THE NEW YORKER

ONWARD AND UPWARD WITH THE ARTS

FUTURE READING

Digitization and its discontents.

by Anthony Grafton

NOVEMBER 5, 2007

In 1938, Alfred Kazin began work on his first book, "On Native Grounds." The child of poor Jewish immigrants in Brooklyn, he had studied at City College. Somehow, with little money or backing, he managed to write an extraordinary book, setting the great American intellectual and literary movements from the late nineteenth century to his own time in a richly evoked historical context. One institution made his work possible: the New York Public Library on Fifth Avenue and Forty-second Street. Kazin later recalled, "Anything I had heard of and wanted to see, the blessed place owned: first editions of American novels out of those germinal decades after the Civil War that led to my theme of the 'modern'; old catalogues from long-departed Chicago publishers who had been young men in the eighteen-nineties trying to support a little realism." Without leaving Manhattan, Kazin read his way into "lonely small towns, prairie villages, isolated colleges, dusty law offices, national magazines, and provincial 'academies' where no one suspected that the obedient-looking young reporters, law clerks, librarians, teachers would turn out to be Willa Cather, Robert Frost, Sinclair Lewis, Wallace Stevens, Marianne Moore."

It's an old and reassuring story: bookish boy or girl enters the cool, dark library and discovers loneliness and freedom. For the past ten years or so, however, the cities of the book have been anything but quiet. The computer and the Internet have transformed reading more dramatically than any technology since the printing press, and for the past five years Google has been at work on an ambitious project, Google Book Search. Google's self-described aim is to "build a comprehensive index of all the books in the world," one that would enable readers to search the list of books it contains and to see full texts of those not covered by copyright. Google collaborates with publishers, called Google Publishing Partners—there are more than ten thousand of them around the world—to provide information about books that are still copyright protected, including text samples, to all users of the Web. A second enterprise, the Google Library Project, is digitizing as many books as possible, in collaboration with great libraries in the U.S. and abroad. Among them is Kazin's beloved New York Public Library, where more than a million books are being scanned.

Google's projects, together with rival initiatives by Microsoft and Amazon, have elicited millenarian prophecies about the possibilities of digitized knowledge and the end of the book as we know it. Last year, Kevin Kelly, the self-styled "senior maverick" of*Wired*, predicted, in a piece in the *Times*, that "all the books in the world" would "become a single liquid fabric of interconnected words and ideas." The user of the electronic library would be able to bring together "all texts—past and present, multilingual—on a particular subject," and, by doing so, gain "a clearer sense of what we as a civilization, a species, do know and don't know." Others have evoked even more utopian prospects, such as a universal archive that will contain not only all books and articles but all documents anywhere—the basis for a total history of the human race.

In fact, the Internet will not bring us a universal library, much less an encyclopedic record of human experience. None of the firms now engaged in digitization projects claim that it will create anything of the kind. The hype and rhetoric make it hard to grasp what Google and Microsoft and their partner libraries are actually doing. We have clearly reached a new point in the history of text production. On many fronts, traditional periodicals and books are making way for blogs and other electronic formats. But magazines and books still sell a lot of copies. The rush to digitize the written record is one of a number of critical moments in the long saga of our drive to accumulate, store, and retrieve information efficiently. It will result not in the infotopia that the prophets conjure up but in one in a long series of new information ecologies, all of them challenging, in which readers, writers, and producers of text have learned to survive.

As early as the third millennium B.C., Mesopotamian scribes began to catalogue the clay tablets in their collections. For ease of reference, they appended content descriptions to the edges of tablets, and they adopted systematic shelving for quick identification of related texts. The greatest and most famous of the ancient collections, the Library of Alexandria, had, in its ambitions and its methods, a good deal in common with Google's book projects. It was founded around 300 B.C. by Ptolemy I, who had inherited Alexandria, a brand-new city, from Alexander the Great. A historian with a taste for poetry, Ptolemy decided to amass a comprehensive collection of Greek works. Like Google, the library developed an efficient procedure for capturing and reproducing texts. When ships docked in Alexandria, any scrolls found on them were confiscated and taken to the library. The staff made copies for the owners and stored the originals in heaps, until they could be catalogued. At the collection's height, it contained more than half a million scrolls, a welter of information that forced librarians to develop new organizational methods. For the first time, works were shelved alphabetically.

Six hundred years later, Eusebius, a historian and bishop of the coastal city of Caesarea, in Palestine, assembled Christian writings in the local library. He also devised a system of cross-references, known as "canon tables," that enabled readers to find parallel passages in the four Gospels—a system that the scholar James O'Donnell recently described as the world's first set of hot links. A deft impresario, Eusebius mobilized a team of secretaries and scribes to produce Bibles featuring his new study aid; in the three-thirties, the emperor Constantine placed an order with Eusebius for fifty parchment codex Bibles for the churches of his new city, Constantinople. Throughout the Middle Ages, the great monastic libraries engaged in the twin projects of accumulating large holdings and, in their scriptoria, making and disseminating copies of key texts.

The rise of printing in fifteenth-century Europe transformed the work of librarians and readers. Into a world already literate and curious, the printers brought, within half a century, some twenty-eight thousand titles, and millions of individual books—many times more than the libraries of the West had previously held. Reports of new worlds, new theologies, and new ideas about the universe travelled faster and more cheaply than ever before. The entrepreneurial world of printing made much use of the traditional skills of learned librarians. Giovanni Andrea Bussi, a librarian of the papal collection of Sixtus IV, also served as adviser to two German printers in Rome, Conrad Sweynheym and Arnold Pannartz, who began printing handsome editions of classical texts, edited, corrected, and sometimes prefaced by Bussi. Like many first movers, Bussi and his partners soon found that they had overestimated the market, with disastrous financial results. They were not the last impresarios of new book technologies to experience this kind of difficulty.

Still, the model of scholars advising printers became normal in the sixteenth century, even if, in later centuries, the profit-driven industry of publishing and the industrious scholarship of the libraries gradually became separate spheres. Remarkably, this ancient model has been resurgent in recent years, as sales of university-press books have dwindled and the price of journal subscriptions has risen. With electronic publishing programs, libraries have begun to take on many of the tasks that traditionally fell to university presses, such as the distribution of doctoral dissertations and the reproduction of local book and document collections—a spread of activities that Eusebius would have found natural.

Fast, reliable methods of search and retrieval are sometimes identified as the hallmark of our information age; "Search is everything" has become a proverb. But scholars have had to deal with too much information for millennia, and in periods when information resources were multiplying especially fast they devised ingenious ways to control the floods. The Renaissance, during which the number of new texts threatened to become overwhelming, was the great age of systematic note-taking. Manuals such as Jeremias Drexel's "Goldmine"—the frontispiece of which showed a scholar taking notes opposite miners digging for literal gold—taught students how to condense and arrange the contents of literature by headings. Scholars well grounded in this regime, like Isaac Casaubon, spun tough, efficient webs of notes around the texts of their books and in their notebooks—hundreds of Casaubon's books survive—and used them to retrieve information about everything from the religion of Greek tragedy to Jewish burial practices. Jacques Cujas, a sixteenth-century legal scholar, astonished visitors to his study when he showed them the rotating barber's chair and movable bookstand that enabled him to keep many open books in view at the same time. Thomas Harrison, a seventeenth-century English inventor, devised a cabinet that he called the Ark of Studies: readers could synopsize and excerpt books and then arrange their notes by subject on a series of labelled metal hooks, somewhat in the manner of a card index. The German philosopher Leibniz obtained one of

Harrison's cabinets and used it in his research.

For less erudite souls, simpler techniques abridged the process of looking for information much as Wikipedia does now. Erasmus said that every serious student must read the entire corpus of the classics and make his own notes on them. But he also composed a magnificent reference work, the "Adages," in which he laid out and explicated thousands of pithy ancient sayings—and provided subject indexes to help readers find what they needed. For centuries, schoolboys first encountered the wisdom of the ancients in this predigested form. When Erasmus told the story of Pandora, he said that she opened not a jar, as in the original version of the story, by the Greek poet Hesiod, but a box. In every European language except Italian, Pandora's box became proverbial—a canard made ubiquitous by the power of a new information technology. Even the best search procedures depend on the databases they explore.

From the eighteenth century, countries, universities, and academies maintained research libraries, whose staff pioneered information retrieval with a variety of indexing and cataloguing systems. The development of the Dewey decimal system, in the eighteen-seventies, coincided with a democratization of reading. Cheap but durable editions like those of Bohn's Library brought books other than the Bible into working-class households, and newspapers, which in the late nineteenth century sometimes appeared every hour, made breaking news and social commentary available across all social ranks.

In the nineteen-forties, Fremont Rider, a librarian at Wesleyan University, prophesied that material was multiplying so quickly that it would soon overflow even the biggest sets of stacks. He argued that microphotography could eliminate this problem, and that, by multiplying the resources of any given library, it also offered the promise of truly universal libraries. As old universities expanded and new ones sprouted in the nineteen-fifties and sixties, generous funding enabled them to buy what was available on film or microfiche and in reprint form. Suddenly, you could do serious research on the Vatican Library's collections not only in Rome but also in St. Louis, where the Knights of Columbus assembled a vast holding of microfilm.

But the film- and reprint-based libraries never became really comprehensive. The commercial companies that did most of the filming naturally concentrated on more marketable texts, while nonprofit sponsors concentrated on the texts that mattered to them. No over-all logic determined which texts were reprinted on paper, which were filmed, and which remained in obscurity. Some efforts, like Eugene Power's STC project, which distributed on microfilm twenty-six thousand early English books, transformed research conditions in certain fields. Others simply died. As the production of books and serials exploded in the sixties, library budgets failed to keep pace. A number of reprint publishers ended up, like Bussi and his associates, drowning in unsold books. The venders of microfilm kept going—so efficiently, and enthusiastically, that they persuaded librarians and archivists to destroy large quantities of books and newspapers that could have been preserved.

The current era of digitization outstrips that of microfilm, in ambition and in achievement. Few individuals ever owned microfilm readers, after all, whereas many people have access to P.C.s with Internet connections. Now even the most traditional-minded scholar generally begins by consulting a search engine. As a cheerful editor at Cambridge University Press recently told me, "Conservatively, ninety-five per cent of all scholarly inquiries start at Google." Google's famous search algorithm emulates the principle of scholarly citation—counting up and evaluating earlier links in order to steer users toward the sources that others have already found helpful. In a sense, the system resembles nothing more than trillions of old-fashioned footnotes.

The Google Library Project has so far received mixed reviews. Google shows the reader a scanned version of the page; it is generally accurate and readable. But Google also uses optical character recognition to produce a second version, for its search engine to use, and this double process has some quirks. In a scriptorium lit by the sun, a scribe could mistakenly transcribe a "u" as an "n," or vice versa. Curiously, the computer makes the same mistake. If you enter *qualitas*—an important term in medieval philosophy—into Google Book Search, you'll find almost two thousand appearances. But if you enter "qnalitas" you'll be rewarded with more than five hundred references that you wouldn't necessarily have found. Sometimes the scanner operators miss pages, or scan them out of order. Sometimes the copy is not in good condition. The cataloguing data that identify an item are often incomplete or confusing. And the key terms that Google provides in order to characterize individual books are sometimes unintentionally comic. It's not all that helpful, when you're thinking about how to use an 1878 Baedeker guide to Paris, to be told that one of its keywords is "fauteuils."

But there are even more fundamental limitations to the Google project, and to its competitors from Microsoft and Amazon. One of the most frequently discussed difficulties is that of copyright. A conservative reckoning of the number of books ever published is thirty-two million; Google believes that there could be as many as a hundred million. It is estimated that between five and ten per cent of known books are currently in print, and twenty per cent—those produced between the beginning of print, in the fifteenth century, and 1923—are out of copyright. The rest, perhaps seventy-five per cent of all books ever printed, are "orphans," possibly still covered by copyright protections but out of print and pretty much out of mind. Google, controversially, is scanning these books although it is not yet making them fully available; Microsoft, more cautiously, is scanning only what it knows it can legitimately disseminate.

Google and Microsoft pursue their own interests, in ways that they think will generate income, and this has prompted a number of major libraries to work with the Open Content Alliance, a nonprofit book-digitizing venture. Many important books will remain untouched: Google, for example, has no immediate plans to scan books from the first couple of centuries of printing. Rare books require expensive special conditions for copying, and most of those likely to generate a lot of use have already been made available by companies like Chadwyck-Healey and Gale, which sell their collections to libraries and universities for substantial fees. Early English Books Online offers a hundred thousand titles printed between 1475 and 1700. Massive tomes in Latin and the little pamphlets that poured off the presses during the Puritan revolution—schoolbooks, Jacobean tragedies with prompters' notes, and political pamphlets by Puritan regicides—are all available to anyone in a major library.

Other sectors of the world's book production are not even catalogued and accessible on site, much less available for digitization. The materials from the poorest societies may not attract companies that rely on subscriptions or on advertising for cash flow. This is unfortunate, because these very societies have the least access to printed books and thus to their own literature and history. If you visit the Web site of the Online Computer Library Center and look at its WorldMap, you can see the numbers of books in public and academic systems around the world. Sixty million Britons have a hundred and sixteen million public-library books at their disposal, while more than 1.1 billion Indians have only thirty-six million. Poverty, in other words, is embodied in lack of print as well as in lack of food. The Internet will do much to redress this imbalance, by providing Western books for non-Western readers. What it will do for non-Western books is less clear.

A record of all history appears even more distant. If you were going to make such a record available, as the most utopian champions of digitization imagine, you would have to include both literary works and archival documents never meant for publication. It's true that millions of these documents are starting to appear on screens. The online records of the Patent and Trademark Office are a boon for anyone interested in its spectacular panorama of the brilliance and lunacy of American tinkerers. Thanks to the nonprofit Aluka archive, scholars and writers in Africa can study on the Web a growing number of African records whose originals are stored, inaccessibly, elsewhere in the world. Historians of the papacy can read original documents of the early Popes without going to Rome, in a digitized collection of documents mounted by the Vatican Secret Archives. But even the biggest of these projects is nothing more than a flare of light in the still unexplored night sky of humanity's recorded past. ArchivesUSA, a Web-based guide to American archives, lists five and a half thousand repositories and more than a hundred and sixty thousand collections of primary source material. The U.S. National Archives alone contain some nine billion items. It's not likely that we'll see the whole archives of the United States or any other developed nation online in the immediate future—much less those of poorer nations.

The supposed universal library, then, will be not a seamless mass of books, easily linked and studied together, but a patchwork of interfaces and databases, some open to anyone with a computer and WiFi, others closed to those without access or money. The real challenge now is how to chart the tectonic plates of information that are crashing into one another and then to learn to navigate the new landscapes they are creating. Over time, as more of this material emerges from copyright protection, we'll be able to learn things about our culture that we could never have known previously. Soon, the present will become overwhelmingly accessible, but a great deal of older material may never coalesce into a single database. Neither Google nor anyone else will fuse the proprietary databases of early books and the local systems created by individual archives into one accessible store of information. Though the distant past will be more available, in a technical sense, than ever before, once it is captured and preserved as a vast, disjointed mosaic it may recede ever more rapidly from our collective attention.

Still, it is hard to exaggerate what is already becoming possible month by month and what will become possible in the next few years. Google and Microsoft are flanked by other big efforts. Some are largely philanthropic, like the old standby Project Gutenberg, which provides hand-keyboarded texts of English and American classics, and the distinctive Million Book Project, founded by Raj Reddy, at Carnegie Mellon University. Reddy works with partners around the world to provide, among other things, online texts in many languages for which character-recognition software is not yet available. There are hundreds of smaller efforts in specialized fields—like Perseus, a site, based at Tufts, specializing in Greek and Latin—and new commercial enterprises like Alexander Street Press, which offers libraries beautifully produced collections of everything from *Harper's Weekly* to the letters and diaries of American immigrants. It has become impossible for ordinary scholars to keep abreast of what's available in this age of electronic abundance—though *D-Lib Magazine*, an online publication, helps by highlighting new digital sources and collections, rather as material libraries used to advertise their acquisition of a writer's papers or a

collection of books with fine bindings.

The Internet's technologies, moreover, are continually improving. Search engines like Google, Altavista, and HotBot originally informed the user about only the top layers of Web pages. To find materials buried in such deep bodies of fact and document as the Library of Congress's Web site or JSTOR, a repository of scholarly journal articles, you had to go to the site and ask a specific question. But in recent years—as anyone who regularly uses Google knows—they have become more adept at asking questions, and the search companies have apparently induced the largest proprietary sites to become more responsive. Specialist engines like Google Scholar can discriminate with astonishing precision between relevant and irrelevant, firsthand and derivative information.

Alfred Kazin loved the New York Public Library because it admitted everyone. The readers included not only presentable young scholars like his friend Richard Hofstadter but also many wild figures who haunted the reading rooms: "the little man with one slice of hair across his bald head, like General MacArthur's . . . poring with a faint smile over a large six-column Bible in Hebrew, Greek, Latin, English, French, German," and "the bony, ugly, screeching madwoman who reminded me of Maxim Gorky's 'Boless.' " Even Kazin's democratic imagination could not have envisaged the hordes of the Web's actual and potential users, many of whom will read material that would have been all but inaccessible to them a generation ago.

And yet we will still need our libraries and archives. John Seely Brown and Paul Duguid have written of the so-called "social life of information"—the form in which you encounter a text can have a huge impact on how you use it. Original documents reward us for taking the trouble to find them by telling us things that no image can. Duguid describes watching a fellow-historian systematically sniff two-hundred-and-fifty-year-old letters in an archive. By detecting the smell of vinegar—which had been sprinkled, in the eighteenth century, on letters from towns struck by cholera, in the hope of disinfecting them—he could trace the history of disease outbreaks. Historians of the book—a new and growing tribe—read books as scouts read trails. Bindings, usually custom-made in the early centuries of printing, can tell you who owned them and what level of society they belonged to. Marginal annotations, which abounded in the centuries when readers usually went through books with pen in hand, identify the often surprising messages that individuals have found as they read. Many original writers and thinkers—Martin Luther, John Adams, Samuel Taylor Coleridge—have filled their books with notes that are indispensable to understanding their thought. Thousands of forgotten men and women have covered Bibles and prayer books, recipe collections, and political pamphlets with pointing hands, underlining, and notes that give insights into which books mattered, and why. If you want to capture how a book was packaged and what it has meant to the readers who have unwrapped it, you have to look at all the copies you can find, from original manuscripts to cheap reprints. The databases include multiple copies of some titles. But they will never provide all the copies of, say, "The Wealth of Nations" and the early responses it provoked.

For now and for the foreseeable future, any serious reader will have to know how to travel down two very different roads simultaneously. No one should avoid the broad, smooth, and open road that leads through the screen. But if you want to know what one of Coleridge's annotated books or an early "Spider-Man" comic really looks and feels like, or if you just want to read one of those millions of books which are being digitized, you still have to do it the old way, and you will have to for decades to come. At the New York Public Library, the staff loves electronic media. The library has made hundreds of thousands of images from its collections accessible on the Web, but it has done so in the knowledge that its collection comprises fifty-three million items.

Sit in your local coffee shop, and your laptop can tell you a lot. If you want deeper, more local knowledge, you will have to take the narrower path that leads between the lions and up the stairs. There—as in great libraries around the world—you'll use all the new sources, the library's and those it buys from others, all the time. You'll check musicians' names and dates at Grove Music Online, read Marlowe's "Doctor Faustus" on Early English Books Online, or decipher Civil War documents on Valley of the Shadow. But these streams of data, rich as they are, will illuminate, rather than eliminate, books and prints and manuscripts that only the library can put in front of you. The narrow path still leads, as it must, to crowded public rooms where the sunlight gleams on varnished tables, and knowledge is embodied in millions of dusty, crumbling, smelly, irreplaceable documents and books. ♦

ILLUSTRATION: TOM GAULD