THE NEW YORKER

ANNALS OF LAW

Google's Moon Shot

Jeffrey Toobin

February 5, 2007

The quest for the universal library.

Every weekday, a truck pulls up to the Cecil H. Green Library, on the campus of Stanford University, and collects at least a thousand books, which are taken to an undisclosed location and scanned, page by page, into an enormous database being created by Google. The company is also retrieving books from libraries at several other leading universities, including Harvard and Oxford, as well as the New York Public Library. At the University of Michigan, Google's original partner in Google Book Search, tens of thousands of books are processed each week on the company's custom-made scanning equipment.

Google intends to scan every book ever published, and to make the full texts searchable, in the same way that Web sites can be searched on the company's engine at google.com. At the books site, which is up and running in a beta (or testing) version, at books.google.com, you can enter a word or phrase—say, Ahab and whale—and the search returns a list of works in which the terms appear, in this case nearly eight hundred titles, including numerous editions of Herman Melville's novel. Clicking on "Moby-Dick, or The Whale" calls up Chapter 28, in which Ahab is introduced. You can scroll through the chapter, search for other terms that appear in the book, and compare it with other editions. Google won't say how many books are in its database, but the site's value as a research tool is apparent; on it you can find a history of Urdu newspapers, an 1892 edition of Jane Austen's letters, several guides to writing haiku, and a Harvard alumni directory from 1919.

No one really knows how many books there are. The most volumes listed in any catalogue is thirty-two million, the number in WorldCat, a database of titles from more than twenty-five thousand libraries around the world. Google aims to scan at least that many. "We think that we can do it all inside of ten years," Marissa Mayer, a vice-president at Google who is in charge of the books project, said recently, at the company's headquarters, in Mountain View, California. "It's mind-boggling to me, how close it is. I think of Google Books as our moon shot."

Google's is not the only book-scanning venture. Amazon has digitized hundreds of thousands of the books it sells, and allows users to search the texts; Carnegie Mellon is hosting a project called the Universal Library, which so far has scanned nearly a million and a half books; the Open Content Alliance, a consortium that includes Microsoft, Yahoo, and several major libraries, is also scanning thousands of books; and there are many smaller projects in various stages of development. Still, only Google has embarked on a project of a scale commensurate with its corporate philosophy: "to organize the world's information and make it universally accessible and useful."

In part because of that ambition, Google's endeavor is encountering opposition. A federal court in New York is considering two challenges to the project, one brought by several writers and the Authors Guild, the other by a group of publishers, who are also, curiously, partners in Google Book Search. Both sets of plaintiffs claim that the library component of the project violates copyright law. Like most federal lawsuits, these cases appear likely to be settled before they go to trial, and the terms of any such deal will shape the future of digital books. Google, in an effort to put the lawsuits behind it, may agree to pay the plaintiffs more than a court would require; but, by doing so, the company would discourage potential competitors. To put it another way, being taken to court and charged with copyright infringement on a large scale might be the best thing that ever happens to Google's foray into the printed word.

Though Google has more than ten thousand employees—about fifty new ones are hired each week—and a market capitalization of more than a hundred and fifty billion dollars, the company cultivates the air of a college campus at its headquarters, in Silicon Valley. Now and then, there are self-consciously wacky stunts, like Pajama Day, which happened to take place when I visited. (The event

was to be madcap within reason; supervisors were told to convey the message that "pajamas means 'pajamas,' not 'what you sleep in.' ") When I met with Sergey Brin, a co-founder of Google, he was wearing bright-blue p.j.s, with the company's logo stitched on the breast pocket.

The story of how Brin and Google's other co-founder, Larry Page, met as graduate students in computer science at Stanford in the mid-nineties, and devised a series of elegant software algorithms that allowed Web searchers to find relevant information quickly and efficiently, has become part of Silicon Valley lore. Less well known is that, at the time, Brin and Page were also working on Stanford's Digital Library Technologies Project, an attempt, funded by the federal government, to organize different kinds of stored information, including books, articles, and journals, in digital form. "There was an attitude in computer science that putting things on dead trees was obsolete and getting it all into a searchable, digital format was a quest that had to be accomplished someday," Terry Winograd, a Stanford professor who was a mentor to Page and Brin, said.

After founding Google, in 1998, Page and Brin—who are now in their mid-thirties and worth around fourteen billion dollars each—began to talk about how to include books in the company's database. Page, in particular, embraced the idea of putting books online; at one point, he set up a primitive lab in his office, with a scanner and a page-turning machine. "I think it was motivating to have those kinds of aspirations, but nobody really took it seriously," Brin told me. The men were less interested in making it easy for people to obtain the full texts of books online than in making accessible the information those books contained. "We really care about the comprehensiveness of a search," Brin said. "And comprehensiveness isn't just about, you know, total number of words or bytes, or whatnot. But it's about having the really high-quality information. You have thousands of years of human knowledge, and probably the highest-quality knowledge is captured in books. So not having that—it's just too big an omission." As Marissa Mayer put it, "Google has become known for providing access to all of the world's knowledge, and if we provide access to books we are going to get much higher-quality and much more reliable information. We are moving up the food chain."

In 2002, Google quietly made overtures to several libraries at major universities. The company proposed to digitize the entire collection free of charge, and give the library an electronic copy of each of its books. "Larry is an undergrad alum here at Michigan, and he knew we were already interested in digitizing the library as part of our preservation efforts," John Wilkin, an associate university librarian at Michigan, told me. "There was a lot of back-and-forth between Google and us in the process. We wanted to insure that the materials wouldn't be damaged and that what came out could be used as a preservation surrogate. They started experimenting with different ways of copying the images, and we started a pilot project in July, 2004. We've been getting better, going faster. We're doubling our output all the time." The Michigan library holds seven million volumes, and Wilkin believes that Google will have copied the entire collection in about six years.

Last month, at the New York Public Library, Google hosted a conference on the future of the publishing industry. About four hundred people—mainly publishing executives and agents—attended, most of them grimly aware of the simultaneous lethargy and panic that have characterized their industry's response to the digital age. Nearly all attempts to sell books in an electronic format have been disappointing, and now Google appeared to be encroaching on the publishers' domain. The implicit message of the conference was summed up by a quotation from Charles Darwin that was projected on a screen: "It is not the strongest of the species that survive, nor the most intelligent, but the ones most responsive to change." As Laurence Kirschbaum, a longtime publishing executive who recently became a literary agent, told me at the conference, "Google is now the gatekeeper. They are reaching an audience that we as publishers and authors are not reaching. It makes perfect sense to use the specificity of a search engine as a tool for selling books."

Google thought so, too, and designed the books project accordingly. In addition to forming partnerships with libraries, the company has signed contracts with nearly every major American publisher. When one of these publishers' books is called up in response to search queries, Google displays a portion of the total work and shows links to the publisher's Web site and online shops like Amazon, where users can buy the book. "We are helping the publishers reach consumers that otherwise might not have known about their books and helping them market their books by giving limited but relevant previews of the books," Jim Gerber, Google's director of content partnerships, told me. "The Internet and search are custom made for marketing books. When there are a hundred and seventy-five thousand new books each year, you can't market each one of those books in mass market. When someone goes into a search engine to learn more about a topic, that is a perfect time to make them aware that a given book exists. Publishers know that 'browse leads to buy.' " (Google says that it does not take a cut of sales made through its books site.)

Still, on October 19, 2005, several leading publishers, including Simon & Schuster, the Penguin Group, and McGraw Hill—all of which are partners in Google

Book Search—filed a lawsuit against the company, seeking to stop the project. The publishers don't object to Google's plan for helping them sell new books, but they assert that the library component of the project is illegal. They claim that Google's "massive, wholesale and systematic copying of entire books still protected by copyright" infringes on the publishers' rights. They demand that Google stop further copying and "destroy all unauthorized copies made by Google through the Google Library Project of any copyrighted works." (The Authors Guild filed its lawsuit around the same time.) The publishers, who have the support of the Association of American Publishers, are suffering from a version of the problem that John Kerry had in the last Presidential campaign: they are for Google Book Search at the same time that they are against it.

Copyright law dates to the birth of the Republic. Article I of the Constitution assigns Congress the right to pass laws "securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." The first copyright law was passed in 1790, and it has been frequently and confusingly amended over the years, most recently in the Sonny Bono Copyright Term Extension Act of 1998, which extended copyright terms by twenty years. (The law is also known as the Mickey Mouse Protection Act, because the Walt Disney Company, seeking to protect its copyright on early animated classics like "Steamboat Willie," lobbied heavily for it.) The twisted history of copyright law has insured an awkward passage into the digital age.

The legal assertion at the core of Google's business plan is its purported right to scan millions of copyrighted books without payment to or permission from the copyright owners. Approximately twenty per cent of all books are in the public domain; these include books that were never copyrighted, like government publications, and works whose copyrights have expired, like "Moby-Dick." Google has simply copied such books and made them available on the Web. Roughly ten per cent of books are copyrighted and in print—that is, actively being sold by publishers. Many of these books are covered by Google's arrangement with its publisher partners, which allows the company to scan and display parts of the works.

The vast majority of books belong to a third category: still protected by copyright, or of uncertain status, and out of print. These books are at the center of the conflict between Google and the publishers. Google is scanning these books in full but making only "snippets" (the company's term) available on the Web. (Google searches turn up only the search term and about twenty words on either side of it.) Copyright law has never forbidden all "copying" of a protected work; scholars and journalists have long been allowed to quote portions of copyrighted material under the doctrine of fair use. Google maintains that the chunks of copyrighted material that it makes available on its books site are legal under fair use. "We really analogized book search to Web search, and we rely on fair use every day on Web search," David C. Drummond, a senior vice-president at Google who is overseeing the response to the lawsuits, told me. "Web sites that we crawl are copyrighted. People expect their Web sites to be found, and Google searches find them. So, by scanning books, we give books the chance to be found, too." (Google also has an "opt out" policy, which allows copyright holders to request that specific titles be omitted from the company's database.)

However, according to the plaintiffs in the cases against Google, the act of copying the complete text amounts to an infringement, even if only portions are made available to users. "What they are doing, of course, is scanning literally millions of copyrighted books without permission," Paul Aiken, the executive director of the Authors Guild, said. "Google is doing something that is likely to be very profitable for them, and they should pay for it. It's not enough to say that it will help the sales of some books. If you make a movie of a book, that may spur sales, but that doesn't mean you don't license the books. Google should pay. We should be finding ways to increase the value of the stuff on the Internet, but Google is saying the value of the right to put books up there is zero."

Google asserts that its use of the copyrighted books is "transformative," that its database turns a book into essentially a new product. "A key part of the line between what's fair use and what's not is transformation," Drummond said. "Yes, we're making a copy when we digitize. But surely the ability to find something because a term appears in a book is not the same thing as reading the book. That's why Google Books is a different product from the book itself." In other words, Google says that being able to search books on its site—which it describes as the equivalent of a giant library card catalogue—is not the same as making the books themselves available. But the publishers cite another factor in fair-use analysis: the amount of the copyrighted work that is used in the creation of the new one. Google is copying entire books, which doesn't sound "fair" to the plaintiff publishers and authors. "Traditional copyright analysis says that a transformation leads to the creation of a new and independent work, like a parody or a work of criticism," Jane Ginsburg, a professor at Columbia Law School, said. "Copying the entire work, which is what Google is doing, does not preclude a finding of fair use, but it does fall outside the traditional paradigm."

Harvard, Stanford, and Oxford have prohibited Google from scanning copyrighted works in their collections, limiting the company to books that are in the

public domain. Because of the opacity of copyright law, and the extension of protections mandated by the 1998 act, it's not always clear which works are still protected. (Copyright status can become murky when authors die or publishing houses go out of business.) Stanford has drawn a line at 1964 and prohibited Google from copying most works published since that date. "When Google got sued, we got nervous," Michael A. Keller, the university librarian at Stanford, told me. "We're not a public institution. We don't have any state immunity from being sued ourselves, so we started sorting out the stuff that we know is public domain." (Several of the public institutions that are Google's partners, including the Universities of Michigan, California, Virginia, and Texas at Austin, are allowing the scanning of copyrighted material.)

The chief engineer of Google's system for scanning books in the library collections is Dan Clancy, who joined the company after eight years at NASA, where he supervised teams of Ph.D.s. working on problems related to artificial intelligence. Google provides its employees with free food twenty-four hours a day, and Clancy, a tall, shambling man with a shock of white-blond hair, conducted most of our conversations with bits of granola bar clinging to his shirt.

"Previously, when people have done scanning, they always were constrained by their budget and their scale," Clancy told me. "They had to spend all this time figuring out which were the perfect ten thousand books, so they spent as much time in selection as in scanning. All the technology out there developed solutions for what I'll call low-rate scanning. There was no need for a company to build a machine that could scan thirty million books. Doing this project just using commercial, off-the-shelf technology was not feasible. So we had to build it ourselves."

Google will not discuss its proprietary scanning technology, but, rather than investing in page-turning equipment, the company employs people to operate the machines, I was told by someone familiar with the process. "Automatic page-turners are optimized for a normal book, but there is no such thing as a normal book," Clancy said. "There is a great deal of variability over books in a library, in terms of size or dust or brittle pages." (To needle Google, several blogs have posted images from the books site that include the scanners' fingers.) Google will not reveal how much it is spending on the books project. In 2005, Microsoft announced that it would spend two and a half million dollars to scan a hundred thousand out-of-copyright books in the collection of the British Library. At this rate, scanning thirty-two million books—the number in WorldCat's database—would cost Google eight hundred million dollars, a major but hardly extravagant expenditure for a multibillion-dollar corporation.

Copying all those pages presents many difficulties, but writing software to make the books useful to searchers is even harder. "The scanning technology is boring," Clancy said. "The real challenge is to get somebody something that they are actually interested in, inside a book. Web sites are part of a network, and that's a significant part of how we rank sites in our search—how much other sites refer to the others." But, he added, "Books are not part of a network. There is a huge research challenge, to understand the relationship between books."

Still, the basic search protocols function well. A search for "Heart of Darkness" leads immediately to Joseph Conrad's novel, which is not as obvious as it sounds, considering how common the words in the title are. As Clancy said, "If you put in 'Heart of Darkness,' we have to know that you're looking for the novel, not a book about lighting conditions in cardiac surgery. So how do we do that? We rank some words more important than others. The title may matter more than the content, so we may weight that more. You could also look at what other people have searched for, so if everyone who searched for 'Heart of Darkness' clicked on the novel, we might figure that you probably will, too."

The most important data for ranking searches, Clancy explained, may come from Web pages that link to books in Google's database. (For instance, if links on the phrase "Clinton's autobiography" direct users to a copy of "My Life" on the books site, there is a high probability that people who use the same search terms will also want this result.) "We just started, and we need to make these books networked, and we need people to help us do that," Clancy said.

Google's database contains many books in languages other than English, but for now they must be searched in the original tongue. On the company's Web site, there is already a primitive translation feature, and it may someday be enhanced to allow books to be rendered in another language at the touch of a button. "In terms of democratization, you want to be able to access information," Clancy told me. In places like the Arab world, where few titles are translated into the local languages each year, he said, access to the world's books could have a substantial impact. "We are talking about a universal digital library," Clancy went on. "I hope this world evolves so that there exists a time where somebody sitting at a terminal can access all the world's information."

Such messianism cannot obscure the central truth about Google Book Search: it is a business. Google has pledged not to show advertising next to the pages of library books, but the company does sell advertising alongside

search results that lead to books obtained from publishers. Google's prospects for producing revenue from the books project appear rather modest, but the company has often made a profit on ventures that initially seemed unlikely to be lucrative. "We've had this fortunate streak that when we've done things that have impacted our users and society as a whole—positively, in a significant way—we've been rewarded by that downstream in some way, even though we may not have envisioned exactly what it was right offhand," Sergey Brin told me. "We didn't have ads when we first put up Web search. It wasn't clear it was great business when we started search. In fact, the companies that were doing search were moving away from it. But we just thought it was important, and we thought that where there was a will there would be a way. And in fact it turned out to be a great way to make money—doing search with targeted advertising. And I think you'll find the same sort of thing here."

The key legal question is whether the courts will allow Google to continue to scan copyrighted material without permission. But the schedule of the lawsuits may turn out to be as significant as the merits of the cases, which are before Judge John E. Sprizzo. In keeping with the stately pace of federal litigation, the depositions of witnesses are to begin sometime this year, and the parties will be allowed to file motions for summary judgment—in Google's case, to dismiss the suits—in early 2008. Then there could be a trial. If the cases are appealed, they could linger well into the next decade.

However, most people involved in the dispute believe that a settlement is likely. "The suits that have been filed are a business negotiation that happens to be going on in the courts," Marissa Mayer told me. "We think of it as a business negotiation that has a large legal-system component to it." According to Pat Schroeder, the former congresswoman, who is the president of the Association of American Publishers, "This is basically a business deal. Let's find a way to work this out. It can be done. Google can license these rights, go to the rights holder of these books, and make a deal."

The terms of such a deal aren't hard to imagine. The Authors Guild is concerned that pirated copies of the books on Google's site could leak to the public, and so the organization would insist on security measures. (Sadly, for writers and publishers, demand for their products has never been robust enough to generate a major piracy problem.) As for distribution of the proceeds from the site, Google might agree to share revenue with publishers, in the way that radio stations pay for the music they play; publishers could receive a fee based on a statistical analysis of how often their books are viewed. Google could pay in cash or in kind, with advertising.

But a settlement that serves the parties' interests does not necessarily benefit the public. "It's clearly in both sides' interest to settle," Lawrence Lessig, a professor at Stanford Law School, said. "Businesses in Internet time can't wait around for years for lawsuits to be resolved. Google wants to be able to get this done, and get permission to resume scanning copyrighted material at all the libraries. For the publishers, if Google gives them anything at all, it creates a practical precedent, if not a legal precedent, that no one has the right to scan this material without their consent. That's a win for them. The problem is that even though a settlement would be good for Google and good for the publishers, it would be bad for everyone else."

Libraries have recognized for some time that they must adapt to the digital age, and many have taken steps in that direction. In 1995, Stanford founded the HighWire Press, which now provides electronic access to more than a thousand scholarly journals. A few years later, Stanford digitized most of its card catalogue, and circulation of its books increased by fifty per cent. "Once our students could sit in their dorm rooms and find out what we had in the library, they sought out more books," Michael Keller, the university librarian, says. Individual libraries sometimes received grants to scan specific collections—in 2001, the New York Public Library used federal money to digitize a substantial portion of the collection at its Schomburg Center for Research in Black Culture—but a comprehensive effort seemed inconceivable. According to Paul LeClerc, who has been the president of the New York Public Library for the past thirteen years, "For the first decade of my tenure, I was always asked, 'Weren't libraries going to go online?' And I'd say of course we want to do it, but it's not going to happen, because no one is going to give us the money to do it. Nowhere on the horizon was that amount of money predictable or identifiable. Then came Google. This struck us as being the quickest, the fastest, and the most efficient way of getting large-scale additions to our collections online for free use."

Among Google's potential competitors in the field of library digitization are members of the Open Content Alliance, which facilitates various scanning projects around the country and overseas. Funded largely by Microsoft and the Alfred P. Sloan Foundation, the O.C.A. has formed alliances with many companies and institutions, including the Boston Public Library, the American Museum of Natural History, and Johns Hopkins University. For the moment, though, the O.C.A.'s members are copying only material in the public domain (and works from copyright owners who have given explicit permission), which limits the scope of the projects substantially.

Google's advantage may well be cemented if the company settles its lawsuits with the publishers and authors. "If Google says to the publishers, 'We'll pay,' that means that everyone else who wants to get into this business will have to say, 'We'll pay,' " Lessig said. "The publishers will get more than the law entitles them to, because Google needs to get this case behind it. And the settlement will create a huge barrier for any new entrants in this field."

In other words, a settlement could insulate Google from competitors, which would be especially troubling, because the company has already proved that when it comes to searches it is not infallible. "Google didn't get video search right—YouTube did," Tim Wu, a professor at Columbia Law School, said. (Google solved that problem by buying YouTube last year for $1.6 billion.) "Google didn't get blog search right—technorati.com did," Wu went on. "So maybe Google won't get book search right. But if they settle the case with the publishers and create huge barriers to newcomers in the market there won't be any competition. That's the greatest danger here."

The most striking thing about Pajama Day at Google was how few people participated. Most of the rank and file saw the stunt for the manufactured fun that it was. They came to work in their usual slacker uniforms of jeans and T-shirts—which are, in their way, as conformist as white shirts and ties were at I.B.M. in the nineteen-sixties. Google, as its employees seem to recognize, cannot pretend to be anything other than a large and powerful corporation.

It's easy to mock Google's unofficial motto— "Don't be evil"—but there is nothing evil about Google Book Search. At the same time, there is nothing inherently virtuous about it. Google has succeeded because, on the whole, it has developed excellent products; it's folly to judge the company's behavior on moral grounds. Its shareholders certainly don't.

Nor can publishers and authors, who are struggling for a way to survive in a new age, portray their conflict with the company as one between good and evil. The dual status of several leading publishers as both partner and adversary to Google underscores their desperate need to hedge their bets in a digital world that they have yet to master. The publishers' complaint against Google states that "the Publishers support making books available in digital form so that those books can be, among other things, researched through electronic means." That may be true in theory, but trade publishers, in particular, have been slow to embrace new technology, especially for out-of-print books; Google will almost certainly bring more attention to these works than their own publishers have.

The law is supposed to resolve issues like these—between self-interested parties with reasonable claims and legitimate arguments. But the rules of copyright are so ambiguous, and the courts so slow, that the judicial system serves largely to implement the law of the jungle. "There is a real opportunity to move books into the digital arena," Marissa Mayer told publishers during the conference at the New York Public Library. "And we are going to do it together."

Google shareholders may criticize the companyâ€™s moonshot projects like self-driving cars and smart contact lenses as big wastes of money. But chairman Eric Schmidt says the futuristic projects guarantee Googleâ€™s future success. Schmidt, speaking at the companyâ€™s annual shareholder meeting Wednesday, defended the experiments as necessary for creating new and potentially blockbuster businesses.