Published in *Scholarly Publishing*, 21 (1989), 11-26.
Author's address: daniel.eisenberg@bigfoot.com

Problems of the Paperless Book

Daniel Eisenberg

We are, with electronic publication, approximately where printing was in the year 1469.[1] The foundations have been established; the technology is spreading from central to peripheral areas; only a minority of holdouts remain completely opposed. There is a sense that great changes are coming.

The advantages of digital publication seem obvious. Production time and cost should decrease substantially. There will be no printer or typesetter, no inventory or warehouse. Neither will there be paper to yellow nor bindings to break. Texts, copied as needed, will never go out of print. They will be electronically searchable and otherwise manipulable by the reader, and research on texts will become easier and better.[2] Yet the promised land of computer publication remains surprisingly and frustratingly distant. Paper consumption continues to rise, and typesetting is anything but obsolete. This article will examine the reasons for this state of affairs, and propose solutions.

---

[1] For an introduction to the early years of print and their many parallels to our present situation with electronic media, see Chapter 1 of Martin Lowry's *The World of Aldus Manutius. Business and Scholarship in Renaissance Venice* ([Ithaca]: Cornell University Press, 1979).

[2] "There is clear evidence of an increase in both research quantity and quality in Classics since the advent of the *Thesaurus Linguae Graecae* data bank" (Theodore F. Brunner, "Data banks for the humanities: Learning from *Thesaurus Linguae Graecae*," *Scholarly Communication*, Number 7 [Winter, 1987], pp. 1, 6-9, at p. 9).

The type of digital publication studied is the scholarly edition. Scholarly editions are expensive to typeset yet have limited sales potential; most scholars needing scholarly texts are already familiar with computers and have access to them. Neither is true of mass-market books and their readers; scholarly publishers must be the trailblazers. Scholarly editing is also the field in which the philosophical problems of publishing are examined. If one has worked out the problems involved in an edition, one can extract from it the procedures to do any type of publication. Indeed, every publication can be thought of as an edition.

I. Distribution

The form in which the text is transmitted to the reader has important implications. Digital publication envisions reading of the text on some type of display screen, with paper copies of the text only produced if needed. The device which transfers an electronic text from its storage medium and displays it on a screen is a computer.

Digital publication of texts must be compatible with the installed base of personal computers which are now available to most scholars. Of course the computers can and will be improved, but change will be gradual. The investment in, existing software for, and familiarity with this installed base of machines make it unlikely that specialized hardware (such as a "bookcard reader" already on the market) can be successful for any but very limited applications. Furthermore, a personal computer is, with no or minor modifications, suitable for reading texts. In the following I am assuming that the computer used for reading texts contains or is connected to a hard disk or similar large-capacity storage device; some features will require a programmable video board, such as the widely-available Enhanced Graphics Adaptor or EGA. A program to facilitate control and improve the display of the text will be discussed later.

Discussions of distribution of electronic texts have so far have focused on two media: CD-ROMS and online distribu-

tion. Both of these have serious shortcomings.

The CD-ROM, a device of huge storage capacity adapted from the data-intensive field of music, has captured wide attention because it offers such a dramatic decline in the cost per byte of data capacity. The disparity between the reduced production costs and the very high prices of existing CD-ROM publications has been instructive about the value of data.

The CD-ROM is actually too capacious a medium for texts. The works of any author would only take up a small portion of one. The creation and purification of literary data in the quantity to fill up a single CD-ROM is a major enterprise, requiring the efforts of many scholars over a period of years. Most uses of CD-ROMS to date have been for reference books or other large collections of data already in print. As individuals have so far had little reason to purchase CD-ROM readers for their computers, they are found primarily in libraries and business settings.

An additional problem with CD-ROMS is that they require manufacturing at a specialized plant (and manufacturing capacity is still scarce). While the cost per disc is small, under ten dollars, the cost for the first unit is several thousand dollars. Thus one loses of the potential advantages of electronic publication, the "printing" of copies as needed.

A second possibility is the distribution of texts, stored in a central computer, via telephone lines. There are actually two slightly different types of usage mixed here. With the first, the text remains in the host computer, and users download (retrieve) it a page at a time, as it is read. The second is the downloading of texts in their entirety, to be read or studied while no longer connected to the host computer. Both of these are good options for distribution of serial material subject to constant updating,[3] but they are not so valuable for publication of editions. Archival storage has yet to be

---

[3] Daniel Eisenberg, "The Electronic Journal," *Scholarly Publishing*, 20 (1988), 49-58.

established, and international access is expensive and sometimes unreliable. Transmission is relatively slow, and only moderate increases in speed seem possible in the near future.[4] To download a book in its entirety can take from very roughly fifteen minutes up to several hours, during which time a computer and a telephone line are tied up. On-line reading not only keeps a telephone line occupied but is hindered by the small delays in transmitting commands and receiving the text requested. If these force pauses in one's reading of literary texts they would be annoying, perhaps even crippling.

A further deficiency of both these forms of distribution is that a publisher would depend upon the skills, machines, prices, and very likely decisions of an outside company. In contrast with printers and binders, there are a relatively small number of manufacturers of CD-ROMS and jobbers of on-line data, and equipment to be self-sufficient is prohibitively expensive for a publisher. Data jobbers (*Dialog*, *BRS*, etc.) keep a high proportion of gross receipts; the pricing policies of some journal publishers and data vendors, the subject of vociferous outrage from academic libraries, illustrate well the economic risks of permitting outside control over data. The Rushdie incident, in which bookstore chain managers were able to withdraw *The Satanic Verses* from many stores at once, illustrates the intellectual dangers of centralization.

I admire and respect the position of those in the computer industry who wish to distribute texts on-line free of censor-

---

[4] The terminology of modems has led people to overestimate their speed. The most common speed of modem, and the one fast enough to make the discussion even possible, has a "baud rate" of 1200; 2400 is the highest speed in general use at present or likely to be so in the near future. These figures refer to the number of bits transmitted per second. As each letter consists of 8 bits, 2400 bits per second is 300 letters per second, or perhaps 50 words per second or 3000 words per minute. Anything but a short book would tie up a telephone and a computer for a significant block of time. These speeds also ignore data lost to overhead (bits which control the flow of data) and telephone noise.

ship or control. However, a mainframe computer is a very vulnerable and immovable facility. The computer industry has not had--yet--experience with the economic, legal, political, and even terrorist pressures which have brought us compressed, colorized, and "edited" movies, the removal of books from school libraries, or, in some countries, suppression of material critical of the government. Should on-line publication of racist, pornographic, libelous, atheist, anarchist, or communist material be permitted? If so, should there be an age limit for access to such material, and should parents be able to override it up or down? Who will decide what is racist and what is not, what is pornographic and what is not? How will the lawyers' bills be paid? Would on-line publishers care to face the to-do that would ensue upon the centralized publication of the new Whitman, Lautréamont, or Agee? Are they willing to go to jail, as printers and traditional publishers have?

What society needs and will implicitly demand is a system that, like our present publishing industry, is decentralized. One that permits inexpensive publication, using common equipment. This has already taken place in a similar technology, home video. Videodiscs, which can be recorded only in special centralized plants, have survived only in industrial applications. The success of the video tape recorder is due precisely to the fact that purchasers could make and exchange their own recordings.[5]

This leads to the suggestion that the medium of choice for digital editions is a low-tech, inexpensive, widely available one: 3.5" and 5" floppy disks. They are cheap, almost every computer can read them, and most important, the same computers can create them without any modification. On my

---

[5] Indeed, it is widely recognized in the industry that a main reason for the early success of the video tape recorder was its adaptability to pornographic use. That X-rated movies are now widely available, even routine, is a direct result of the technical impossibility of controlling their creation and reproduction on the home video recorder.

own computer, a very ordinary IBM-compatible, I can copy in a minute a disk containing approximately 65,000 words.[6] In eight hours, a minimally-trained employee could produce almost 500 copies. That is more copies than a small publisher can sell in a day. Inexpensive equipment for faster copying is readily available.

II. The Identity and Sale of the Text

Digital data can be copied indefinitely, with automatic detection and correction of errors, if any. This is not possible with photographic, xerox, or printed reproductions, or analog (conventional) audio and video tapes, in which each generation of copies brings a decline in quality. Copying of data cannot be limited if any machine with a floppy disk drive or a modem is permitted access to it. Even if one adulterates the text with identifying information, pirates could easily delete it.[7] This, in turn, has economic implications. If one's text can be copied indefinitely, how could one then sell more than the first copy? And if one can not sell copies, how can the costs of editing and publishing be recovered? Would not digital publishing be destructive of the publishing industry?

---

[6] 360,000 letters per disk ÷ 5.5. Simple utility programs can divide larger texts into disk-size units for distribution, and reassemble them later.

[7] There are various techniques to conceal information such as a serial number within a digital text while providing minimal interference to users. If each chapter heading is followed by a varying number of spaces or nulls, they would, taken together, form a serial number; following a letter or number with a back-space code (control-H) makes it invisible to some software; after the end of file marker (control-Z) but within the disk area reserved for the file there is usually room for identifying information also inaccessible to much software. Any of these can be easily read with a powerful editor or disk utility. They can of course be deleted or altered, but it requires a knowledgeable pirate to do so.

Fears of this sort reflect an oversimplified view of publishing. The publisher does not just supply a text, but guarantees that text: that all copies are identical, that the text corresponds to the one which was reviewed in the scholarly journal, that the edition is accurate and suitable for scholarly or classroom purposes, that the notes are current and sound. A pirate can sell a reproduction (of course the pirate would have at least as much to fear from other pirates as did the original publisher), but cannot provide a guarantee.

The need for such a guarantee, for a person or organization vouching for the accuracy of a text, can only increase. Software documentation circulated via computer bulletin boards or duplicators of "public domain" material typically has missing underscoring and italics, poor spelling, punctuation, and capitalization, and formatting ranging from simple to incompetent. More serious errors, such as truncation, are not rare. The producers of these computer files have no idea what an em-dash is, much less why they might need one or how to code it. The proliferation of scanners means that scanned texts, taken from varying or unspecified editions, with the errors inherent to scanning indifferently or painstakingly removed, will soon be circulating. The identification of an undocumented electronic text is much more difficult than with a printed text, for which one has typography and paper to offer some guidance about date and provenance. To use an undocumented text as a basis for research is of course the scholarly equivalent of flying in an uninspected airplane. Horror stories are sure to follow.

As it is impossible to embody a guarantee within the electronic text itself, it must be external. It is the integrity, reputation, and technical skill of the publisher and the scholarship of the editor of the text which will make it salable. It is possible for an on-line facility to guarantee the texts it makes available, although it would have to become a true publisher, with editors and reviews of its products. For an existing conventional publisher, it is the labels on the disks which will communicate the guarantee. Labels can easily be serialized, and their counterfeiting can be detected and

prevented with conventional technology. Holographic emblems with an adhesive backing, used with videotapes, are easily applied and resistant to copying.

The software industry provides grounds for optimism for would-be publishers of digital texts. As computer programs are inherently copiable the software industry has complained for years about theft: loss of sales due to shared copies rather than open selling of their product, for which conventional laws are sufficient. Technical means of restricting copying have been all but abandoned; they were burdensome for the legitimate customer, yet only briefly hindered the thief. Stolen software, users have realized, comes without manuals, assistance from the publisher with problems, and information about program improvements and new products. There has been support for royalties among the more enlightened computer users, a recognition that paying for software is in their own long-term interests, and new forms of payment, such as site licenses, have been implemented at the request of customers. Some smaller companies have invited users to copy and share their programs, viewing such distribution as free advertising. Most important, the software industry has thrived despite what was, only a few years ago, seen as a serious threat.[8] The foreign situation lags behind that in the U.S., but progress is reported.

III. The Architecture of the Electronic Edition

An electronic text of the King James Bible first was offered for sale to the public, as an ordinary business product,

---

[8] Software, one should note, is sold on floppy disks. None of the several attempts to sell software on-line has so far been viable, nor has there been, to my knowledge, a CD-ROM devoted to commercial software.

in 1982.[9] Various Biblical and classical texts have become available in the following six years.[10] Very recently, Shakespeare and some Library of America texts have become available.[11] Yet these electronic publications do not include notes, variant readings, or introductions (other than brief introductions in hard copy). They are devoid of italics, superscripts, and special characters. They have been monolingual, and primarily in English.

The problem of how to remove these limitations has been a major stumbling block to electronic publication. If, for example, one includes annotations in the same file as the text,

---

[9] The earliest advertisement known to me is on p. 100 of the January-February, 1984, issue of *Profiles*, a now-defunct magazine of the KayPro Corporation. The price was $200, which included a retrieval program called The Word. The company, Bible Research Systems of Austin, Texas, has confirmed by telephone that the product was first sold in 1982.

[10] Nine different bibles are currently available for sale, at prices ranging from $9 to $259, some including software (*Computer Shopper*, April 1989, pp. 312, 520, and 522; unspecified "brand new adult novels on disk" are offered on p. 519 for $11.95 each). Various classical texts in Greek, Hebrew, Arabic, Sanscrit, and other languages, digitalized by the University of Pennsylvania, are available from Gamma Productions, 710 Wilshire Boulevard, Suite 609, Santa Monica, CA 90401.

[11] Some works of Shakespeare are available from Shakespeare on Disk, Hollow Road, PO Box 299H, Clinton Corners, NY 12514 (according to an advertisement in *PMLA*, 104 [1989], 270). The Riverside text of Shakespeare ($299), eight Library of America editions, the King James and New International Bibles, and "The Constitution Papers" are available from the Electronic Text Corporation (5600 North University Avenue, Provo, Utah 84604), which in its literature requests suggestions concerning titles or specific editions to be published. The latter company is the publisher of a text indexing and retrieval program called WordCruncher.

as most word processors do, then the file becomes much longer and any program to analyze the text must be programmed to ignore notes. A word with a hyphen inserted at the end of a line (called a "soft hyphen") is to the computer a different word. Italics, bold face, superscript, line numbers adulterate the text with data not part of the author's words.

This leads to the suggestion that all subsidiary codes and information, except an alphabet identifier[12] and a backspace or overstrike code,[13] be placed in external files. The text would thus be in as nearly pristine condition as possible, ready for any type of electronic searching or analysis. It would be usable by a wide variety of programs, including the

---

[12] If one did not have a change of alphabet code in the file, the computer could not tell whether a given string was, for example, "mañana" (in "Spanish ASCII," discussed below) or "ma|ana" (in "US ASCII"). One would not know whether one was looking at a Greek word or the English characters which happened to occupy the same positions as the Greek letters.

[13] ASCII code 8. Simple equipment would display only the second character, that following the overstrike character; some sorting programs will correctly ignore a diacritic or letter if it is followed by backspace. More advanced equipment would combine two shapes to form a single one, a capacity which video boards do not ordinarily have today, yet easy to implement. The replacement of ´ plus i or ^ plus i by the dotless í and î is a function of the projected output (reading) program, as is the combination of ae to produce ä (in German) or æ (in Latin), the ss to produce ß in German, and fi to produce the ligature fi. It requires an astronomical number of codes, which will not sort correctly, to assign separate codes to `a', `a' plus acute accent, `a' plus macron plus acute accent, `a' plus grave accent, `a' plus macron plus grave accent, `q' plus tilde (a common Renaissance abbreviation), ya plus shadda plus fatha plus nun (Arabic), omega plus iota subscript plus smooth breathing plus acute (Greek), etc.

The overstrike can also be used to produce cancelled letters, underscore, and to reproduce misprints, which editors sometimes need to discuss.

simple “TYPE” command.[14]

There are already extensive precedents for the distribution of a group of files as a unit. A disk can of course contain many files, which can be arranged in directories; when one purchases software one usually purchases such a collection of files. Simple utility programs, already in wide use, combine related files into superfiles called “libraries” and “archives.” The same utility programs also reverse the process. It would be also a simple matter to incorporate into such collective files the heirarchy referred to below.

What files would the electronic edition consist of? For an edition of a text existing in a single version, whose alternate readings could be contained in notes, a directory containing the following files is proposed:[15]

1. The text file;

---

[14] In digital texts prepared for input to typesetting equipment, carriage returns are used only at the end of paragraphs, and each paragraph is thus a single long line. However, it is a simple matter to convert unneeded carriage returns into spaces if the need arises, and most typesetters working with texts produced on word processors already have procedures for doing this. Therefore, so as to make the text displayable and printable with the simplest of commands, it is proposed that carriage returns (carriage return-line feed pairs, actually) be included in the text file no less than every 60 characters. Two carriage return-line feed pairs would separate paragraphs or stanzas; a single pair followed by tab would separate lines of verse. This scheme is used by simple word processors.

[15] Variant texts can be handled within the textual annotations, which can indicate texts to be deleted, alternate texts, anything except extensive relocation of passages. An edition with relocated text would require processing so as to relocate the corresponding annotations; WordPerfect’s “generation” process does something similar. Parallel texts in which no single text can be labeled as the more correct “master text” require that each text be in its own subdirectory with its own notes, and that an artificial numbering scheme be constructed in the main directory.

2. A table of contents or guide to the chapters or other divisions of the text file, allowing one to move rapidly to the section desired, and providing data for a status line or running head or foot;
3. A file containing the attributes modifying the text, on which more shortly;
4. An index/concordance file of the words in the text file;
5. A file with cataloguing information;
6. A file with the copyright notice;
7. A file with a blurb or brief summary of the contents and characteristics of the edition;
8. A file with the publisher's catalogue and order blank.

Some other items part of an edition, while they are dependent on the main text, are texts in their own right. For example, a historical introduction to a work is itself a text and may require notes. The author of a text may provide additional text in the form of notes (Nabokov's *Pale Fire*). Even editorial notes can themselves be annotated, explaining the reasons for an emendation, or, with older notes especially, examining what was meant or the procedures by which the annotation or emendation was created. Annotations are sometimes issued separately from the texts they comment on. Professor X's edition of fifty years ago may need an up-to-date set of annotations and emendations by Professor Y.

In other words, notes are themselves texts. This reality can be supported if a hierarchical structure is created, in which each type of secondary material is permitted the same treatment as the main text. Notes would then have their own title pages, copyright notices, introductions, indexes, and could themselves be annotated, as could the notes to the notes, and if needed the notes to the notes to the notes, as has happened with Biblical texts. What is proposed is an open structure.

Such a hierarchical, tree-like structure is in fact part of the MS-DOS operating system used on all IBM-compatible

computers.[16] This operating system uses subdirectories, names and addresses of which are included in the directory just superior. In the main directory of the edition, then, in place of a file one includes a reference to another directory, on the same disk, containing the subordinate file of notes and all its peripheral information. Thus subdirectory includes, if needed, another reference to a doubly subordinate directory with notes on the notes.

Thus, to continue the contents of the main or root directory:

9. One or more references or disk addresses for subdirectories containing files of textual notes and variants. One such subdirectory might contain, say, the information needed to reconstruct the first edition. Another might contain an edition revised by an author. "Variants" might consist of a translation of the text. Readings produced by different editors could be contained within different such directory groups; the program governing the reading would permit the reader to select one of these alternate texts. Each subdirectory would contain:

a. the contents of the notes, keyed to words of the text, or to paragraphs or changes of speaker, in the case of a translation;
b. the attribute file for the notes;
c. an index/concordance file for the notes;
d. the summary description of the notes (the editor who prepared them, the public to which they are addressed);
e.f.g. Cataloguing information, copyright notice, blurb for the notes, etc.

---

[16] This discussion is based on the MS-DOS operating system. It is my understanding that similar provisions exist with the UNIX and Macintosh operating systems, with which I am less familiar; it is easier to emulate IBM on the latter than the other way around. Some of the features proposed have been implemented in the "Hypertext" programs of the Macintosh.

h. Reference or disk address of a subdirectory containing notes on the notes;

10. One or more references or disk addresses for subdirectories containing annotations, each of which directories would contain the same type of files as the textual notes. Digitalized illustrations, adornments, or musical examples would be annotations to points or passages in the text;

11. One or more references or disk addresses for subdirectories containing other types of notes (a page by page analysis of sources for a medieval historical work, for example; identification of the parts of speech or lemmas [roots] of words of the text);

12. One or more references or disk addresses for subdirectories containing essays on or introductions to the work;

13. One or more references or disk addresses for subdirectories containing notes created by the reader.

IV. The reading program and the attributes file

The proposed reading program would take the text file, with its simple formatting, and display it in a more sophisticated fashion. Right and left margins would be set, as would indentation. Text would be justified if desired, and the reading program would hyphenate as needed. On request it would display explanatory annotations and the variant readings included in the file(s) of textual notes.

The reading program would also interpret and express the contents of the attributes file referred to above. It is a condensed file, paralleling in its contents the text file, containing data about the text on a section, word or character level. This file would include information needed to enhance the display of the text. This includes codes indicating the languages of the text (necessary for correct hyphenation, among other uses), markers labelling titles, subtitles, and similar text parts,

the hyphenation of words which require sentence analysis,[17] references to standard paragraph or page numbers for classical texts, information to be inserted to the left or the right of the text (standard line numbers, the name of the character speaking, in drama), and reader-inserted markers (where one was when one stopped reading, for example; the electronic equivalent of the folded-down page corner).

The attributes themselves are modifications of the text: to indicate it is important (in traditional typography done with italics, bold, or larger type, on typewriters and some computer printers by underscoring and "shadow" printing), to indicate it is unimportant (cancelled letters; smaller type), to mark it as distinctive (indentation, super- and subscript). The screen expression of these attributes is set by the output program in accordance with the hardware and the preferences of the reader. Important text might be shown on one system by true underscoring, on another by a particular color of letter or background, on another by large letters, and on another by blink. The ability to see all possible combinations of attributes is desirable; as the number of colors people can readily distinguish is limited, the ability to show a two-tone background would be helpful. Some on-screen attributes would need to be reserved for reader-marked emphasis (the equivalent of the marginal line or highlighting), to indicate visually that textual, explanatory, or other notes exist for a word or passage, or for special text characteristics (in epic poetry, that a line is or is not formulaic). The design of the screen display--its font, colors, dimensions, letter size, status line and so on--requires the attention of graphics designers.

Printing was unable to reproduce some attributes found in manuscripts: the use of colored inks, for example. Similarly,

---

[17] Pro-ject (verb) and proj-ect (noun); pro-gress (verb) and prog-ress (noun); at-trib-ute (verb) and at-tri-bute (noun); as-so-ci-ate (verb) and as-so-ciate (adjective, noun); sa-ke (the Japanese beverage) and sake (purpose). The hyphenation of Spanish is so blissfully simple, though arbitrary, that one wonders if the complexity of English's more precise system is worthwhile.

in an electronic text one would have to abandon, as prohibitively complex, some characteristics of type. These include: the selection of specific type faces, sizes, or colors (such decisions to be determined by the reading program); reproducing the layout of earlier printed versions of the text;[18] the use of vertical, slanted, curved, inverted, or mirror image type; the use of alternate shapes to represent the same letter, unless these could be mechanically selected by the output program (the final letters of Arabic and Hebrew); the setting of type into designs (the mouse's "tail" in *Alice in Wonderland*); the breaking of lines in the middle of words, as with:[19]

now
i can tell
of being swept b
y a god a michael
angelo's david a
man of such phys
ical perfection,
one could not be
lieve him human

Illustrations accompanying the text could accommodate any such material. One could "read" such a book by moving from illustration to illustration.

---

[18] The textuality (interpretative significance) of typography and page layout was the subject of papers at the 1988 Modern Language Association convention (*PMLA*, 103 [1988], 960-61).

[19] From "An Idyll" by Ana Castillo, quoted by Norma Alarcón, "The Sardonic Powers of the Erotic in the Work of Ana Castillo," in *Breaking Boundaries. Latina Writing and Critical Readings* (Amherst: University of Massachusetts Press, 1989), pp. 94-107, at p. 100.

V. Alphabets

Computer character sets or alphabets do not receive much attention. When the topic comes up, it is usually centered around problems in printing. ("How do I get this character printed on paper?") Once solved the matter is forgotten until the next crisis. As a result of this inattention, the situation is chaotic.

Because of the binary system on which all digital computers operate, 256 (2^8) is a logical size for a character set. 16 (2^4) is equally logical, but it is too small, and 65,536 (2^{16}) is also logical, but it is too large. 256 is within reason. All personal computers, regardless of how many characters they can display or print, work internally with sets of 256 characters.

Older equipment was only able to use 128 of the 256 possibilities.[20] A standard set of 128 characters and control codes was formalized as the American Standard Code for Information Interchange, known by its acronym ASCII and sometimes referred to as "standard ASCII" or "US ASCII" (though there are no other "ASCII"'s than the US one). This is in fact used for all personal computers, although parts of it have been altered. It includes the diacritics ´ (the acute accent, also used as apostrophe, not the other way around), `, ^, and ~; it also includes 31 non-printing characters intended to control hardware, communications, and data.

From this relatively standardized beginning a generation ago things have deteriorated. Hardware manufacturers, *motu propio*, adapted the 128-character ASCII set to equipment sold in other countries; there are now at least 10 foreign

---

[20] The 256 characters are the number of possibilities one has with eight "bits," each a microscopic electronic switch. The two possible states (on or off) for each of the eight switches gives a total of 256 possible combinations (2^8). The earlier equipment used one of the eight switches as a check to prevent internal errors (a parity check), thus leaving only 128 possibilities (2^7).

sets.[21] There has been no standardization for the 128 new characters made available by improved hardware. The IBM PC Graphics set has been the most successful, and is sometimes erroneously referred to as "extended ASCII" or even "ASCII" pure and simple; it has been widely used on IBM-compatible computers. However, PostScript, used in digital typesetting, has its own set, as does the Roman-8 character set found on laser printers. Advanced word processing programs have gone beyond the 256 in various ways. Text formatting languages, oriented towards printed output, have devised cumbersome, English-language, but unambiguous codes to represent additional characters within the original 128.

As a result, we now have the following situation. To represent an ñ within the ASCII set one combines a ~, an overstrike code, and an n. Other sets use single-character combinations: the so-called Spanish ASCII uses character 124, the IBM PC Graphics set uses 164, PostScript uses 4, the Roman-8 set uses 183, and a Star printer I own uses 222.[22] As the tilde is seldom used for its original purpose, it, and various of the control codes of the original ASCII, have been put to contrasting purposes by different pieces of hard- and software. As a result, chaos exists with all characters outside the basic English alphabet and punctuation. The standardization of the original ASCII and the relatively successful PC Graphics set contributed greatly to the success of personal computers. Our present state of affairs casts a pall over electronic publishing.

---

[21] The following are known to me: France, Germany, U.K. (using the pound in place of the number sign), Denmark I and II, Sweden, Italy, Spain, Japan, and Norway. These sets vary from one hardware manufacturer to another, and the deficiencies of some seemingly reflect ignorance of the languages.

[22] The formatting language TeX uses an n followed by \tilde, the University of Chicago recommends <tid>n, and the Association of American Publishers' Electronic Manuscript Standard uses ñ.

It is impossible to devise a new standard set of 256 characters serving all purposes. (There is no way to incorporate both a Greek and a Roman alphabet within 256 characters, for example.) Nor is it a reasonable goal to seek to convert existing or future electronic texts in these various character sets into a single standard, nor would such a conversion be simple: would one convert printed-style quotation marks “ ” into typewriter-style quotation marks " ", or the other way around?[23] Instead, a meta-system encompassing all present sets is needed. It is proposed, therefore, that the varying existing sets all be accommodated, along with others yet to be determined A 0 following the change of alphabet marker would indicate an improved default set.[24] A 1 would indicate the PC-Graphics set, a 2 the Roman-8 set, a

---

[23] Printing-style quotation marks require, for aesthetically pleasing display, a variable character width, something not usually implemented on computer screens. Computer-displayed text more resembles typewriter copy than printing, and will continue to do so for the near future. The resolution of today’s best monitors is only about 100 dots per inch, far below the 300 of ordinary laser printers or the 2400 used in typesetting.

[24] Such a set would include a non-break (hard) space, and a minus (a dash which should not be used to end a line, such as with a negative number -10); I cannot see the need, however, for soft hyphens and returns. It would include a full set of diacritics, distinguishing the acute accent from the apostrophe and the grave from the open single quotation mark. It does not need both guillaumettes « » (PC Graphics 174 and 175) and quotation marks, as they perform the same function. (In the U.S., the « » might be *displayed* as “ ”. It would be easy to add to word processors the automatic pairing of these codes, and other pairs such as parentheses, while typing.) Nor does it need country-specific characters (such as the Spanish Peseta symbol, PC Graphics character 158), which would be left for national character sets. Some characters would be left unassigned, for use with “illustrations” of unique character shapes, specific to a particular text, the data for which would be stored at the beginning of the attributes file.

3 the Postscript set, 4 the "French ASCII," 5 the "German ASCII," another dingbats (special characters), and so on. Numbers would be reserved for orthographers to design better sets for the various world languages, others could be assigned to professions (mathematics; astronomy; music). The International Standards Organization exists to keep track of them. All characters used since the inventing of printing (and scholarly editors need to use all of them at one time or another) will fit within such a system. To the extent possible these sets would duplicate and overlap each other. One number of the 256 would be reserved for expansion (if the alphabet marker were 31, which would cause as little conflict as any, then a second 31 would signal a second set of 256 alphabets).

VI. The Prospects for Scholarly Editions

Converting a manuscript into a printed book was a complicated process, involving not just mechanical work but judgement. Which works should be published first? (The classics were in fact published first.) What manuscript should be chosen, if there are many available, as the basis for making 1000 new copies? Should one conserve the readings of the different manuscripts examined, so future scholars could check one's decisions? Are the manuscripts themselves worth conserving? (Usually they were discarded.) Which scribal abbreviations should be conserved in type?[25] Should one perhaps restore defective language, standardize spelling and punctuation? If the latter, what are the standard spelling and punctuation to be? Establishing a consensus on these ques-

---

[25] $, ¢, £, %, @, &, and # were finally conserved. The tilde was restricted to the n in Spanish, the a and o in Portuguese. Other abbreviations, such as a crossed "p" to indicate "per," were abandoned. Superscript letters of abbreviation were originally abandoned, reintroduced on a limited basis much later (the ª and º of Spanish, ème of French, occasionally others).

tions took several generations.

Much the same sort of decisions await us now. It should be clear from the above that the fantasy of inserting a printed book into a machine and receiving out the other end an electronic version of the same is just that, a fantasy.[26] Broken or mispositioned type and uneven inking make error-free scanning of many books impossible. The ways in which notes, headings, and bibliographies are printed in different editions will make inputting of them a matter of considerable expertise, requiring software of its own. There are also such problems as distinguishing soft hyphens from hard, quotation marks from apostrophes, pro-ject from proj-ect, and the identification and correction of true misprints and obsolete spellings in the original. Only with a multi-volume set of uniform format can standardization be achieved.

In short, just as many manuscripts were never printed, many printed books are never going to be converted into electronic format. There is a shortage of skilled people, and these limited resources will be concentrated where there is most demand. The selection of books to be converted is a matter which will occupy considerable scholarly attention over the coming generation, as will the related question of whether the editions being scanned can or should be improved. In some cases it might make more sense to reedit a work than to reissue an existing edition; reediting will also be much easier with all these tools at our disposal. The reediting of texts was one of the most long-lasting contributions of the

---

[26] A derivative of this fantasy is the confidential hope of the president of an ivy league university [Brown University] that the expensive, bulky, and ever-growing library will be replaced by a computer data bank and decentralized terminals. The new "library" would be a facility like the fire station (his analogy), to which one rarely needed to go in person. A generation ago it was of course microfilm that was going to replace the library.

Renaissance.[27]

[27] This article was written using the outline processor Kamas (Knowledge and Mind Amplification System), published by Kamasoft, P.O. Box 5549, Aloha, OR 97007.