Wired | http://www.wired.com/news/business/0,1367,60948,00.html November 2003

The Great Library of Amazonia

by Gary Wolf

The fondest dream of the information age is to create an archive of all knowledge. You might call it the Alexandrian fantasy, after the great library founded by Ptolemy I in 286 BC. Through centuries of aggressive acquisition, the librarians of Alexandria, Egypt, collected hundreds of thousands of texts. None survives. During a final wave of destruction, in AD 641, invaders fed the bound volumes and papyrus scrolls into the furnaces of the public baths, where they are said to have burned for six months. "The lesson," says Brewster Kahle, founder of the Internet Archive, "is to keep more than one copy."

Kahle recently gave a copy of his digital archive of 10 billion Web pages to a new library in Alexandria. On a visit to the city last year, he sat down with Suzanne Mubarak, the wife of Egypt's president, and discussed his gift, which has all the advantages of a modern electronic resource: It can be instantly updated, easily searched, and endlessly replicated. Mubarak, with diplomatic politeness, allowed that she was impressed. Still, she ventured a protest: "But I love books!"

Therein lies a problem. Books are an ancient and proven medium. Their physical form inspires passion. But their very physicality makes books inaccessible to the multi-terabyte databases of modern Alexandrian projects. Books take time to transport. Their text vanishes and their pages yellow in a rash of foxing. Most important, it's still shockingly difficult to find information buried in books. Even as the Internet has revived hope of a universal library and Google seems to promise an answer to every query, books have remained a dark region in the universe of information. We want books to be as accessible and searchable as the Web. On the other hand, we still want them to be books.

An ingenious attempt to illuminate the dark region of books is under way at Amazon.com. Over the past spring and summer, the company created an unrivaled digital archive of more than 120,000 books. The goal is to quickly add most of Amazon's multimillion-title catalog. The entire collection, which went live Oct. 23, is searchable, and every page is viewable.

To build the archive, Amazon CEO Jeff Bezos has had to unravel a tangle of technological and copyright problems. His solution promises to remake the publishing business and give Amazon a powerful new weapon in its battle against online competitors such as Yahoo, Google, and eBay. But the most interesting thing about the archive is the way it resolves the paradox of the book, respecting its physical form while transcending its limits.

I recently drove to a home in Silicon Valley and spent a few hours digitally searching the text of books. My host was Udi Manber, an Israeli-born computer scientist and author of a popular textbook, *Introduction to Algorithms: A Creative Approach*. Ten years ago, while developing a seminal piece of Unix search software called agrep, Manber came up with a concept for an information tool he has yet to build. It was supposed to search the mess of papers on his desk. The idea that you could perform a digital search of physical objects has long fascinated him. "Why not have users take pictures of their bookshelf?" Manber asked when we first met. "We could scan the images, extract the titles, and then let them search the entire text of the books they own."

The notion of Amazon scanning all of its books but allowing users to search only those they own is a clever way around the central barrier to creating a digital archive: Copyrights are distributed among tens of thousands of publishers and authors. But when Manber told Bezos his idea, he found the Amazon founder ready to work on a grander scale. Bezos wanted his customers to be able to search everything.

In his small, ranch-style Palo Alto house, Manber and I sit side by side at a table near the kitchen as he begins typing my queries into his laptop. The computer is connected to a prototype of the archive, which at the time of my visit is scheduled to go live in a few weeks. Within seconds, I am captivated. The experience reminds me of how I felt a decade ago, when I first began browsing the Web. Back then, the Web was still small, and most of my time was spent peeking into the homepages of physicists and engineers. Even so, the power of the new network was unmistakable. The thrill didn't come from the content of the pages but from the structure of the Web itself, its obvious scalability and ease of navigation.

Amazon's new archive is more densely populated than the early Web was, but it's still far from complete. With its 120,000 titles, the archive has about as many books as a big brick-and-mortar store. Still, this is plenty to create a familiar sensation of vertigo as an expansive new territory suddenly opens up.

The more specific the search, the more rewarding the experience. For instance, I've recently become interested in Boss Tweed, New York's most famous pillager of public money. Manber types "Boss Tweed" into his search engine. Out pop a few books with *Boss Tweed* in the title. But the more intriguing results come from deep within books I never would have thought to check: *A Confederacy of Dunces,* by John Kennedy Toole; *American Psycho,* by Bret Easton Ellis; *Forever: A Novel,* by Pete Hamill. I immediately recognize the power of the archive to make connections hitherto unseen. As the number of searchable books increases, it will become possible to trace the appearance of people and events in published literature and to follow the most digressive pathways of our collective intellectual life.

From the Hamill reference, I link to a page in the afterward on which he cites books that influenced his portrait of Tweed. There, on the screen, is the cream of the research performed by a great metropolitan writer and editor. Some of the books Hamill recommends are out of print, but all are available either new or used on Amazon.

With persistence, serendipity and plenty of time in a library, I may have found these titles myself. The Amazon archive is dizzying not because it unearths books that would necessarily have languished in obscurity, but because it renders their contents instantly visible in response to a search. It allows quick query revisions, backtracking, and exploration. It provides a new form of map.

Getting to this point represents a significant technological feat. Most of the material in the archive comes from scanned pages of actual books. This may be surprising, given that most books today are written on PCs, e-mailed to publishers, typeset on computers, and printed on digital presses. But many publishers still do not have push-button access to the digital files of the books they put out. Insofar as the files exist, they are often scattered around the desktops of editors, designers, and contract printers. For books more than a few years old, complete digital files may be lost. John Wiley & Sons contributed 5,000 titles to the Amazon project—all of them in physical form.

Fortunately, mass scanning has grown increasingly feasible, with the cost dropping to as low as $1 each. Amazon sent some of the books to scanning centers in low-wage countries like India and the Philippines; others were run in the United States using specialty machines to ensure accurate color and to handle oversize volumes. Some books can be chopped out of their bindings and fed into scanners, others have to be babied by a human, who turns pages one by one. Remarkably, Amazon was already doing so much data processing in its regular business that the huge task of reading the images of the books and converting them into a plain-text database was handled by idle computers at one of the company's backup centers.

The copyrights to these titles are spread among countless owners. How was it possible to create a publicly accessible database from material whose ownership is so tangled? Amazon's solution is audacious: The company simply denies it has built an

electronic library at all. "This is not an ebook project!" Manber says. And in a sense he is right. The archive is intentionally crippled. A search brings back not text, but pictures—pictures of pages. You can find the page that responds to your query, read it on your screen, and browse a few pages backward and forward. But you cannot download, copy, or read the book from beginning to end. There is no way to link directly to any page of a book. If you want to read an extensive excerpt, you must turn to the physical volume—which, of course, you can conveniently purchase from Amazon. Users will be asked to give their credit card number before looking at pages in the archive, and they won't be able to view more than a few thousand pages per month, or more than 20 percent of any single book.

Manber has built a powerful, even mind-boggling tool, then added powerful constraints. "The point is to help users find a book," says Manber, "not to make a new source of information."

Bezos is vehement on this point. He has sold publishers on the idea that digitizing hundreds of thousands of copyright books won't undermine the conventional book-selling business. "It is critical that this be understood as a way to get publishers and authors in contact with customers," he says in an interview at Amazon's Seattle headquarters. "We're perfectly aligned with these folks. Our goal is to sell more books!"

Bezos has some good evidence to back up his argument. Amazon has consistently added features that have proven to increase book sales. Through its customer reviews, used-book business, and personalized recommendations, the company constantly puts its customers in contact with new titles. Amazon is a machine that stimulates the acquisitive urge of readers. It appeals to their specialized interests.

It makes people buy books. But Amazon's scheme would never work if users really wanted their books in digital form. The magic of the archive lies in the assumption that physical books are irreplaceable. The electronic text is simply an enhancement of the physical object.

The Amazon project—dubbed Search Inside the Book—represents a bold step toward the dream of a universal library. Bezos refuses any such allusion. But outlines of the Alexandrian fantasy can clearly be made out in Amazon's innocent book-purchasing tool. The company's success at launching a massive archive of digital books will undoubtedly fuel enthusiasm for overturning the current publishing and copyright regime.

I first talked with Brewster Kahle a dozen years ago over a plate of execrable spaghetti in the kitchen of a flat in San Francisco's Mission District. The apartment served as the headquarters of WAIS, one of the first Internet search engines, and Kahle was sharing a dinner with me and several of his employees. He was in his twenties then—thin, with a mass of unruly curly hair, a rapid manner of speech, and an unguarded expression. Kahle was already one of the great enthusiasts of universal information access. A few months earlier, he had left his job at Thinking Machines, the legendary builder of massively parallel supercomputers, and devoted himself full-time to refining and selling WAIS.

To Kahle, it was obvious that vast amounts of useful material could be shared with the general public via computers, but the Web did not yet exist, and most of the major databases were not linked. You couldn't do a comprehensive search. WAIS was meant as a remedy, and it proved a modest success. However, its most significant contribution to Kahle's evangelical mission was a byproduct of the stock market bubble: In the spring of 1995, AOL bought WAIS for $15 million in stock. AOL stock soared, and Kahle became rich.

With his money, Kahle started the Internet Archive, while also creating another company that offered a clever Web search tool called Alexa. In 1999, as the bubble continued to expand, Alexa was sold to Amazon for $250 million in stock, and Kahle became richer. He's now committed to public service. The computers of the Internet Archive are in a Mission district warehouse. The headquarters are in a ramshackle house in the Presidio, a decommissioned Army base near the Golden Gate Bridge. The office

is one of those classic engineer-idealist domains, where programmers go up the fire stairs because the inside stairs are broken, and age-old pizza molders in the refrigerator.

When I call Kahle to ask if I can come talk to him about the state of digital libraries, he says, "Sure, I'm free right now." I find him substantially the same. "What's the average lifespan of a Web page?" Kahle asks me when we meet, and then answers himself: "One hundred days!" He has a slightly accusatory tone, as if I share in the general neglect that led to the erasure of history—or would, if the machines owned by the Internet Archive weren't so busily preserving it.

The goal of the archive is to save digital information and make it accessible to all. But what, exactly, is digital information? As the entertainment industry has learned to its great chagrin, digital information might take the form of music, or movies, or even books. To Kahle, this is a good thing. The products of human knowledge ought not to be squirreled away where they can't ever be found.

Kahle hates the idea that when people think of information, they think only of what's accessible via Google. "Seventy-one percent of college students use the Internet as their research tool of first resort," he says, citing figures from a 2001 PEW Internet Study. "Personally, I think this number is low. For most students today, if something is not on the Net, it doesn't exist."

And yet most books are not on the Net. This means that students, among others, are blind to the most important artifacts of human knowledge. For many students, the Internet actually contracts the universe of knowledge, because it makes the most casual and ephemeral sources the most accessible, while ignoring the published books. "It's shameful," Kahle continues, "because we have the tools to make all books available to everybody. You need three things. Technically, you need storage and connectivity. Storage is easy. For under $10 million, you can store all published works of humankind back to the Sumerian tablets. The last time they tried this was in Alexandria, and they had an innovative storage mechanism, too. They had papyrus, and papyrus was astonishing compared to clay tablets. But we can do better than the Alexandrians, because we also have connectivity. I have traveled in Uganda and in rural Kenya and seldom been more than one day's walk from an Internet café. It is technologically possible for most kids in the world to have access to all the books in the world."

The third item on Kahle's list has nothing to do with technological know-how; it's simply political will. Here, he finds the situation mixed. "We live in an open society in which the concept of widespread knowledge is embraced as a goal of governance," he says. "Just look at our libraries. Public libraries spend $7.6 billion a year; academic libraries spend another $5 billion." That's the good news. The budgets are hard evidence of a public commitment to the Alexandrian ideal. But on the other hand, almost none of this money goes to digitizing books.

The Internet Archive turned Kahle into an expert in managing huge databases of publicly accessible information. Now—in partnership with Carnegie Mellon University, the National Science Foundation, and the governments of India and China—his goal is to create a digital archive of 1 million books. Books from the US are packed into containers and shipped to India to be scanned and proofread, then the digital files go to the Internet Archive and the books are returned to the owners. Kahle and his partners are hoping to have about 100,000 online by the end of the year, making this project almost as big, at least numerically, as Amazon's effort. "We chose a million books because it's a big number," admits Kahle. "It's something you can strive for."

But in reality, the Million Book Project lags far behind Amazon's effort. For one thing, libraries have been slow to lend parts of their collections. And even then, the project concentrates on digitizing those that are out of copyright. Libraries and nonprofits don't have much leverage with publishers, and since the goal is fully readable online text, there is no system to protect the interests of copyright holders. As a result, many of

the titles being digitized by the Million Book Project are government documents, old texts, and books from India and China, where copyright laws are less stringent.

Kahle is happy to sidestep the problem of digitizing commercially successful books. He has no wish to antagonize the publishing industry. What he hates is that the Million Book Project cannot legally digitize countless books that aren't generating money for anybody. US libraries hold about 30 million unique volumes. No one knows how many of those books continue to be protected by copyright or are available from commercial publishers. Still, Kahle says, "they can't be digitized because the copyrights can't be cleared, and the copyrights can't be cleared because it's too much work to identify the copyright holders. Some people call them abandonware. I call them orphans."

"Amazon is taking a cut at the commercially available titles," continues Kahle. "We are going for the public domain titles. But who is taking care of the orphans? Nobody."

This is no longer true. Kahle's plea on behalf of orphaned books, stripped of its sentiment and restated in the rational voice of finance, exactly expresses the logic of Amazon's Alexandrian venture. Latent within the new archive is a business model for selling books that, with a little legal help, ought to vanquish orphanhood forever.

The publishing industry has made great strides since the Roman era. Movable type was invented in 11th-century China, then reinvented in 1450 in Germany. In 1886, Ottmar Mergenthaler created an automatic typesetting machine. In 1983, we got desktop publishing. But publishers continue to edit books using four colors of pencil, and the idea of freely accessible digital files conjures nightmares of a peer-to-peer disaster among media corporations. Things are even going backward—Barnes & Noble recently announced it would stop selling ebooks.

In this context of change, confusion, and fear, Jeff Bezos is forced to behave like a politician. Talk of a universal library elicits no enthusiasm from him. When I mention it, he counsels caution and patience. "You have to start somewhere," he says. "You climb to the top of the first tiny hill, and from there you see the next hill. It's difficult to see what's beyond before you have climbed the first hill."

When I met Bezos the first time, in 1996, there was no masking the radical nature of his ideas. At the time, he was trying hard to prove that you could create a major retail company on the Web. Skeptics abounded, and he answered them with vivid descriptions of the future. All of Amazon's important innovations—starting from the concept of a Web bookstore—have suggested a profound change in the bookselling business, a change that makes it possible to earn a profit by selling a much wider variety of books than any previous retailer, including many titles from the so-called long tail of the popularity curve. "If I have 100,000 books that sell one copy every other year," says Steve Kessel, an Amazon VP, "then in 10 years I've sold more of these, together, than I have of the latest Harry Potter."

In fact, Amazon doesn't have to wade far into the shallows to begin remaking the book business. Books are abandoned by publishers long before their sales are reduced to one copy every other year. Under the current publishing system, a title becomes inefficient at thousands of sales per year. An electronic archive through which readers can find books is an essential counterpart to Bezos' original vision of an infinitely big bookstore, just as Internet search engines are essential to the fragmented, increasingly diverse cultures of the Web.

This vision implies that readers will someday be able to purchase books that are printed at the time they are ordered. On a small scale, that phase of the revolution has already been quietly accomplished. As part of the Million Book Project, Kahle has created an Internet Bookmobile that produces decent-quality paperbacks of out-of-copyright books for about $1 each. The bookmobile consists of a Ford Windstar minivan with a satellite dish, a computer, a printer, and a binder. Meanwhile, last spring, Amazon announced a partnership with Ingram Industries' Lightning Source subsidiary,

a print-on-demand company that offers more than 100,000 titles—with a list that grows by hundreds each week. Lightning Source is the high end, Kahle's Internet Bookmobile is the low end; both operate on the premise that tiny runs of books can be affordably made and sold.

With these tools, the concept of out of print is becoming obsolete. A copyright-friendly archive that allows all books to be easily found plus a books-on-demand printing network gives publishers an economic motive for reactivating entire back catalogs. As for books whose copyright holders cannot be found—Kahle's orphans—this is where the copyright law needs to change. A sensible solution advanced by copyright scholar (and Wired columnist) Lawrence Lessig and written into a bill before Congress requires that copyright be renewed every 50 years for a token sum. Anybody who can't be bothered to pay a dollar or two to hold on to a copyright loses the work to the public domain.

And who will digitize the books once they've been claimed by their copyright holders or lapsed into the public domain? Project Gutenberg, the first book-digitizing initiative, has put about 10,000 titles online and is rapidly accelerating its effort. The Million Book Project, launched by Carnegie Mellon computer scientist Raj Reddy, is eager for more volumes. Lessig, in partnership with Stanford University librarian Michael Keller, will soon announce a free program to digitize any out-of-print book whose copyright holder wants to make it available to the public. And of course, Manber invites copyright holders to offer nonexclusive searching and browsing rights to Amazon, which will digitize the titles and offer access to audiences forever. "Give the books to me," Manber says. "I am glad to do it."

The original vision of a digital archive of all knowledge renounced paper volumes; physical books were seen as antiquated, like papyrus or clay tablets. But if electronic archives prove to raise the value of physical books, a new dream may replace the old one. After talking with Manber, I raise this question with Kevin Kelly, a Wired founding editor who spent part of his summer trying to establish a private cooperative library of digital books. The digital titles in Kelly's library would match the physical books on his shelf. "The idea of ebooks was to do away with paper," he says. "But really, you want to add dimensionality to a physical object rather than take it away. You want an enhanced physical world."

In this enhanced physical world, the logic of the book business is transformed. Human attention is limited, and a massive number of newly browsable books from the long tail necessarily compete with the biggest best-sellers, just as cable siphons audience from the major networks, and just as the Web pulls viewers from TV.

This shifts power away from the people who own finite sets of copyrighted material and toward the people who offer access to information about where this material can be found. Information about books, not ownership of copyrights, becomes a new center of power. Manber is correct when he says that Amazon's Search Inside the Book is not an ebook project. It is merely a catalog. But a decade of Internet history proves that the catalog is exactly what you want to own.

Of course, Amazon is not merely part of the book business. The Internet as a whole is going through a similar transition. Revenue for Internet companies increasingly comes from users seeking to buy something—from transactions made in the physical world. Even Google, which doesn't sell anything directly, earns most of its cash from advertisers whose messages pop up when users search for info about specific products, such as computers or cars—or books. With retail at the center of the Internet industry, Google is a key competitor because customers begin their online shopping trips at search engines that offer neat algorithms for comparing prices across multiple vendors. Everybody—Yahoo!, eBay, AOL, Microsoft, and, of course, Amazon—wants to be the site of first resort.

All the leading retail sites have better knowledge of their customers than Google.

But Google is the leading Internet information tool, period. Google is a window onto the entire Web. On the other hand, the contents of books may be the only publicly accessible data set with the potential to match Google's Web index both for size and utility. Search Inside the Book makes Amazon the sole guide to tens and ultimately hundreds of millions of pages of information. And while Google's business is vulnerable to any competitor that builds a better search engine, Amazon's book archive is the product of negotiated contracts with hundreds of publishers. Amazon has cornered the market on information that was once hidden away in books. The burden of the physical—the fact that the database Amazon uses is linked into a complex system involving real things—gives it a stunning, if perhaps temporary, advantage.

This fall, Amazon announced that it was forming A9.com, a new company devoted exclusively to search technologies. Manber, who leads it, came up with the name by running a simple compression algorithm on the word algorithms. Algorithms begins with A and is followed by nine other letters. When Manber explains the name to me, he notes mischievously that another word can be identically compressed: Alexandria.

Amazon's Alexandrian scheme hinges on the insight that physical books can be turned into electronic databases and then—in the retail process—turned back into physical books. This is one of the boldest maneuvers yet in an intense commercial competition, but for all its cunning, this is a civilized, even civilizing war, one that builds libraries rather than burns them.