Copyright and Research in Google Book Search

by Benjamin J. Keele

Benjamin Keele is a reference librarian at the William and Mary Law Library, where he provides research assistance and teaches legal research. He earned his law degree from the Indiana University Maurer School of Law and his master of library science degree from the Indiana University School of Library and Information Science. He and is co-author of the forthcoming *Librarian's Copyright Companion, Second Edition*. His research interests are copyright, privacy law, and scholarly communications.

Many researchers — even trained professionals — often use the Google search engine to begin searches for information. Google's many products enable researchers to search public websites, scholarly articles, and even patents. One vast area of information not yet thoroughly indexed by Google is print books. Google Book Search (also at times referred to as Google Books, Google Print and Google Library Project) is the company's effort to digitize and index the world's print literature.

Google's digital corpus is remarkable and valuable, but due to copyright law Google cannot simply give or sell access to all these books. This article will briefly review the history of Google Books, including the litigation and proposed settlement. Then the article will discuss how copyright affects the scope and functionality of Google Books.

Google Books Litigation

In December 2004, Google announced plans to scan and index the text of millions of print books. Books were borrowed from major research libraries and scanned. Publishers could also make agreements with Google to include digital copies of their books. The full text was made available of books in the public domain due to expired copyrights. Only snippets — a few sentences relevant to the search terms — were made available from books that possibly were still under copyright protection. A publisher can consent to Google providing greater access to its books.[1]

In September and October 2005, authors and publishers of books digitized by Google sued the company for infringement of their copyright privileges. Google asserted that digitizing books to make them searchable, show snippets, and display relevant ads was not infringement because it was fair use.[2] After years of negotiating, in October 2008 the parties announced a proposed settlement of the class-action lawsuit.[3]

Under the first proposed settlement, Google would have a license to digitize and display up to 20 percent of virtually every book in the world. Full text of books would be accessible through individual purchase or institutional access. Libraries would receive public access terminals that would provide free access to the digital books, although there would be a fee for printing from the database. Libraries could also purchase subscriptions to provide access beyond the dedicated terminals. Revenue would be shared between Google and a new Books Rights Registry. The registry would be a royalties-collecting nonprofit organization charged with distributing royalties to publishers and authors and seeking out the rights holders of orphan works — books for which the rights holders are unknown.

Google would also be permitted to make further uses of the digital books, in addition to selling digital copies and online subscriptions. As it does for indexed web content and email in its Gmail service, Google would display relevant ads next to the book content. The digital corpus would also let Google further refine its search algorithms, making it more competitive in the search market. Finally, Google would provide a copy of the digital corpus to two universities that would provide computational access to nonprofit researchers. This text-mining research would be valuable to, among others, linguists, historians, and computer scientists.[4]

Rights holders who do not want their books to appear in Google Book Search could direct Google to remove them. The system is opt-out. Unless Google hears otherwise, it would be able to digitize books, include them in the digital collection and display 20 percent of each book's contents. This default rule is especially important because it gives Google access to orphan works for which it would be difficult to locate every rights holder and negotiate for permission.

The sweeping proposed settlement drew a number of objections, most notably from the

Department of Justice, which expressed concerns about antitrust. Since the settlement would give Google retroactive cover for digitizing millions of books, possible competitors would have to mimic Google by scanning books and hoping to get a similar settlement — a significant liability risk. Google is the only company that is digitizing books on such a large scale. Microsoft scanned books for a time, but later bowed out. Having the settlement apply only to Google would present a major barrier to any possible competitors.

After the Department of Justice recommended against approval of the first settlement, the parties produced an amended settlement to address some of the concerns. While the original settlement would have included books from any country, the amended settlement includes books from only the United States, Canada, Australia, and the United Kingdom. A fiduciary to represent the interests of rights holders for unclaimed books would also be added, and the pricing models under which Google could monetize the digital books were further defined.

These amendments did not comfort the Department of Justice and class members, who objected. Objectors claimed that the settlement gave Google an end run around copyright law by permitting it to reproduce books and distribute copies without express permission of the rights holders — especially those who owned the rights to orphan works. The settlement provisions addressing orphan works would impose a solution where Congress would normally act, but had thus far failed to. There was concern that Google would still gain a market position so dominant that competition would be impossible and that libraries and researchers would become dependent on Google for access to digital book content.[5] Due to the objections, in March 2011, Judge Denny Chin of the U.S. Court of Appeals for the Second Circuit rejected the amended settlement. Chin suggested that an opt-in system, rather than an opt-out one, would be more acceptable.

Chin's decision places the parties in a difficult position, because the opt-out system made the settlement attractive to both sides. Google would incur significant transaction costs analyzing each book to determine if it is still under copyright protection and who holds the copyright. For books that may be more than seventy years old, this can be a challenge. Paperwork has been lost, publishers have dissolved or merged with others, and authors have died. There are probably many successor companies and heirs that technically own copyrights but have no knowledge of them. For some books, the trail has gone so cold there may be no way to definitively say who holds the copyrights. Even if after a diligent search a relatively small percentage of books are truly orphans, in a universe of millions of books a fraction is still a large number. The settlement would have let Google commercialize these orphan books without identifying and negotiating with rights holders. The Book Rights Registry, funded through revenues shared by Google, would handle that chore.

The publisher and author parties likewise would rather avoid the expense of tracking down rights holders, and including the orphan works would increase the value of the institutional databases from which they would derive some revenue. Money earned from the unclaimed works would be entrusted to an unclaimed works fiduciary that would hold the funds in escrow for ten years, after which the money could be donated to charities.[6] If the works were claimed in that time, then the money can be paid to the rights holders.

If a new settlement were opt-in, then unclaimed works would not be available for Google to index or provide through institutional subscriptions. The first amended settlement had already excluded most foreign works, so further limiting the database to books in the public domain or permitted by their rights holders would make the database much less valuable than it would have been under the original settlement agreement.

The parties will now have to decide whether such a settlement is worthwhile. Other options are dropping the litigation or proceeding to trial. Either way, without orphan works legislation from Congress, it appears that a monolithic collection of digital books is unlikely to be available soon. In the meantime, Google has proceeded with scanning books and offering snippet views, as it has from the beginning.

Research in Google Books

Even though copyright law has placed limits on what Google can do with its digital copies of books, Google Books is still a valuable research tool. Google has two groups of sources for its digital books: libraries and publishers. Books from libraries are scanned, while books from publishers might be scanned or supplied as digital files. The full text of each digitized book is indexed and searchable, but for some books Google only has bibliographic information, such as title, author, and publisher. Searching in Google Books is simi-

lar to searching in Google's standard Web search engine. Researchers type in search terms and see what Google's algorithms find, or use Boolean operators to further limit the search. An advanced book search interface lets researchers search for books by title, author, publisher, or even International Standard Book Number. One can browse books by subject, although the subjects are general enough that using them alone would be unwieldy. For example, the subject "law" has about forty-five million results. Using that subject heading and adding the search terms "copyright" and "fair use" reduces results to seventeen thousand hits, of which at least the first few dozen are relevant books. For researchers already familiar with using Google to search the Web, searching in Google Books will require relatively little practice.

Once one finds relevant books, there are three levels of view. For books that are not protected by copyright or for which Google has rights holder permission, the full text is accessible. Public domain books can be downloaded as PDFs. Rights holders can also opt for a limited view, in which certain pages or chapters, but not the entire book, are viewable for free. For books that are copyrighted and for which Google has no permissions, a snippet view shows a few sentences that contain the search terms. One cannot print pages from the Google Books interface.

For most books, Google links to book vendors from which paper copies can be purchased and a link to the book's WorldCat record so a copy in a library can be located. Public domain and rights-holder-authorized books can also be added to a Google eBooks library. Through this service one can purchase ebooks that are stored on Google's servers and accessible through a computer and other ebook reader devices.

Google maintains that its scanning and indexing efforts are fair use, but it does not contend that it may make the full text of a copyrighted book freely available without rights holder permission. Google Books provides full-text access to books that are very old (and thus have expired copyrights) or relatively new (and have active rights holders to grant permission). This makes Google Books a great source for old books that may be difficult to borrow from a library. Books published between 1923 (the year before which virtually all copyrights have expired) and the last few decades are less likely to have more than snippet view. For those volumes Google Books will be mostly useful as a tool to find bibliographic information and links to libraries and booksellers with a paper copy. Until the orphan works problem is legislatively addressed, it will be difficult for Google to expand its full-text coverage of books from this period.

Another limit on Google Books is its spotty quality control. Partly due to the large scale of its operations, some scans are fuzzy or skewed, and major metadata issues have been found. Some books have been marked with incorrect publication dates or subject headings, for example.[7] Google automatically runs optical character recognition software to make the scans searchable, but the software makes mistakes and humans do not review the transcriptions for errors. Google Books generally works well, but researchers should not assume that its contents are thoroughly edited and they should be prepared to do more research if an anomalous result is found.

Google Books is an ambitious project to make print books accessible online. Copyright concerns have slowed the project's progress. Depending on how one views copyright, the law has prevented universal access to much of the world's printed literature, or preserved authors' and publishers' proper rights to control and benefit from their creative works. Either way, Google has given researchers, authors, publishers, and Congress reason to reconsider how copyright can fulfill its constitutional purpose to "promote the progress of science and the useful arts."[viii] In the meantime, Google Books is a relatively intuitive and familiar way to search and access books. Its contents are not perfect, but researchers should think of Google Books when looking for books on legal and nonlegal topics.

For Further Information

- Official Google Book Search website: http://www.google.com/googlebooks/about.html
- Official Google Book Settlement website: http://www.googlebooksettlement.com/
- American Library Association website on Google Books: http://wo.ala.org/gbs/
- The Public Index, a project at New York Law School to monitor the Google Book Settlement: http://thepublicindex.org/

Endnotes:

1 Kate M. Manuel, *The Google Library Project: Is Digitization for Purposes of Online Indexing Fair Use Under Copyright Law?*, CRS Rep. R40194, at 1 (2009).

2 For analyses of the merits of Google's fair use claim, see Hannibal Travis, *Google Book Search and Fair Use: iTunes for Authors, or Napster for Books?*, 61 U. MIAMI L. REV. 601 (2006), and Melanie Costantino, *Fairly Used: Why Google's Book Project Should Prevail under the Fair Use Defense*, 17 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 235 (2006).

3 Much of this description of the settlement agreements is drawn from Pamela Samuelson, *Google Book Search and the Future of Books in Cyberspace*, 94 MINN. L. REV. 1308 (2010).

4 Google has already launched its Books Ngram Viewer, which lets one search a set of digital books for the occurrence of words and phrases over time. More information can be found at http://ngrams.googlelabs.com/info and in Jean-Baptiste Michel, et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, 331 SCIENCE 176 (2011).

5 More detailed reviews of objections to the settlement can be found in Jonathan Band, *Guide for the Perplexed Part IV: The Rejection of the Google Books Settlement*, http://www.arl.org/bm~doc/guideiv-final-1.pdf (Apr. 1, 2011), and Pamela Samuelson, *Academic Author Objections to the Google Book Search Settlement*, 8 J. ON TELECOMM. & HIGH TECH. L. 491 (2010).

6 Pamela Samuelson, *The Google Book Settlement as Copyright Reform*, 2011 WIS. L. REV. 480, 524.

7 Geoffrey Nunberg, *Google's Book Search: A Disaster for Scholars*, CHRON. HIGHER EDUC., Aug. 31, 2009, *available at* http://chronicle.com/article/Googles-Book-Search-A/48245/; Norman Oder, *Google, 'The Last Library,'; and Millions of Metadata Mistakes*, LIB. J., Sep. 3, 2009, *available at* http://www.libraryjournal.com/article/CA6687562.html.

8 U.S. CONST. art. I, sec. 8, cl. 8.