by Jeffreys Copeland and Haemer

PAUL STODDARD

"[T]ake no riffe raffe bookes for such would prove a discredit to our Librarie." – Thomas Bodley in a letter to Thomas James, as James was organizing the Bodleian Library at Oxford University

A Short History of Reading

Consider the book. It is an object of remarkable ubiquity that follows us through our lives. In our houses, we have examples ranging from thick pasteboard with the teeth marks of infants and dogs to sympathy books from family members' funerals. They have been with us, in one form or another, since the dawn of civilization. But why are they in this form now? And what form will they likely take in the future?

Parchment to Paper

"Littera scripta manet" – The written word remains. And it has. There are examples of ancient manuscripts dating back thousands of years, the most famous of which are the Dead Sea Scrolls. (Aside: How are these ancient scrolls connected to modern rock music? Through Miles–no relation–Copeland, father of rock impresario Ian and drummer Stewart, who was the CIA Station Chief in Beirut during the '50s. When the Dead Sea Scrolls were found, he was the first person to whom they were offered for sale.) Some of those ancient examples didn't survive because they were printed on an invention from the Nile river valley, paper, which was not durable; others failed to survive because they were housed in the ancient library at Alexandria, whose destruction was the most vile act of terrorism to occur before the current century.

But paper was vital. Because pre-Renaissance Europe didn't have the technology to produce it, books were written on parchment, which was much more difficult to manufacture. On the other hand, Islam had paper, which allowed the Moorish library at Cordova to house 400,000 volumes in 1085, while Europe's knowledge was still hidden in monasteries during the plague.

Through the Middle Ages, books were created by skilled scribes, who painstakingly copied the text at the rate of one or two books per year. But by the mid-1400s, craftsmen in Northern Europe had independently invented a trick that had been known in Korea and China for some time. By making molds of letters, lining them up, inking them and pressing paper against them, Johannes Gutenberg invented printing with movable type. The type used by Gutenberg was designed to emulate the handwriting of scribes in the monasteries. For the next 500 years, the purveyors of Western knowledge toiled at places like the Clarendon Press in Oxford, England.

By 1469, the printing press reached Venice, and by the end of the century, Aldus Manutius was printing large editions there. In those days, we hasten to add, "large edition" meant more than 200 books. Again, the absence of paper prevented larger press runs. But what made Aldus a success was type cut for him by Francesco Griffo. Because the workload in the papal chancery at the Vatican had become so heavy, the scribes developed a cursive style of handwriting that allowed them to write quickly and to pack more words per piece of parchment. Griffo's design was based on this style of handwriting and became known as italic type.

It allowed Aldus to print on smaller pages and, as a result, he began producing octavo editions–books containing sheets that were folded into eighths before cutting–which fit conveniently into saddle bags and, thus, became easily portable to other cities in Europe.

In 1603, the Bodleian Library opened in Oxford, endowed by King James I with income from land in Berkshire and London. (King James, of course, later commissioned the English translation of the Bible that bears his name.) The Bodleian Library was open six hours a day, six days a week and was one of the first libraries with a catalog of its holdings.

By the time of the American Revolution, paper was being mass-produced in places like Italy and Holland, and printing presses had become so numerous that books were common–if not as widespread as today. Thomas Jefferson owned the largest library in the American colonies, which housed 6,000 volumes in eight languages, including Russian. Jefferson's library became the foundation for the Library of Congress after the initial library was destroyed during the War of 1812.

But the spread of knowledge was continuing. From our childhood, we remember a drawing in a Time Life book about the volume of scientific knowledge. In 1750, there were about 10 scientific journals in the world, represented in this drawing by a scientist in a tricorn hat holding a stack of paper under his arm; by 1830, the number had increased to about 100, represented by a waist-high stack next to the scientist; by 1970, the number had increased to 10,000–of which 300 were abstract journals–illustrated by a pile towering over the heads of the other two scientists. Douglas Copland (also no relation), in his novel *Microserfs*, echoes this growth by having one of his characters observe:

"*We've reached a critical mass point where the amount of memory we have externalized in books and databases (to name but a few sources) now exceeds the amount of memory contained within our collective biological bodies. In other words, there's more memory 'out there' than exists inside 'all of us.' We've peripheralized our essence.*"

We don't know if this is actually true, but it sounds plausible.

By the 20th century, the technology for producing print media had undergone a change. The five centuries from 1450 to 1950 were the renaissance of movable type. At the end of the 19th century, the invention of the Linotype and Monotype machines allowed movable type to be set mechanically. By the middle of the 20th century, phototypesetting became practical and began to take over. In the early 1960s, computer-driven phototypesetting began to appear at newspapers and large publishing houses, driven by mainframes. The advent of the cheap laser printer coupled with the cheap personal computer in the 1980s effectively killed movable type as a production medium. (There are some pockets of resistance: The county newspaper for Saguache in Colorado's San Luis Valley still uses a Linotype machine.)

The Book or the Text?

In his book of essays *Being Digital*, Nicholas Negroponte repeatedly explores the difference between physical artifacts and their computer representation. His mantra is "bits or atoms?" When Negroponte places a value of $2 million on his laptop computer, does it reflect the value of the physical object, which cost about $2,000, or the value he places on the data stored on its hard disk? Is it the atoms of the book or the data contained within it that are actually important?

While we consider some books to be important as physical objects–there is the Robert Frost first edition with a dust jacket by Alan Haemer on one of our shelves and a century-old edition of *Alice's Adventures in Wonderland* on another–we have to agree with Professor Trefusis, whose amanuensis, Stephen Fry, captured him explaining in *The Liar*:

"*Books are not holy relics. Words may be my religion, but when it comes to worship, I am very low church. The temples and the graven images are of no interest to me. …The world is fond of saying that books should be 'treated with respect.' But when are we told that words should be*

treated with respect? From our earliest years we are taught to revere only the outward and visible."

Clearly, it's the contents of a book that are important, which is one of Ray Bradbury's lessons in *Fahrenheit 451*. We need the data, not the object. What does this mean for us?

Appearing on an Internet Near You

Because the bits are more important than the atoms of a book, we've already started seeing some interesting developments. One of the first was Project Gutenberg, based at Illinois Benedictine University (`http://www.gutenberg.net`), which has been working since 1971 to make great texts that are in the public domain available online. We carry around copies of various texts from Project Gutenberg to read on the road. This month, it's *Moby Dick*. In a similar vein, Alexandria Digital Literature (`http://www.alexlit.com`) is a commercial project to bring modern fiction, mostly science fiction, to the Internet. (Other ventures in the online world are literary, but have nothing to do with digital literature: The ever-present Amazon.com and Barnes and Noble online bookstores are examples.)

Indeed, there appears to be a small movement by authors who aren't interested in dealing with large publishing houses to put their work up on the Web. Libertarian science-fiction writer J. Neil Schulman has done so (`http://www.pulpless.com`). The updated edition of Bruce Sterling's *The Hacker Crackdown* is available in electronic, rather than printed form (`http://lonestar.texas.net/~dub/sterling.html`). Peter McWilliams' *Ain't Nobody's Business If You Do* is available in both print and Web editions (`http://www.mcwilliams.com`).

This sort of samizdat is important and has a long tradition. Indeed, we'd argue that not only does information want to be free, but that political freedom is a product of information freedom, which is how the Founding Fathers understood it.

Furthermore, self-publishing has interesting implications for edition creep. As far back as 1985, Marsh Heinrichs at Addison-Wesley Publishing Co. expressed concern about editorial review and the lack of clear edition markings when scholarly publishing began to appear on the Internet.

Online publishing also has interesting implications for copyright law. Only now, and in fits and starts, are legislators starting to come to grips with differing needs for the protection of digital representations of intellectual property. How do we protect online literature and scholarly work? Certainly, it is not feasible for all of it to be protected (or unprotected) under terms similar to that of the GNU software license. It is not acceptable to wait for all work to fall into the public domain either–certainly in the fields of science and technology, by the time 50 years have passed the content is important more for historical purposes. Unfortunately, the U.S. Congress seems more concerned with extending intellectual property protection for Mickey Mouse than worrying about copyright in cyberspace.

All concerns about content aside, we as software people want to have some knowledge of what form that content will take. There are a number of ways we could present our online books and each has its shortcomings.

Flat ASCII is the obvious choice (it's certainly the form used by Project Gutenberg) except that it doesn't allow for illustrations, or easily handle different-width displays or reformatting on the fly.

We could adopt a proprietary format, like the ones used by the what-you-see-is-what-you-get word processing programs from Microsoft Corp. and Corel Corp., except you've already heard our rants about closed formats that require you to have commercial software. (Indeed, Haemer's response to receiving such data is to silently return it to the sender using `procmail`.)

PostScript has several advantages. It's actually a page description format, rather than claiming to be a markup language, like the word-processing tools

(attempt). It also has cheap (or free) interpreters available, such as GhostScript. Furthermore, given the electronic text in PostScript, the publisher's intentions about rendering into print are very clear. But PostScript is really a printing standard, not a display one.

A step beyond PostScript is Adobe Systems Inc.'s Portable Document Format (PDF), supported by the Acrobat Reader and other freeware readers. It is more compact and is intended to be read online, while supporting all the features of PostScript. However, we keep seeing examples of PDF that are nothing more than scanned images of printed pages, which seems to defeat the purpose of the technology.

On the other hand, the original open-standard format for the reading of formatted text online is HTML, which we use every day when we retrieve data from the Web. We have the advantage of being able to change HTML formatting on the fly when we render it in a window of a different size, for example. But because the reader (or browser in this case) is tunable, it's possible the publisher's (or author's) intentions on how the information is to be displayed will be ignored. Worse, because HTML is a lightweight markup language, there are no standards for displaying equations and little support for layout more complicated than running text with interspersed pictures. Designer David Siegel suggests a raft of workarounds to these and other problems on his Web site (`http://www.dsiegel.com`). Follow-ons, like eXtensible Markup Language (XML), also address some of HTML's limitations.

Last, there are the open, intermediate formats from our favorite text-processing programs such as typesetter-independent `troff`'s intermediate form (also used by `groff`) and TeX's device-independent (DVI) file format. But using any of these for an online reading format requires some interpretation and font standardization.

The overall problem with each of these is annotation. None of them easily supports bookmarks, marginal notes, dog-eared corners or little yellow sticky notes on the pages.

The problem is worse than just format. Portability becomes an issue. Think of all the different things you read. This magazine. The newspaper. The Web. The latest Tom Clancy novel. A Supreme Court decision. Bruce Schneier's *Applied Cryptography*. The Camel book. *Hamlet* via Project Gutenberg. *Hamlet* in book form. The `sendmail` documentation. Your email. A textbook. And that's not even counting the stuff you could have "read" through other means, like listening to Jon Krakauer's *Into Thin Air* on tape in your car. It would be convenient if they could all be read using the same tools and if those tools could change mode, so you could read a book, or it could read itself to you, or you could just look at the pictures. (A technical audio book sounds like an oxymoron, but T.V. Raman, a blind computer science student, completed his Ph.D. dissertation at Cornell University four years ago in which he explored techniques for having a voice synthesizer read mathematics from technical text.)

The existence of any of the electronic forms for reading books would change the economics of publishing. It would be possible to have a large backlist because there would be no tax consequences. Indeed, nothing would ever go out of print.

However we choose to render our online books, we need to heed the observation of Stephen Walli, vice president of research and development at Softway Systems Inc. (who has the thankless task of being Copeland's boss): "Every useful application outlives the platform on which it was originally developed and deployed." In other words, any form we choose is going to be one we'll be stuck with for quite a while.

A useful statistic at this juncture, so that you can understand where the marketing effort might be spent: According to *Fotune* magazine, textbooks account for roughly one quarter of the total book sales in the United States–low-volume, high-cost specialty books, for the most part–and audio books account for one tenth. We wish we could ignore the huge collection of celebrity gossip we see at our local bookstore and the appearance of books of comic strip cats on *The New York Times* best-seller list, but those seem to be a staple of the bookselling industry.

The existence of any of the electronic forms for reading books would change the economics of publishing. It would be possible to have a large backlist because there would be no tax consequences, that is, there would be no inventory of physical books at the warehouse to be taxed. Indeed, nothing would ever go out of print. The new economics of publishing would also make it possible for there to be midlist authors again, a nice feature of publishing that disappeared about the time Messrs. Harcourt and Brace stopped running the firm that bears their names. Similarly, it would mean that the quarter of all books represented by textbooks would not weigh quite so much on our shelf or in our children's backpacks.

Nonetheless, there's a more serious problem. How do we actually read text in any of these forms?

Software Requires Hardware

We've come the long way 'round to briefly talk about the Dynabook, the notional computer developed by Alan Kay at Xerox PARC to provide a self-contained, personal database of books and information about the size of a three-ring notebook. In the Dynabook, the bits are the thing. Books can be read at the same resolution in print as on the high-resolution screen. New data and

annotations can be added using a keyboard. While it's not exactly Kay's vision, the laptop computer has filled the niche for personal data retrieval he had in mind.

But Moore's Law, which tells us that computers will cost half as much in 18 months, marches on. And we can use the ideas of science fiction to make some guesses about where technology can take us.

Certainly, Arthur C. Clarke's novel *Imperial Earth* is as much about his notions for a global computer and communications system as it is about politics or exploration. Most of the features of Clarke's pocket-size user access device can be found in off-the-shelf personal digital assistants (PDAs) today. And many of those PDAs have add-on software to load and read free text.

Similarly, Neil Stephenson's *The Diamond Age* features wafer-thin touchscreen displays. Stephenson's displays are being prototyped by an MIT Media Lab spin-off, which hasn't yet gotten them to the resolution we need for them to display the daily newspaper.

Douglas Adams' famous contender for an electronic book is *The Hitchhiker's Guide to the Galaxy*, which is more commercially successful than the *Encyclopedia Galactica* not only because it fits in your pocket, but because it has the words "Don't Panic" emblazoned on the cover in big, friendly letters.

(On the retro side, we have Isaac Asimov's send-up of *The Double Helix* in his short story, "The Holmes-Ginsbook Device," in which he postulates the invention of paper folded into piles, which can then be read, so that the hero won't miss pretty girls walking past him in the library because he's got his head stuck in a microfilm reader. The formal name of the invention is given in the title, but it's often simply referred to as the "book.")

In the present, we've written some software to emulate the process we want to use for reading text on the screen (more about that later). But, given the prevalence of the technology used in laptop computers, the full problem appears to be on the road to a solution with commercial products expected by the end of the year.

Commercial Reader Hardware

We must confess that Liz Copeland (finally, a relation–Jeff's wife) had the idea for stand-alone reader hardware about two years ago. We couldn't see how the publishing deals would work out, nor did we understand how the economies of scale could allow the finished boxes to be sold at a reasonable price. So, now we owe Liz a couple million dollars and an apology, because by the time you read this, there should be special-purpose electronic book hardware available. And if you're wondering what to get us for Christmas…

The main offerings are the Rocket eBook by NuvoMedia Inc., Palo Alto, CA (`http://www.rocketbook.com`), and the SoftBook System from SoftBook Press, Menlo Park, CA (`http://www.softbook.com`). Both offerings provide storage for books, a high-resolution LCD screen, connection to the outside world and a source of text to read.

The Rocket eBook is the lighter of the two, weighing in at 20 ounces, or the weight and size of a 200-page trade paperback. It claims to have a 20-hour battery life. You can browse an electronic bookstore on the Web and download the text through your PC into the Rocket eBook using the serial cable connected to its recharging cradle.

The SoftBook is a bit larger, with a bigger screen and weighs in at just under three pounds, featuring a 9.5-inch screen and a five-hour battery life. The additional weight results from being self-contained, with a built-in modem and firmware to connect directly to the SoftBook Web site.

Both products have touch screens. Both companies have methods of downloading arbitrary text to their hardware, so you can load documentation or text from Project Gutenberg with equal ease. Both boxes store large amounts of text–4,000 pages, in the case of the Rocket eBook.

It's unclear how online distribution of books from major publishers will shake out. SoftBook intends to act as a bookstore for its publishing partners, which includes Random House Inc. and Simon & Schuster. NuvoMedia will act as a book distributor; its investors include German publishing giant Bertlesmann, which also owns Random House and Barnes and Noble Inc.

Clearly these boxes fill an interesting niche: It's much easier to curl up in bed with a box that weighs half a kilogram and is the size of a paperback than with a laptop computer. *Fortune* magazine quotes Michael Hart, the founder of Project Gutenberg, as saying, "The thing I really want is the paperback-sized text reader you can buy at Kmart for $20." He's exactly right. Most people don't want to carry a laptop computer around with them. Once the reader hardware is a commodity product, we will finally have truly ubiquitous electronic text. The good news is that day appears to be nearing.

Finishing Up

We'd like to thank our friends Steve Hughes, recently retired from Nova Financial Services, and Gary Brown on the editorial staff of the *Palm Beach Post*, who started us thinking a little more seriously about the notion of online reading. Also a tip of the hat to Guy Lillian, an attorney in New Orleans who took the Luddite view in favor of paper and made us rethink some assumptions.

Next time, we'll show you the reader for ASCII text that we originally developed to allow us to read back issues of the RISKS Digest during boring parts of standards meetings.

Until then, happy trails.

Jeffrey Copeland *(*`copeland@alumni.caltech.edu`*) lives in Boulder, CO, and works at Softway Systems Inc. on UNIX internationalization. He spends his spare time rearing children, raising cats and being a thorn in the side of his local school board.*

Jeffrey S. Haemer *(*`jsh@usenix.org`*) works at QMS Inc. in Boulder, CO, building laser printer firmware. Before he worked for QMS, he operated his own consulting firm, and did a lot of other things, like everyone else in the software industry.*

Note: The software from this and past Work columns is available at `http://alumni.caltech.edu/~copeland/work`.