32 million ever copyrighted volumes – are close to valueless for commercial purposes. Yet, they have significant public value – especially for present and future generations of researchers and students. Preservation of access to society's cultural, scholarly and scientific heritage is a primary concern of traditional research librarians; indeed, this is a key motivation for maintaining a healthy public domain to which works granted copyright protection eventually will be returned. It also is the core rationale for not unnecessarily deferring the termination date of the period during which access can be restricted by the legal protection afforded to owners of copyrights. If one wonders what is immediately at stake here, it is possible to begin to get a sense of this by thinking about the following question: Supposing that Google and the OCA today were able to instantaneously digitize all library collections of U.S. copyrighted works, just how many of those texts would be available tomorrow for free reading on the Internet? To answer this and similar questions about the quantitative dimensions of the issues, speculation about the future progress of digital information technologies is not relevant. Instead, as has been intimated by our earlier notice of the small proportion of the world's texts that presently are in the public domain, one must look to the past, and in particular to the 20th century history of copyright legislation.

If this point had not occurred to the more effusive public commentators when the Google "libraries project" first was announced, it certainly was quite apparent to Stanford librarian Elizabeth Edwards when she pointed out that it is not technology but copyright

2 See: *http://www.nytimes.com/2008/10/29/technology/internet/29google.html?r=1&scp=3&sq=out%20of%20court%20copyright%20.*

3 Edwards, Elizabeth (2005), "The Google Deal (Down on the Farm)," Confessions of a Mad Librarian – Digital Issues Archive Blog, January 07, 2005. Available at *http://edwards.orcas.net/blog/archives/2005_01.html.*

law that "will drive what can be fully displayed on the Web."[3] The significance of the past in determining what can be "fully displayed" emerges clearly when the question is rephrased this way: In the course of the coming 20 years, how many more books will have ceased to be under the protection of U.S. copyright law and therefore could be made available via the Web for unrestricted browsing, searching, and free downloading?

Copyright law has not remained static, however. Books registered with the U.S. Copyright Office at different dates in the past will have begun their "published lives" under the terms set by different statutes and will have been treated differently by the subsequent modifications of the law. Consequently, the question we must try to answer is more specific than the one with which we began: What has been the magnitude of the effects of successive revisions of the 1909 copyright statute upon the numbers of books and pamphlets that are not scheduled to become fully accessible on the Internet in each year of the coming quarter century?

Pursuit of this line of inquiry has given us some surprising answers and, as a side-benefit, a way of quantifying the first-order effects of each step in the sequence of changes made during the latter half of the 20th century in the U.S. statutes affecting the term of copyright protection. This latter approach exposes important effects of a number of legislative actions that passed essentially without notice or comment when they were being enacted into law, although they were quite foreseeable and hence could have occasioned debate at that time.

The Course of 20th Century U.S. Copyright Legislation

The driving forces behind 20th century copyright legislation have persistently favored lengthening the term of protection. The first act extending duration was the Copyright Act of 1909, which lengthened the renewal term to 28 years (from 14 years), making the maximum renewal period 56 years (an initial 28-year period followed by a 28-year renewal period). The Act of 1909 dictated the duration of copyright up until 1962, but from then until the end of the century Congress modified the copyright term provisions of the law 11 times. This began with the Act of 1962, which kept out of the public domain the copyrights that were in a renewal term that was due to expire on September 19, 1962. That eleventh-hour congressional "reprieve" for copyright owners was repeated in every year until 1976: the effect being to reset the maximum statutory duration at 75 (or 28 plus 47 years) for works registered before 1978. These annual rituals culminated in the Copyright Act of 1976, which officially set the renewal term at 47 years for all works published before January 1, 1978, and specified that for works published thereafter there would be no renewal requirement – the term of copyright protection would be the author's life plus 50 years. In 1992, Congress eliminated the formality of having to seek renewal while the work was in copyright and automatically granted renewal for all works that had been published between 1964 and 1977.

In contrast with the slight public notice that those changes had attracted, during 1998 the news media devoted considerable attention to the ongoing legislative effort to prevent Disney's copyrights on Mickey Mouse, and much else besides, from "falling into the public domain." After the passage of the Sonny Bono Copyright Term Extension Act (CTEA) in 1998, there came the widely reported attempt to have the 1998 statute overturned on constitutional grounds – in the unsuccessful legal suit *(Eldred v. Ashcroft)* that Lawrence Lessig argued before the U.S. Supreme Court. What the CTEA had done was grant works copyrighted by individuals after 1978 a term of protection limited to the author's "life plus 70 years"

continued on page 4 ...

(thus adding 20 years to the pre-existing "life plus 50") and applied this extension *retroactively* to works still in copyright at the time of the act's passage. Some works that had been published well before 1978 thereby were awarded a total term of protection lasting 95 years.

To quantitatively assess the consequences of this 20th century record of U.S. copyright legislation, we carried out a series of calculations that yielded estimates of the numbers of books kept out of the public domain for additional periods of time by these statutes, separately and collectively. The next section presents highlights of the results and briefly explains the way they were obtained (details of both can be found in the "Further Reading" cited at the end of this brief). Following that we briefly consider some broader implications of these findings for future approaches to legislative "policymaking" affecting protection of intellectual property rights, and copyright protection in particular.

Without the Legislative Extensions of Copyright, How Many More Books Would Have Been in the Public Domain?

We began this exercise by determining the number of books registered after 1902 that are in the public domain and then asked this question: How many books would have been in the public domain if each of the successive changes in the term of copyright acts had not been enacted? That is, we calculate a series of counterfactual outcomes in which each legislative act (in the sequence) did not exist, providing us with an estimate of each law's marginal or incremental effect on the number of books in the public domain. Copyright term extensions may have had another, non-quantified effect: Increased potential rents to copyright owners might have induced more authors to write more books. In a related study, however, we have econometrically modeled the determinants of U.S. copyright registrations and renewals during this era (to 1997) and find that the length of the renewal period has no statistically significant effect on the number of books copyrighted in any given year. So, for purposes of the present empirical exercise we take it as reasonable to set aside possible positive marginal effects of these extensions and to concentrate on the legislation's restrictive effects.

This estimation approach may be illustrated by starting with calculations to assess the immediate marginal impact of the Sonny Bono Act. Had that legislation not been passed, copyrights beyond their 75th year of age would have fallen into the public domain between 1998 and the present; the counterfactual therefore includes all those works among the total number of books in the public domain.

From such counterfactual calculations made also for the acts of 1992, 1976 and 1962 (and its annual *sequelae*) we found that the 1976 and 1992 acts have had a surprisingly large negative impact on the number of books available in the public domain. This predictable result scarcely drew mention in contemporary discussions of either of the latter two bills, nor subsequently when interest in the broader issues surrounding term extensions was stirred by the litigation over the 1998 Sonny Bono Act. By comparison, the incremental effects of the acts of 1998 and 1962 (et seq.) remain quite negligible for the period stretching as far into the future as 2027. We estimate that combined effects of the copyright acts of 1976 and 1992 alone will have kept some 1.5 million volumes out of the public domain by 2010; by 2027 the cumulative number of copyrighted titles that returned to the public domain (absent just those two acts) would have been twice as large, approximately 3 million – more than double the number that can be anticipated under the presently existing statutes. In other words, for all books registered since 1902, the estimated number of books that will not have come into the public domain by 2027 as a result of the 1976 and 1992 acts will be approximately 15 percent greater than the cumulative number that will become

available for free downloading from the Web by that date.

The effects just described flowed from the removal of the copyright renewal option for books registered after 1964. Between the 1950s and the mid-1970s, the average probability of a book's copyright actually being renewed when its initial term expired was rising from approximately .05 to the neighborhood of .17. The effect of the congressional legislation therefore was to impose longer periods of protection upon the portion of the expiring copyrighted titles (declining secularly from c. 0.95 towards c. 0.80) that otherwise could have been expected to fall into the public domain. Consequently, the estimates cited above are approximations that rest on the assumption that (in the absence of the new statutes' automatic renewal provisions) copyright owners would have continued to adhere to the average rates at which they had been allowing protection to lapse after the initial term set by prevailing law.

How Many Books Have Been (Unambiguously) Kept Out of the Public Domain by Legislative Extension of Copyright, and For How Long?

The fact that many books registered between 1923 and 1964 are in the public domain because their copyrights were not renewed is not immediately apparent for individual works, and the transaction costs associated with searching for and determining a book's current copyright status (and the identity of the current owner when protection remains in force) can be prohibitively expensive. This consideration has obliged the organizations involved with the OCA to play it safe by restricting scanning to books that unambiguously are out of copyright. Indeed, even though at present most of the books published in the 1923-1964 interval were not renewed and are no longer in copyright, the high costs of determining their status in each case appears to have been sufficient also to have dissuaded Google from showing more than a snippet under the terms of the recently settled infringement suits.

This situation makes it of interest to determine the numbers of books that as a result of 20th century U.S. copyright legislation did not *clearly* enter the public domain in each year of this period. The answers here will be different from those presented previously, because, having altered the question, it is no longer necessary to include in the total an estimate of the number of titles that were actually renewed; we can simply add the number of books that had become eligible for renewal in each year to the total number of books whose original term of protection was still in force, or had been extended by changes in the copyright statutes. The main actors responsible for reducing the stock of works unambiguously in the public domain are the 1962 Act (and its sequels) and the 1998 CTEA, both of which lengthened the maximum possible duration of protection by extending the renewal period of works still in copyright. Figure 1 (page 6) displays our estimates of the incremental effects of those statutes for each year between 1957 and 2007, showing the effect of the acts of 1962 and its sequels in the divergence between the solid and dashed lines during the 1963-77 interval, with the second major divergence occurring as the CTEA took effect.

One can turn to Figure 2 (page 6) for a summary view of the total numbers of works that have enjoyed various periods of legislatively *extended* copyright protection (19, 39, or 67 years, or still longer durations) before eventually returning to the public domain. The gray bars in this chart indicate the number of books whose entry into the public domain was unambiguously delayed for the indicated period of years (and the legislative acts responsible). Those constitute the *lower-bound* estimates of the works that – having been published in a particular time-interval – actually were held back for a certain time by the effects of either an individual act or interaction between more than

continued on page 6 ...

FIGURE 1

Cumulative Books Entering the Public Domain After 1958 Incrementing for Effect of Each Law, Books Registered 1902–Present

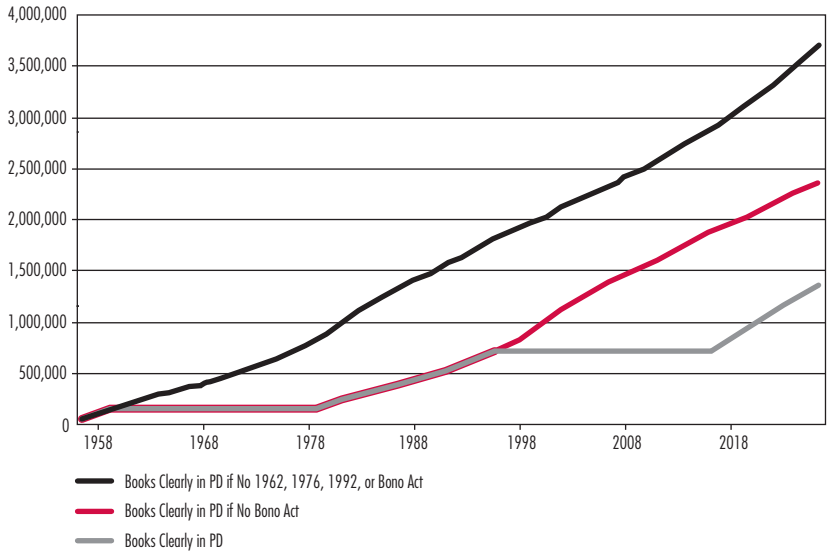
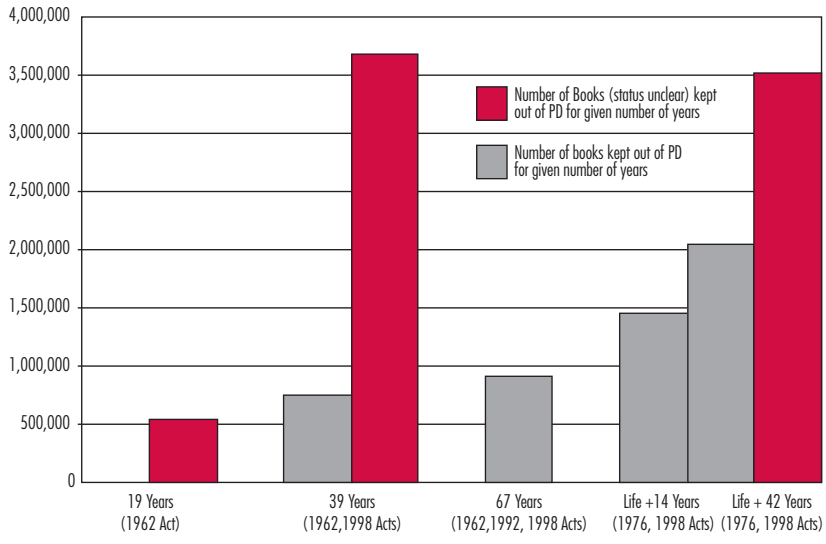


FIGURE 2

Total Number of Books Registered During 1902–1999 Whose Copyright Lives Were Extended by Legislation, Grouped by Additional Number of Years Held Out of the Public Domain Under the Indicated Statutes



one statutory change. The red (and tallest) bars show the corresponding *upper-bound* estimates – on the supposition that full renewal rights were exercised by the copyright holders in every eligible case. The differences between light and dark for each of the duration-groups reveal how large is the volume of works whose status remains ambiguous, requiring costly inquiries to ascertain whether or not they had been renewed.

Looking at the gray bars alone, and supposing (conservatively) that the representative book's author had on average 35 years of life remaining after its publication, one can conclude that by repeatedly extending the term of copyright these five legislative acts have combined to produce *at least 300 million book-years* of delayed public domain access.

Some Broader Implications of the Quantitative Findings

There is a nice anecdote that strikingly illuminates the potential burdens that consumers of "creative works" have incurred as a result of extensions of copyright protection, by showing what happened in one instance where some of those costs were miraculously avoided – due to an oversight on the part of the copyright owner. The copyright on Frank Capra's 1946 film *It's a Wonderful Life*

was not renewed upon the expiration of its initial 28-year term, quite inexplicably, since in 1974 the cost of the renewal registration was negligibly small. The film had been largely ignored when it was first released and was barely remembered, except by Jimmy Stewart's most dedicated fans. Only during the latter half of the 1970s, following its "accidental fall" into the public domain, did it ascend rapidly to its present perennial place in popular television-programming for the Christmas holiday season. This story was brought up during a 1991 congressional hearing, but not to indicate how much the viewing public might benefit by abridgments of the duration of copyright protection. Quite the opposite, it was cited as a reason for making renewal automatic rather than discretionary – so that such "tragic" losses of profits by inattentive copyright owners would ever after be prevented!

Gaining a more precise understanding of the impacts of intellectual property rights upon the advancement of and access to the diverse forms of knowledge shared by human cultures is one of the larger goals toward which the research reported here has been directed. Aside from the intrinsic interest of that large and complicated question, it is important to deepen our understanding of the variety of situations in which this issue cannot be ignored. The statutory copyright regime, in particular, almost certainly will continue to undergo modifications (as will other institutional structures that impinge upon the production of cultural and scientific information and the benefits that flow from their distribution); at the same time, the law's present and likely future configuration will continue to shape the pace and direction of tomorrow's technological innovations.

That reciprocal dynamic, however, is not guaranteed to unfold in ways that will enable society to benefit most fully from the technical capacities afforded by enhanced telecommunication network infrastructures and networked digital information applications.

Without gaining a better understanding of these complicated interactions, and hence a greater ability to anticipate and assess how the process is likely to proceed when driven largely by the pursuit of conflicting private interests, it will remain very difficult to discern when and how best to direct attention to the broad range of longer-term societal interests that are at stake. To continue with these blinkered conditions for policymaking would seem to be a recipe for ultimately frustrating not only hopes of the kind that animate the laudable quest for a freely accessible universal library but also the variety of more practical collaborative enterprises undertaken between the keepers of traditional libraries and the providers of digital scanning facilities and search services.

Further Reading

This "brief" is based on research supported by a Rockefeller Foundation grant to Stanford University for SIEPR's Knowledge, Networks and Institutions for Innovation Program (KNIIP) and reported in Paul A. David and Jared Rubin, "Restricting Access to Books on the Internet: Some Unanticipated Effects of U.S. Copyright Legislation", SIEPR Policy Paper No. 07-036 (March 2008), at *http://siepr.stanford.edu/papers/pdf/07-36.html*. The author's journal article under the same title has been published in the *Review of Economic Research on Copyright Issues*, 5(1), July 2008: pp. 23-53.

SIEPR

About SIEPR

The Stanford Institute for Economic Policy Research (SIEPR) conducts research on important economic policy issues facing the United States and other countries. SIEPR's goal is to inform policymakers and to influence their decisions with long-term policy solutions.

Policy Briefs

SIEPR Policy Briefs are meant to inform and summarize important research by SIEPR faculty. Selecting a different economic topic each month, SIEPR will bring you up-to-date information and analysis on the issues involved.

SIEPR Policy Briefs reflect the views of the author. SIEPR is a non-partisan institute and does not take a stand on any issue.

For Additional Copies

Please see SIEPR website at *http://SIEPR.stanford.edu.*

SIEPR *policy brief*

A publication of the
Stanford Institute for Economic Policy Research
Stanford University
579 Serra Mall at Galvez Street
Stanford, CA 94305
MC 6015

Non-Profit Org.
U.S. Postage
PAID
Palo Alto, CA
Permit No. 28