Chapter 4

Digitizing the Library of Congress

It is often said that the Library of Congress is on the Internet. My students, their anxious parents, journalists, educators, naked self-promoters, and my teenage daughter have made this claim so frequently that I thought I might do some small service both for beleaguered reason and under-appreciated librarians by taking it seriously. In fact, the Library of Congress is where it has always been, firmly, solidly in Washington D.C. But what would be necessary to digitize it? As it happens, quite a lot.

Before we get into that, though, let’s take a quick look at the claims:

- "The outposts on the information roadway are already well established. They constitute everything from the Library of Congress to the Center for Auto Safety." (*Information Today*, 4/23/94)

- "…Glenbrook North High School in the affluent Chicago suburb of Northbrook, Ill., … is equipped with hundreds of personal computers, … where students routinely tap into the Internet, the worldwide network linked to resources like the Library of Congress." (Tacoma *News Tribune*, 5/1/94)

- "K-12 students can research their reports at the Library of Congress from anywhere in the country." (*Electronic Learning*, 1994)

- "Then I discovered the Library of Congress’ World Wide Web site…and the remote and stuffy hardcopy depository was transformed into an accessible, dynamic, multi-dimensional electronic clearinghouse.” (*Link-Up*, 11/95)

- "If we had the information superhighways we need, a school child could plug into the Library of Congress every afternoon and explore a universe of knowledge.” (Al Gore, quoted in Richard Stoll’s *Silicon Snake Oil*, 1995)

- "So I took him aside and showed him the Vatican's online library, the Sistine Chapel home page, and how you can access the Library of Congress over the Internet." (Seattle *Post-Intelligencer*, 12/15/97)

- "But much of the technology that makes it possible to access the Library of Congress from your living room…." (*Communications of the ACM*, 10/98)

And my personal favorite, though in a slightly different category:

- "The amount of freely available text on the Web is equal to that in the Library of Congress says Steve Lawrence, a Web reporter at the NEC research institute." (*USA Today*, 7/11/ 00)

Here Mr. Lawrence dispenses with the physical Library altogether, invoking only its size. No matter what makes up this "freely available text," there's a lot of it—as much, in fact, as you'll find in the Library of Congress. It is as if the Chamber of Commerce claimed that the local Wall Mart, thrown up last summer in a highway hell of fast food, oil change, and discount furniture franchises, is "equal to" the cathedral of Santa Croce in Florence because it has the same volume.

Finally, no romp through the wonders of computing would be complete without the future tense:

- "So in 25 years, you'll be able to get the sum total of all human knowledge on a personal device." (*Barron's Online*, 11/13/00)

Though the Library of Congress is not explicitly cited, I include this quote because the Library, in its mythical gown, is a metonym for "the sum total of all human knowledge." Predictions like this[1] have been meal tickets ever since Alvin Toffler, advisor to Newt Gingrich and other visionaries, wrote *Future Shock* thirty years ago. These come from

[1] The author of the Barron's quote, one Greg Blonder, late of AT&T, is nearly off the chart. "Within ten years," he tells us, "a medical ATM … [will] constantly monitor your physical condition and … be networked with databases and your personal physician. It … [will] be able to take action if it detects an incipient heart attack or an infection getting out of control." I'll bet the docs are standing in line for this one, as if pagers weren't bad enough. For my part, if an around-the-clock virtual physical examination is the price of a long life, then let me go gentle into that good night.

just a quick survey of what happens to be easily available from my university library. Not a representative sample of printed opinion, of course. But you get the idea.

So, what's really available online? Let's begin with the Library of Congress itself. First, what used to be called the author/title card catalog is now available over the World Wide Web. This is no small undertaking. The Library owns millions of books. Unfortunately, I cannot access it this afternoon. I don't know if the world's citizens have chosen just this moment to investigate the Library of Congress, if too many students on campus are using Napster's descendents to steal from heavy metal bands right now, or if that bizarre mixture of shopping, chit-chat, vanity publication, and pornography that accounts for the bulk of Internet traffic has caused a jam somewhere between me and Washington. I could find out. But it really doesn't matter. What does matter is that the experience falls far short of the promise. Nevertheless, the catalog, at least in theory, is available to me.

Second, the Library has made many of its special collections available over the Internet. On a better day, through The American Memory Project, I could find baseball cards, northern California folk music from the thirties, and portraits of the presidents and their wives. This is all interesting, especially if you're like my eighth grade daughter, and need some filler for a report that you've put off writing until the eleventh hour. It is not, however, the Library of Congress, not even close.

In fact, card catalogs and special collections, like some of James Boswell's papers, housed in Yale's Beinecke Library[2], classic poetry, a handful of novels, and a growing

[2] At the end of an article describing this project, one of its designers says "…further assessment and discussion has clarified the library's understanding of the utility of maintaining Web access to less used digital files." In effect, it's not worth the trouble. The Italians have a nice expression for this: *Tra dire al fare c'e mezzo di mare.* Between the saying and the doing, there is half a sea.

selection of scholarly and trade journals just about do it for the digital library. I don't want to deny the usefulness of providing specialized materials like these in digital form. I am quite happy to learn from my office that a book is not available in the university library. I'm also happy to be able to do a full-text search of *Henry IV, Part I* and the poetry of Andrew Marvell. I'm happy for the many reference services, the indexes, the databases, the *Encyclopedia Britannica*, a selection of periodical literature, all now available to me over the World Wide Web. But these, were they printed, would stack neatly in the back corner of one of the Library of Congress's several buildings.   They don't begin to approximate what the Library calls "the single most comprehensive accumulation of human expression every assembled."

You may have noticed that books are missing from the roster of materials now available electronically. Though one part of this has to do with the technical and legal matters that I will discuss later, a large part is surely due to the ideology that swirls about computer enthusiasts like San Francisco fog. Several years ago, when my own university, captivated by the digital dream, built a new library, then called the Center for Information Technology, an aging and well-loved English professor was moved to ask the library director: “Yes, all these computers are nice, but where are the books?”   I learned later, to answer the professor's simple question, that the new library’s collection of books is approximately the size of the old library’s collection. The building, though, filled with all manner of electronic gadgetry, could house the old one in its west wing.

Well, to be precise, books are not completely missing from the roster. Project Gutenberg has more than a thousand volunteers around the globe encoding books from the public domain, that is, books published before the early twenties. What does Michael

Hart, its founder, have to show for this 30 year labor of love that *Publishers Weekly* calls "the Alexandria Library of the digital age?" About 2,500 titles online. The Library of Congress owns 18 million books. On the commercial end—we're not talking about a digital library here, but rather a digital bookstore—Barnes and Noble has joined Microsoft and begun *selling* digital books over the Web. Customers may download these, read them on the screen, or ask that individual copies be printed on high-speed presses. According to *The New York Times* (8/7/00), Barnes and Noble will begin by offering 2,000 titles and add about 150 per week, an effort that may well drive what remains of local booksellers out of business. Nevertheless, a couple thousand books about diet, mental health, flat abs, and the stock market, along with some current and classic fiction, do not the Library of Congress make, especially if we have to pay for them.

In fact, substantial technical and legal obstacles stand in the way of a fully digitized library, obstacles that are, for practical purposes, insurmountable given the state of the art. Let's begin with formatting issues. Nearly all books and journals these days have a digital alter ego somewhere, in some format. I am writing this essay with Microsoft Word. I will send it to the publisher in this format who will transcribe it, with much human intervention, into another digital format. Were there a concerted effort to build digital libraries, all new books and articles might enter a library collection in a standard digital format. A format is a set of codes that tells the reader's computer how the writer would like his work displayed. A standard format is simply one that is widely agreed upon.

Sounds easy enough. But in the seething world of computing, products come into existence, live their short lives, and depart with the speed and volatility of subatomic particles. Formatting languages are no exception. Standard Generalized Markup Language has emerged as a good choice for text. It has not caught the attention of Microsoft, however, which does not offer SGML as one of Word's formatting options. Microsoft Word, itself, was well on its way to becoming a de facto, though proprietary, digital standard until the Justice Department discovered that Microsoft is a monopoly. Yet another possibility is Postscript, the format developed by the Adobe Corporation for its Acrobat Reader. Still another is Hypertext Markup Language, the format used by Web browsers. There are many, many other candidates.

Even if we could settle on a standard format for current books and journals, we would be a long way from having the Library of Congress available online. The library claims to own 119 million items. The greatest bulk of these do not exist in digital form, standard or otherwise, for the simple reason that most of them were acquired long before computers existed. This means that they have to be digitized. For text, there are really only two possibilities: re-key the entire document, a fabulously expensive proposition, or mechanically scan it into digital form. Michael Lesk, one of the pioneers of the digital library, describes scanning a 900-page book. At fifteen seconds per page, it should have taken, he says, about four hours. The words "should have taken" ought to be engraved on golden tablets and mounted in the Smithsonian. Everything in computing takes longer than it should have taken. In Lesk's case, his four hours stretched into "most of the day." In the end, he had a digitized book. What he didn't have was the original whose binding had to be cut off to pass it through the mechanical feeder.

Still, if we are willing to destroy books in order to save us a trip to the library, scanning will do the trick. In 1992, Lesk says that researchers at Cornell found that they could scan 19th century books for about 40 dollars per volume. At this rate, the cost for the Library's 18 million books would be about three quarters of a billion dollars, over ten times what the Library has collected for its National Digital Library since 1996. And this is just the cost of scanning for books alone. Since we are talking about the next generation of libraries, not the next generation of heat-seeking missiles, this is a stunning sum of money.

When I was a boy, our parish priest, ominously named Fr. Fate, told us a story about the hopelessness of the damned. He asked us to imagine a steel ball the size of the earth and a bird that flew past this great ball just once a year, brushing it with its wing, ever so lightly. "When the ball is worn to nothing,"--he paused here, letting us imagine the uncountable centuries required--"eternity has just begun." Like the damned, having spent onto a billion dollars scanning the Library's entire collection of books, we have miles to go before we sleep.

Scanning produces a digital image where the many shades of light and dark that constitute a printed page are represented, in effect, as ones and zeroes. The more closely these ones and zeroes are packed together, that is, the higher the resolution, the more accurately the scanned image resembles the original. Resolution is represented in dots per inch, dpi to the initiated. Fax images have a resolution of 200 dpi, while laser printers print at 300 dpi. For technical reasons, scanning an image is more difficult by a factor of two than printing it. So to produce a scanned image of laser printer quality requires a resolution of 600 dpi. There are three reasons why we care about this.

First, the higher the resolution, the more storage that is required and, though disk storage is cheap, it's neither free nor infinite. A scanned image requires lots more storage space, in fact, several orders of magnitude more, than a simple encoding of text. The reason is simple. A word, say in extended ASCII format, is a sequence of eight 1s and 0s, or bits, per letter. A scanned image, on the other hand, is something like a photograph. The letters are not encoded but, rather, represented as shades of light and dark. As an experiment, I transformed the previous page into ASCII and found that it requires 14,480 bits of storage. At 8 bits to a byte, the standard unit of storage in a computer, and 1024 bytes to a KB, read "kilobyte," that page requires 1.76 KB. But when I scan it at 600 dpi, though, the resulting file is 2,490 KB in size.

Second, the size of the digital image is directly related to how quickly that image can be transmitted over the Web. Right now, my 56K modem will, in theory, download 57,344 bits per second. Assuming that there is no network overhead, an incorrect assumption, the previous page in ASCII format could be downloaded in less than a second. The scanned page, however, would require 356 seconds. Michael Lesk, in his book *Practical Digital Libraries,* provides another example. The three letters "CAT" require 3 bytes. A line drawing of a cat requires around a kilobyte. A scanned photograph of a cat, however, requires 13 KB of storage, about 4,500 times as much as the word. This is exactly why a flashy Web page with lots of complicated graphics at the top composes much more slowly on your computer than a boring one that relies heavily on text.

Third, and perhaps, most importantly, you cannot do a full-text search of a scanned image, even if the image is of text. Recently, I wanted to find where in Shakespeare these lines occur:

> Some are born great, some achieve greatness, and some
> have greatness thrust upon 'em.

Because the collected works are digitized as text, I had no trouble discovering that these lines are spoken by Malvolio in *The Twelfth Night.* Had Shakespeare's collected works merely been scanned and placed online, I would have been forced to use a concordance, the kind of low-tech index that has been around for generations. Since full-text searching is one of the strongest arguments for a digital library, we have somehow to move from a scanned to a text-based image.

Clever programmers have been hard at work on this problem. After digitizing a document, we run it past another piece of software that uses a technique called optical character recognition. OCR software is usually included with even inexpensive scanners. The good news is that OCR is surprisingly accurate. Good software can correctly extract characters from scanned images 99 percent of the time. Suppose a book you are scanning has about 15 five-letter words per 45 line page. Ninety-nine percent accuracy would produce about three dozen errors per page that must be corrected by a real, live person. The bad news is that this is expensive. The Journal Storage Project of the Andrew W. Mellon Foundation has found that it can scan, optically read, and correct economics journals at about 39 cents per page. This triples the Cornell price of 40 dollars per 300 page volume. The cost of digitizing the Library's collection of books has now jumped to over two billion dollars. As Senator Dirksen of Illinois once said, "A billion here, a billion there, and pretty soon you're talking about real money."   The Library's total budget for 1999 was 620 million dollars.

But we're still not finished. Digital media have surprisingly short lifetimes. According to Jeff Rothenberg in a report for the Council on Library and Information Resources, testing done at the National Media Lab indicates that a randomly chosen tape, disk, or CD "is unlikely to have a lifetime of even five years."   Before you start arguing that your copy of *The White Album* still plays 10 years after you bought it, think about what we are discussing here. Even if this number is off by a factor of 10, it still means that in 50 years the Library's entire collection, digital copies of all 119 million items, will have to be transferred to new media.

We've jumped from the pot to the fire this time. One of the arguments for digitizing entire collections is that they are disappearing even as we dither. This is the so-called "brittle paper" problem. Lisa Fox, in her book, *Preservation Microfilming*, claims that inexpensive paper, mass-produced from wood pulp since the mid-nineteenth century, disintegrates so quickly that one-third of all research library materials are unusable, and 80 million books in North American Libraries are threatened with destruction. The situation is even worse for newspapers. Though one might question these numbers (as the *New Yorker*'s Nicholson Baker does in a recent book) the sad fact is that paper, however fragile, seems to last longer than digital media.

But the physical evanescence of digital media is only the beginning of the nightmare. Have you noticed that computers these days can't read 5.25 inch disks? If you did not have the foresight to transfer your book, your dissertation, your tax records from 5.25 inch to the current 3.50 inch diskettes, well, you're out of luck. These diskettes lived, prospered, and died a quiet death, all in only 15 years. The situation is much aggravated with software. However, we choose to digitize our libraries, we have to decide upon a piece of software that will both store and allow us to access our collections. The trouble is that software lives about as long as children's fads. Once it's dead, it's dead. You'll get your eight-year old to bond with an old Cabbage Patch doll well before you'll get Microsoft Word to read a WordStar file.

The most obvious solution to this dilemma has some obvious dilemmas of its own. Just keep a copy of the software. But this means keeping a copy of the operating system on which the software runs, which also means keeping the computer on which the operating system runs. These, of course, become obsolete quickly, too, which is why you have to run off to the computer store every couple of years. All of this is compounded with large computers that require considerable expertise to administer. We may well keep our old computers around, a fabulous proposition, but our technicians will have long since moved on. The situation is so bad that Jeff Rothenberg has proposed the original, clever, and utterly baroque idea of encapsulating with the digital document, the software used to create it, the operating system on which the software runs, and a precise description of the current computer. Librarians of the future will use this description to build new software to run on computers not yet imagined that will act like the computer on which the operating system runs. If this sounds complicated, that's because it is. Let's hope our librarians don't toss the books before they get the kinks worked out.

I have not even mentioned cataloguing and, above all, copyright issues. We need to be able to retrieve items quickly, after all, and we need to do it legally. Ever since Anthony Danizzi catalogued the British Museum's holdings in 1831, cataloguers, those unsung and underpaid backroom heroes, have made the modern research library possible. Without them, the Library's treasures would look like a grand version of my daughter's closet. But cataloguing is labor-intensive work. Caroline Arms, in an article in *D-Lib Magazine,* cites one sobering estimate of the time required to catalog the University of California at Berkeley's 3.5 million photographic images: 400 years, even with the entire cataloguing staff press-ganged into the effort. As for copyright, it is true that much of the Library's holdings are already in the public domain, though much is not, including every book on computing itself. Even if we overcome technical obstacles, a very large, very old-fashioned obstacle remains--writers and publishers expect to be paid for their work. They will not be happy when libraries freely distribute it over the Web.

Please let me clarify. I am not saying that a fully digitized research library is unappealing. I don't have a sentimental attachment to the feel of the card catalog or the smell of paper, even to the heft of a book that I'm only using for information. I would be the first to sign up if I could access the entire collection of a major research library, and I could do it without waiting half an afternoon for material to download. Nor am I saying that digitizing a library is technically, or even legally, impossible. I cannot, after all, foresee the future. (A modest admission, but one rarely found among the computer literati.) Maybe some free-spending, book-loving Congress will appropriate money to digitize the Library's entire collection. Perhaps storing digital documents in a time capsule, along with digitizing software, an operating system, and instructions on how to build a virtual computer to run on any possible future computer is not as far-fetched as it seems. Maybe Microsoft will invent a scheme to compensate copyright holders, and library users, content in their digitized world, will not rebel. Maybe writers will give up their royalties for the good of humankind, and publishers, like the state in another imagined future, will simply wither away. Unlikely propositions all, though certainly possible. But I do claim this. The entire collection of the Library of Congress is not available over the World Wide Web. Not even a small part is available. Nor is the collection of any other library. Nor will they be available anytime soon. You can find wonderful things on the Web, along with a lot of nonsense, of course. What you cannot find is "the single most comprehensive accumulation of human expression every assembled," or even a more modest accumulation. You'll have to go to the library for that.

References and Further Reading

Many of the articles and books listed below are either breathlessly enthusiastic about the digital promise or aimed at a specialized audience. I've annotated those that I found most interesting.

Analytical Perspectives: Budget of the U.S. Government, Fiscal Year 2001. Washington D.C.: U.S. Government Printing Office. 2000
My source for the Library of Congress's budget.

Apple, R.W. "Library of Congress is an Internet Hit." *The New York Times*. 16 Feb. 1997

Arms, Caroline. "Historical Collections for the National Digital Library." *D-Lib Magazine*. April 1996

Baker, Nicholson. "A Reporter at Large." *The New Yorker.* 24 July, 2000
A detailed and spirited polemic against the practice of microfilming, then discarding, newspaper collections.

Baker, Nicholson. *Double Fold: Libraries and the Assault on Paper*. New York: Vintage Books. 2002

Balas, Janet. "Bringing Library Collections Online." *Computers in Libraries,* 10 Oct. 1999 pp. 46-48.

Borgman, Christine. "What are Digital Libraries?" *Information Processing and Management,* 35, 1999 pp. 227-243.

Bouche, Nicole. *Digitization for Scholarly Use: The Boswell Papers Project at The Beinecke Rare Book and Manuscript Library.* Washington, D.C.: Council on Library and Information Resources. 1999
An interesting description of what happens when a library begins to digitize a rare book collection.

Brown, Lonny. "The Library of Congress at Your Fingertips." *Link-up,* 12 (6), Nov/Dec. 1995 pp. 21 ff.

Cassidy, Gerald. "Public Policy and the Internet." *PR News* 55 (41). 18 Oct 1999

Chapman, S., Kenney, A. "Digital Conversion of Research Library Materials." *D-Lib Magazine*. Oct. 1996

Entlich, R., Garson, L., Lesk, M., Normore, L., Olsen, J. and Weibel, S. "Making a Digital Library." *D-Lib Magazine*. Dec. 1995

Erskine, Lynn. "Cornell Launches Data Preservation Project." *Humanities* 21(1), Jan/Feb 2000 pp. 50 ff.

Fox, Lisa, ed. *Preservation Microfilming: A Guide for Librarians and Archivists*. Chicago: American Library Association. 1996
Explains, among many other things, the brittle paper problem.

Gladney, H., Mintzer, F. Schiattarella, F. Bescos, J. and Treu, M. "Digital Access to Antiquities." *Communications of the Association for Computing Machinery,* 41(4), April 1998 pp. 49-57.
A beautifully illustrated article that describes possibilities and problems when digitizing historical sources.

Guenther, Kim. "Designing and Managing Your Digital Library." *Computers in Libraries* , 20(1), Jan. 2000 pp. 34-39.

Kirkpatrick, David. "Online Superstore Pushes into Digital." *The New York Times*. 7 Aug 2000

Lancaster, F. W. "Second Thoughts on the Paperless Society." *Library Journal* 124(15), pp. 48-50. 15 Sep. 1999
A retired professor of library science recounts and recants his earlier enthusiasms for a paperless society.

Lesk, Michael. The Organization of Digital Libraries. *Science and Technology Libraries.* 17(3,4), 1999 pp. 9-25.

Lesk, Michael. *Practical Digital Libraries*. San Francisco: Morgan Kaufmann Publishers. 1997
The best single volume on digital libraries by one of the field's pioneers.

Lesk, Michael. "Going Digital." *Scientific American*, Mar. 1997 pp. 58-60.
A good summary of the promises and problems of digital libraries.

Levy, D., Marshall, C. "Going Digital: A Look at Assumptions Underlying Digital Libraries." *Communications of the Association for Computing Machinery,* 38(4), April 1995 pp. 77-84.

Lohr, Steve. "School Lessons Along the Information Highway." Tacoma, WA: *The News Tribune*. 1 May 1994

Lynch, Clifford. “From Automation to Transformation: Forty Years of Libraries and Information Technology in Higher Education.” *Educause Review,* 35(1), Jan/Feb 2000 pp. 60-68.

Manes, Stephen. “Time and Technology Threatens Digital Archives.” *The New York Times*. 7 April 1998
A concise and readable account of digital deterioration.

Nelson, Milo. “Putting Up Overnight at the Information Highway Motel.” *Information Today,* 11(4), April 1994 pp. 23 ff.

Roth, Katherine.“Libraries Consider their Future in Today’s Digital Age.” Vancouver, B.C: *Columbian*. 9 June 1999

Rothenberg, Jeff. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Washington, D.C.: Council on Library and Information Resources. 1998
A good account of the technical issues that digitizers face along with an clever, if baroque, solution.

Salvador, Roberta. “Information System from Library of Congress.” *Electronic Learning,*13(8). 10 May 1994

Simmons, Barbara. “Outlawing Technology.” *Communications of the Association for Computing Machinery,* 41(10), Oct. 1998 pp. 17-18.

Smith, Abby.. “Preservation in the Digital Age.” *American Libraries,* 30(3), March 1999 pp. 36-39.

Starr, Paul. “The Electronic Commons.” *The American Prospect*, 27 Mar. – 10 April 2000 pp. 30-34.
A discussion of some of the problems faced by current libraries and how building a “Global Public Library” might alleviate them if issues of intellectual property are resolved.

Stepanek, Marcia. “Congress Learns What Really is on the Internet.” *Seattle Post-Intelligencer*. 15 Dec 1997

Stoll, Clifford. *Silicon Snake Oil*. NY: Doubleday. 1995
This extravagant non-book is fun to read and a healthy correction to digital boosterism. A chapter on digital libraries makes some of the arguments made here--but in inimitable Stoll form.

Tennant, Roy. “Copyright and Intellectual Property Rights.” *Library Journal,* 124(13), Aug. 1999 pp. 34-36.
Advice on copyright law to those seeking to place materials on-line.

Weller, Sam. “Project Gutenberg Sets 10,000 Book Goal.” *Publisher’s Weekly,* 3 Apr. 2000 pp. 42 ff.

Winter, Ken. “From Wood Pulp to the Web: The Online Evolution.” *American Libraries,* 31(5), May 2000, pp. 70-74.

Witten, I., Nevill-Manning, C. McNab, R., Cunningham, Sally. “A Public Library Based on Full-Text Retrieval.” *Communications of the Association for Computing Machinery,* 41(4), April 1998 pp. 71-75.
Describes features of the New Zealand Digital Library, the most fascinating of which is a melody index of 9,400 international folk tunes that “retrieves music on the basis of notes that are sung, hummed or played.” Unfortunately, the article does not provide a web site where the curious reader can try it out.