Scan This Book!

Library shelf’s worth of specialized information. And as with music playlists, once created, these “bookshelves” will be published and swapped in the public commons. Indeed, some authors will begin to write books to be read as snippets or to be shared with fellow readers. Marginalia can be broadcast. Bibliographies swapped. You might get an alert that your friend Carl has annotated a favorite book of yours and is linking his thoughts to other pages and other books. Once text is digital, books seep out of their bindings and weave themselves together. The collective intelligence of a library allows us to see things we can’t see in a single, isolated book.

Next come the words. Just as a Web article on, say, aquariums, can have some of its words linked to definitions of fish terms, any and all words in a digitized book can be hyperlinked to other parts of other books. Books, including fiction, will be integrated in full, by fans for fans. Once a book's full text is online, anyone can revisit its pages — past and present, multilingual — on a particular subject, then you can have a clearer sense of what we as a civilization, a species, do know and don’t know. The white spaces of our collective ignorance are highlighted, while the golden nuggets of our accumulated wisdom are illuminated.

The arsenal of our current display technology — from handheld gizmos to large flat screens — is already good enough to move books to their next stage of evolution: a full digital scan.

In addition to a link, which explicitly connects one word or sentence or book to another, readers will also be able to add tags, a recent innovation on the Web but already a popular one. A tag is a public annotation, like a keyword or category that can be integrated in full, by fans for fans. Those with a passion for a special subject, obscure author or favorite book will, over time, link up its important parts. Multiply that simple generous act by millions of readers, and the universal library is no longer an island, but the boundaries of all subject fields worldwide who are underserved by ordinary paper books. It is these underbooked — students in Mali, scientists in Kazakhstan, elderly people in Peru — whose lives will be transformed when even the simplest unadorned version of the universal library is placed in their hands.

2. What Happens When Books Connect

The most important, but most obscured, aspects of digital reading have been these contentious questions: Who will get to write the highbrow technology of ink on paper and instead read on cumbersome machines? Or will we keep reading our paperbacks on the beach? For now, the answer is yes to both. Publishers, who have lost millions of dollars on the long-posted e-book revolution that never occurred, while the number of physical books sold in the world each year continues to grow. At the same time, there are already more than a half a billion PDF documents on the Web that people read on computers without printing them out, and all more people now spend hours watching movies on microscopically small cell phone screens. The arsenal of our current display technology — from handheld gizmos to large flat screens — is already good enough to move books to their next stage of evolution: a full digital scan.

Yet the common vision of the library’s future (even the e-book future) assumes that books will remain isolated islands, independent first as one, then just as they are on shelves in your public library. There, each book is pretty much unaware of the other books in the world. The Web, on the other hand, has already been fully digitized, much of it by us. About one tenth of the 300,000 or so movies listed on the Internet Movie Database are now digitized on DVD. But because of copyright issues and the physical fact of the need to turn pages, the digitization of books has proceeded at a relative crawl. At most, one book in 20 has moved from analog to digital. So far, the universal library is a library without many books.

The dream is an old one: to have in one place all knowledge, past and present. It is a familiar hope, in part because long ago we briefly built such a library. The great library at Alexandria, considered one of the seven wonders of the ancient world, was designed as half a million volumes. The ancient Egyptians had already begun a vast, unbroken record of written history — copies of all the works of each generation, from a library's worth of specialized information. And as with music playlists, once created, these “bookshelves” will be published and swapped in the public commons. Indeed, some authors will begin to write books to be read as snippets or to be shared with fellow readers. Marginalia can be broadcast. Bibliographies swapped. You might get an alert that your friend Carl has annotated a favorite book of yours and is linking his thoughts to other pages and other books.

Once a book has been integrated into the new expanded library of mankind, its text will no longer be separate from the text in other books. For instance, in this book, you will find a link to The Atlantic’s article on the history of the Web. In the same way, you can read about your favorite TV show, the history of the Web, or any topic — all books about Sweden, for instance, or books on clocks. Once snippets, articles and pages of books become ubiquitous, ubiquitous and transferable, users will earn prestige and perhaps income for creating an excellent collection.

Libraries (as well as many individuals) aren't eager to relinquish old habits, for the printed book is a kind of technology that supports the power of relationships, which is all links really are. There are about 14 billion Web pages, and each page holds, on average, 10 links. That's a trillion electrified connections coursing through the Web. This tangled relationship is precisely what gives the Web its immense force. The book world of knowledge is about to be transformed by the same devotion of relationships, as each page in a book discovers other pages and other texts. Once text is digital, books keep out of their bindings and weave themselves together. The collective intelligence of a library allows us to see things we can’t see in a single, isolated book.

When books are digitized, reading becomes a community activity. Bookmarks can be shared with fellow readers. Marginalia can be broadcast. Bibliographies swapped. You might get an alert that your friend Carl has annotated a favorite book of yours. A moment later, his links are yours. In a curious way, the universal library becomes very, very, very large single text: the world's only book.


At the same time, once digitized, books can be unraveled into simple pages or be reduced further, into snippets of a page. These snippets will be remixed into rendered books and virtual bookshelves. Just as the music audience now juggles playlists, the book audience will now be able to create “bookshelves” of their own. Some readers may read an entire book in snippets, then compose their own bookshelves. These bookshelves will be published and swapped in the public commons. Indeed, some authors will begin to write books to be read as snippets or to be shared with fellow readers. Marginalia can be broadcast. Bibliographies swapped. You might get an alert that your friend Carl has annotated a favorite book of yours and is linking his thoughts to other pages and other books.

At the same time, someone else can compose a new book from the same snippets. Indeed, a whole new kind of intelligence will emerge on the Web: a form of intelligence common on the Web, but previously foreign to the world of books. This intelligence is measured by the number of times a book is linked to by others on the Web. So far, the number of times a book is linked to is a kind of intelligence common on the Web, but previously foreign to the world of books. This form of intelligence is measured by the number of times a book is linked to by others on the Web. So far, the number of times a book is linked to is a kind of intelligence common on the Web, but previously foreign to the world of books. This form of intelligence is measured by the number of times a book is linked to by others on the Web. So far, the number of times a book is linked to is a kind of intelligence common on the Web, but previously foreign to the world of books. This form of intelligence is measured by the number of times a book is linked to by others on the Web. So far, the number of times a book is linked to is a kind of intelligence common on the Web, but previously foreign to the world of books. This form of intelligence is measured by the number of times a book is linked to by others on the Web. So far, the number of times a book is linked to is a kind of intelligence common on the Web, but previously foreign to the world of books. This form of intelligence is measured by the number of times a book is linked to by others on the Web. So far, the number of times a book is linked to is a kind of intelligence common on the Web, but previously foreign to the world of books. This form of intelligence is measured by the number of times a book is linked to by others on the Web.
The desire of all creators is for their works to find their way into many minds. A text, a melody, a picture or a story succeeds best if it is connected to as many ideas and other works as possible. Ideally, over time a work becomes so entangled in a culture that it exists as a fabric of common culture. This is how ideas spread, in the way that the Bible, Shakespeare’s plays, “Cinderella” and the Mona Lisa are inseparable from ours. This tendency for creative ideas to infuse other works is great in culture. In fact, this commingling of creations is culture.

In prudential days, exact copies of a work were rare for a simple reason: it was much easier to make your own version of a creation than to duplicate someone else's exactly. The amount of energy and attention needed to copy a scroll exactly would be enormous, since each new manuscript required the effort of tracing the original with a stylus. But in the wake of writing machines that could produce texts rapidly at a high quality, and the electronic storage of information, deep as many different authors worked on them and as they migrated from spoken tales to other media (furniture, music, painting). This system worked well for audiences and performers, but the only way for such creators to earn from their works was through the support of patrons.

That ancient economy of ideas has been outweighed at the dawn of the industrial age by the technologies of mass production. Suddenly, the cost of duplication was less than the cost of appropriation. With the advent of the printing press, it was now cheaper to print exact copies of a manuscript than to allow one by hand. Copy makers could profit more than creators. This imbalance led to the technology of copyright, which established a new order. Copyright bestowed upon the creator a property—a temporary monopoly—for 14 years in the United States — over any copies of the work. The idea was to encourage artists and authors to create yet more works that could be cheaply copied and thus fill the culture with public works.

Not coincidentally, public libraries first began to flourish with an ample vault of copies. Before the industrial age, libraries were primarily the property of the wealthy elite. With mass production, every small town could afford to put duplicates of the great works of literature, philosophy, and science in their public libraries. This shift fueled literary creativity and education, and helped to spread the benefits of copyright in the United States because they made creative works toward the public commons sooner, weaving them into the fabric of common culture while still remaining under the author’s copyright. These are now known as “fair uses.”

This wonderful balance was undone by bad intentions. The first was a new copyright law passed by Congress in 1976. According to the new law, creators no longer had to register or renew copyright; the simple act of creating something bestowed it automatically. By default, every print or electronic book is covered by copyright. At first, this was a boon to authors. They gained instant and deep ownership, and artists and authors were happy. But the 1976 act, and various revisions and extensions that followed it, made it extremely difficult to move a work into the public commons, where human creations naturally spread and multiply. Copyright became a tool to lock up valuable intellectual property corporate interests. By the early 1990s, Congress was extending the one-time brief protection enabled by copyright in order to prevent works from entering the public domain. With constant nudging, Congress moved the expiration date from 1972 to 28 by 42 and then to 56.

While corporations and legislatures were moving the goal posts back, technology was accelerating forward. In Internet time, even 14 years is a long time for a monopoly; a monopoly that lasts a human lifetime is essentially an eternity. So when Congress passed the DMCA (Digital Millennium Copyright Act) in 1998, with its original 10-year expiration date, it had 10 additional years to 2009. The net effect of this legislation is that copyright now existed primarily to protect a threatened business model. And because Congress at the same time tucked in a 20-year extension onto all existing copyrights, nothing — no published creative works of any type — will fall out of protection and enter the commons until after 2050. Almost everything created today will not return to the commons until the next century. Thus the streams of shared material that anyone can improve (think: “A Thousand and One Nights” or “Amazing Grace” or “Beauty and the Beast”) will largely dry up.

In the books of the industrial age, the indeterminate extent of copyright has had a perverse effect. It has created a vast collection of works that have been abandoned by publishers, a continent of books left permanently in the dark. In most cases, the original publisher simply doesn’t find it profitable to keep those books in print. In other cases, the publishing company doesn’t know whether it even owns the work, since author contracts in the past were not as explicit as they are now. The state of this abandoned library is shocking: about 75 percent of all books in the library’s holdings are orphaned. Only about 15 percent of all books are in the public domain. A lucky 10 percent are in print. The rest, the bulk of our universal library, is dark.

7. The Moral Imperative to Scan

The 15 percent of the world’s 75 million cataloged books that are in the public domain are freely available for anyone to borrow, imitate, publish or copy wholesale. Almost the entire current scanning effort by American libraries is aimed at this 15 percent. The Million Book Project misses this sly sliver of the pie, as does Google. Because they are in the commons, no law hinders this 15 percent from being scanned and added to the universal library.

At present about 10 percent of the actively published in print will also be scanned before long. Amazon carries at least four million books, which includes multiple editions of the same title. Amazon is slowly scanning all of them. Recently, several big publishers have announced plans to scan all their stocks. Google, meanwhile, has a partnership in a program in which Google scans their books, offers sample pages (controlled by the publisher) to readers and points readers to where they can buy the new book. No one doubts electronic books will become money eventually. Simple commercial incentives guarantee that all in-print and backlist books will be scanned before long.

The major problem for large publishers is that they are not certain that what they actually own. If you would like to assign yourself, pick an out-of-print book from the library and try to determine who owns its copyright. It’s not easy. There is no list of copyrighted works. The Library of Congress doesn’t have an exhaustive list, even of their own inventors (though they say they are working on it). The older, the more obscure the work, the less likely a publisher will have a legal hold on the copyright. A publisher dies along with the publishing company it represents. Even then, unless the publisher actually owned the copyright, making copies of its own publications to be scanned by Google amounts to digital piracy. There’s no legal basis for such a position.

Publishers and libraries have been fighting back. They’ve won the right to challenge an individual challenge in court. The libraries lost their case against HarperCollins and have been forced to stay out of search. (Few do.) Google applies the same principle of opting-out to Book Search. It is up to you as an author to notify Google if you don’t want the company to scan or search your copyrighted material. This might be a reasonable approach when you have connections and are not an orphan. But in practice, it’s a waste of time and money. The problem here is: The copyright is owned by the publisher and the copyright is better owned by the creator. There is no way to resolve this problem. The searches of the public domain are not going to stop. The only way to resolve this problem is to give the creators the copyright to their works.

Google’s plan was to scan the full text of every book in five major libraries: the more than 10 million titles held by Stanford, Harvard, Oxford, the University of Michigan and the New York Public Library. Every book would be indexed, but each would be up in search results in different ways. For out-of-copyright books, Google would show the whole book, page by page. For in-print books, Google would work with publishers and let them decide what parts of their books would be shown and what conditions. For the dark orphans, Google would show only limited snippets. And any copyright holder (author or corporation) who could establish ownership of a supposed orphan could ask Google to remove the snippets for any reason.

At first glance, it seemed genius. By scanning all books (something only Google had the cash to do), the company would advance its mission to organize all knowledge. It would let books be searchable, and it could potentially sell ads on those searches. It was, in short, a solution in search of its problem. After the initial pages of publicity, the author and publisher communities were not as enthusiastic. Google lawyers argued that the snippets the company would display were not the same as a copy or an excerpt. Instead, they argued, the snippets should be viewed as “fair uses.”

Some authors and many publishers found more evil than genius in Google’s plan. Two points outraged them: the virtual copy of the book that sat on Google’s indexing server and Google’s assumption that it could scan first and ask questions later. The first would be a real problem, if not solved, for anyone who wants to sell their book through other companies. The second problem would never get that far.

There is a technical solution to this problem: for the search companies to compile and maintain a common list of no-scan copyright holders. A publisher or author who doesn’t want a work scanned notifies the keepers of the common list once, and that’s all. Google would be prohibited from scanning that work. Even if a publisher or author had rights in a work, and if they didn’t notify the lists, the company could ask the libraries to remove any book from their collections.

The second complaint against Google is much more complex. Google argues that it is nearly impossible to track copyright holders of orphan works, so, it says, it must scan those books first and only afterward honor any legitimate requests to remove the scan. In this way, Google follows the protocols of the Internet. Google scans all Web pages if: it’s on the Web; it’s an image; by default, are born copyrighted. Google, therefore, regularly copies billions of copyrighted pages into its database, regardless of whether you don’t want Google to search your Web site, you can stick code on your site. In order to solve this, you will have to opt out of search. (Few do.) Google applies the same principle of opting-out to Book Search. It is up to you as an author to notify Google if you don’t want the company to scan or search your copyrighted material. This might be a reasonable approach for Google to demand from an author or publisher if Google were the only search engine. But search technology is becoming a commodity, and if it turns out there is any money in it, it is not impossible to imagine a hundred merceless scanners scanning-out-of-print books. Should you use a crutcher to be forced to find and notify each and every book to which you own the copyright. If you don’t notify us, you don’t want it indexed? Is it what you mean?

There is no technical solution to the problem of the search companies to compile and maintain a common list of no-scan copyright holders. A publisher or author who doesn’t want a work scanned notifies the keepers of the common list once, and anyone using the scanning technology would have to remove material that was listed. Since Google, like all the other big search companies — Microsoft, Amazon and Yahoo! — is forestmore a technical solution company, it favors this approach. But the battle never got far.

8. When Business Models Collide

In thinking about the future of book publishing and search, I realized that there are many ways to conceive of this conflict. At first, I thought that this was a misunderstanding between people of the book, who favor solutions by law, and people of the screen, who favor technology as a solution in all problems. But so far they have not worked. At least not in great numbers. And there is only one reason: the hegemony of the copy.

There are dozens of excellent reasons that books should quickly be made part of the emerging Web. But so far they have not been, at least not in great numbers. And there is only one reason: the hegemony of the copy.

Many search engines, including Google, operate in the open in order to compete. Search is a whole new concept, not formalized in version 1.0 of our intellectual-property law. In the words of a recent ruling by the United States District Court for Nevada, search has a “transformative purpose,” adding new social value to what it searches. What search functions are not just keywords but also the inherent value of connection. While any artist recognizes that the value of a creation ultimately rests in the work by the artist and he personally gets from creating it (and for a few artists that value is sufficient), it is also true that the value of any work is increased the more it is shared. The technology of search maximizes the value of a
An article on Page 42 of The Times Magazine today about the future of book publishing misstates the number and type of libraries in which a Chinese company, Superstar, has made digital copies of books. It is 200 libraries of all kinds, not 900 university libraries.

Correction: May 14, 2006

An article on Page 42 of The Times Magazine today about the future of book publishing misstates the number and type of libraries in which a Chinese company, Superstar, has made digital copies of books. It is 200 libraries of all kinds, not 900 university libraries.

Creative work by allowing a billion new connections into it, often a billion new connections that were previously inconceivable. Things can be found by search only if they radiate potential connections. These potential relationships can be as simple as a title or as deep as hyperlinked footnotes that lead to active pages, which are also footnoted. It may be as straightforward as a song published intact or as complex as access to the individual instrument tracks — or even individual notes.

Kevin Kelly is the “senior maverick” at Wired magazine and author of “Out of Control: The New Biology of Machines, Social Systems and the Economic World” and other books. He last wrote for the magazine about digital music.

Search opens up creations. It promotes the civic nature of publishing. Having searchable works is good for culture. It is as good, in fact, that we can now see a new covenant. Copyrights must be counterbalanced by copypathics. In exchange for public protection of a work’s copies (what we call copyright), a creator has an obligation to allow that work to be searched. No search, no copyright. As a song, movie, novel or poem is searched, the potential connections it radiates seep into society in a much deeper way than the simple publication of a duplicated copy ever could.

We see this effect most clearly in science. Science is on a long-term campaign to bring all knowledge in the world into one vast, interconnected, footnoted, peer-reviewed web of facts. Independent facts, even those that make sense in their own world, are of little value to science. (The pseudo- and perennials are nothing less, in fact, than small pools of knowledge that are not connected to the large network of science.) In this way, every new observation or bit of data brought into the web of science enhances the value of all other data points. In science, there is a natural duty to make what is known searchable. No one argues that scientists should be paid when someone finds or duplicates their results. Instead, we have devised other ways to compensate them for their valuable work. They are rewarded for the degree that their work is cited, shared, linked and connected in their publications, which they do not own: They are financed with extremely short-term (20-year) patent monopolies for their ideas, short enough to truly inspire them to invent more, sooner. To a large degree, they make their living by giving away copies of their intellectual property in one fashion or another.

But the reign of livelihoods based on the copy is not over. In the next few years, lobbyists for book publishers, movie studios and record companies will exert every effort to mandate the extinction of the “indiscriminate flow of copies,” even if it means outlawing better hardware. Too many creative people depend on the business model revolving around copies for it to pass quietly. For their benefit, copyright law will not change suddenly.

But it will adapt eventually. The reign of the copy is no match for the bias of technology. All new works will be born digital, and they will flow into the universal library as you might add more words to a long story. The great continent of orphan works, the 250 million older books born analog and caught between the law and users, will be scanned. Whether this vast mountain of dark books is scanned by Google, the Library of Congress, the Chinese or by readers themselves, it will be indexed and copied again. What counts are the ways in which these common copies of a creative work can be linked, manipulated, annotated, tagged, highlighted, bookmarked, translated, enlivened by other media and sewn together into the universal library. Soon a book outside the library will be like a Web page outside the Web, gasping for air. Indeed, the only way for books to retain their waning authority in our culture is to wire their texts into the universal library. But the reign of isolated books born analog and caught between the law and users, will be scanned. Whether this vast mountain of dark books is scanned by Google, the Library of Congress, the Chinese or by readers themselves, it will be indexed and copied again. What counts are the ways in which these common copies of a creative work can be linked, manipulated, annotated, tagged, highlighted, bookmarked, translated, enlivened by other media and sewn together into the universal library. Soon a book outside the library will be like a Web page outside the Web, gasping for air. Indeed, the only way for books to retain their waning authority in our culture is to wire their texts into the universal library.

But the reign of livelihoods based on the copy is not over. In the next few years, lobbyists for book publishers, movie studios and record companies will exert every effort to mandate the extinction of the “indiscriminate flow of copies,” even if it means outlawing better hardware. Too many creative people depend on the business model revolving around copies for it to pass quietly. For their benefit, copyright law will not change suddenly.

But it will adapt eventually. The reign of the copy is no match for the bias of technology. All new works will be born digital, and they will flow into the universal library as you might add more words to a long story. The great continent of orphan works, the 25 million older books born analog and caught between the law and users, will be scanned. Whether this vast mountain of dark books is scanned by Google, the Library of Congress, the Chinese or by readers themselves, it will be indexed and copied again. What counts are the ways in which these common copies of a creative work can be linked, manipulated, annotated, tagged, highlighted, bookmarked, translated, enlivened by other media and sewn together into the universal library.

Kevin Kelly is the “senior maverick” at Wired magazine and author of “Out of Control: The New Biology of Machines, Social Systems and the Economic World” and other books. He last wrote for the magazine about digital music.

Correction: May 14, 2006

An article on Page 42 of The Times Magazine today about the future of book publishing misstates the number and type of libraries in which a Chinese company, Superstar, has made digital copies of books. It is 200 libraries of all kinds, not 900 university libraries.

Creative work by allowing a billion new connections into it, often a billion new connections that were previously inconceivable. Things can be found by search only if they radiate potential connections. These potential relationships can be as simple as a title or as deep as hyperlinked footnotes that lead to active pages, which are also footnoted. It may be as straightforward as a song published intact or as complex as access to the individual instrument tracks — or even individual notes.

Kevin Kelly is the “senior maverick” at Wired magazine and author of “Out of Control: The New Biology of Machines, Social Systems and the Economic World” and other books. He last wrote for the magazine about digital music.

Correction: May 14, 2006

An article on Page 42 of The Times Magazine today about the future of book publishing misstates the number and type of libraries in which a Chinese company, Superstar, has made digital copies of books. It is 200 libraries of all kinds, not 900 university libraries.