To Save or Not to Save: That Is the Question

(prepared for the panel on "Preserving Electronic Writings at the annual meeting of the Association of American Law Libraries, Seattle, Washington, July 11, 1994)

Before I begin talking specifically about the topic we have been assigned to address on this panel, I'd like to preface my remarks by saying that I take my contribution here to be part of an ongoing effort by university presses to become more engaged in the dialogue that librarians, to their credit, began many years ago as they reacted to the growing serials crisis and looked ahead to the day when the emerging electronic networked environment might provide solutions to the crisis or at least alternative ways of doing business that could circumvent the problems libraries faced in the traditional world of print publication. University presses have for a long time existed at the margins of the academic communities in which they reside. Their importance to scholars, particularly in the humanities and social sciences where presses' activities have mainly been concentrated, has of course been recognized, but presses have not had much interaction on campus with other entities that are equally part of the system of scholarly communication--not only libraries but also computer centers, college bookstores, and university printing services.

But times are changing, and over the past few years university presses have awakened from their slumber and, with the support of their association (our AAUP as distinguished from the AAUP of university professors), have made great strides in adding their voices to the dialogue. There have already been three annual conferences jointly sponsored by the AAUP and the ARL on "Scholarly Publishing on the Electronic Networks," and earlier this year the Coalition for Networked Information, in cooperation with the AAUP, endorsed pilot projects at thirteen universities (including my own) where presses have joined with libraries and computer centers to launch experimental programs in electronic publishing, the best known of which so far is probably Johns Hopkins's Project Muse. The AAUP has also been active in other fora where efforts are being made to bring the various interested parties together for discussions about how to make the system work in our present and ever-changing environment. I serve myself, for example, on the subcommittee of the Copyright Clearance Center that is following up the CCC's pilot project on photocopying in universities with the aim of eventually having a viable licensing scheme to offer, probably of the blanket license variety. And I also serve on the Task Force on Copyright Compliance of the Association of American Publishers the aim of which is more to stimulate on-campus dialogue to resolve copyright problems in a mutually beneficial way than it is to plan strategy for litigation. Finally, I might mention that the AAUP has just instituted a new policy committee that will work toward formulating positions for the university press community on issues of central concern to all of us, such as intellectual property, much as the ARL has done so successfully, as recently as last month with its own statement on intellectual property. In these many and diverse ways university presses are thus eager to play the role that the AAU Task Force on Intellectual Property recommended in the report it released in April. "The next crucial phase," the report concluded, "is to build campus consensus and bring other organizations, particularly the AAUP..., into the process."

With that rather lengthy introduction, let me now move on to the topic at hand. What is the role of university presses in "preserving electronic writings," and what are they actively doing about it at the present time? When asked to give a title to this paper long before I had even begun to think about what I might say, I quickly came up with "To Save or Not To Save." That is indeed a question presses are beginning to confront in a more self-conscious way than ever before. One sign of this emerging self-consciousness was a discussion in March on the AAUP's listserv about "longevity of media." It began with a posting from the director of the Ohio State University Press, who asked: "how long does an electronic text last?" He went on to elaborate: "in thinking about archiving our book composition files it occurred to me that I don't have any sense of how long the various storage media for electronic files last. That's not quite true; I have a dim sense that they don't last very long, but that's only based on my own limited experience with diskettes going south, hard drives failing, etc.... For those of you who are storing your composition files, what do you regard as a proper archival medium?"

This posting generated a flurry of responses that carried over several days, but then petered out at the end of that month. In preparing for this talk last week, I resurrected the discussion for another round of several days by asking for replies from presses that are following some kind of policy on saving electronic files to these two questions: "1) what kinds of publications are you saving electronically (books, journals, both, or some of each), and 2) in what type of storage media are you saving them?" What I will report to you today comes from the replies I received to my posting as well as those that were responses to the earlier posting. This is not based on any scientific survey, obviously, but I suspect that what I learned comes pretty close to representing the truth for the full universe of presses even though replies came from only about a dozen, because these dozen include most of the presses that are at the cutting edge of the electronic revolution as it has affected our particular segment of the publishing industry.

To refer to the title of this paper again, I think it's safe to say that presses divide into three groups: one group (which includes my own press) that does no in-house composition and is therefore not even raising the question of whether to save or not; a second group (which would include Ohio State) that is just beginning to ask the question and is trying to formulate answers to it by investigating what storage media might be most appropriate; and a third (which includes such presses as California, Chicago, Colorado, Illinois, Johns Hopkins, Minnesota, MIT, North Carolina, Princeton, Tennessee, and Texas) that already has been saving electronic files. If I were asked to estimate how many presses fall into each category, I'd guess that out of the 100 or so U.S. presses at least half fall into the first group, thirty more fall into the second, and probably only twenty into the third. Even among the third group it might be stretching to claim that all these presses are following anything that could be called a formal policy; as the production manager at Illinois put it, "we have a practice rather than a policy, if that's not too fuzzy a distinction."

It seems pretty clear, just from the great diversity in the responses, that the presses in this third group have come to do what they now do in a more or less ad hoc fashion, devising their policies on the run as a function of the particular production setup in place at each press and the budgetary constraints under which each press has to operate. There is thus little uniformity in storage media used, for example: several presses (Colorado, Johns Hopkins, North Carolina) are storing just on floppy disks, though a few are now getting ready to move everything to Syquest cartridges as the numbers of floppies in the archive steadily mount; one press (Princeton) is using magneto-optical WORM disks; another press (Minnesota) stores everything on a MacIntosh DAT tape; still another press (Illinois) employs 1 GB optical disks.

Chuck Creesy of Princeton provided some useful comparisons among the various alternative storage media one might consider, in terms of both useful life and cost. Here are his estimates:

floppy disks: 5-10 yrs.; 50c/MB
hard drives: 10 yrs.; 80c/MB
WORM disks: 20 yrs.; 6c/MB
recordable CDs: 25 yrs.+; 5c/MB
Syquest cartridges: 10 yrs.; 60c-$1/MB
backup tapes: manufacturer's stated life; 3c/MB

Various other considerations need to be kept in mind. Backup tapes, while the cheapest option, don't make finding files easy and making copies of tapes is "a bit of a hassle unless you have special equipment." Hard drives "are a crap shot"; some crash in the first month, others are still humming away after ten years. Floppies and CDs need to be kept in a temperature-controlled environment to ensure maximum life. Creesy concluded his comparisons with this sobering thought: "Given the relatively short timeframes before most electronic media need to be copied or refreshed, there's still a lot to be said for acid-free paper. One of the surest archival methods may be to print out files on a high-quality laser printer, using a standard typeface in a large point size that can be scanned at a high level of accuracy. It's been downhill ever since we stopped chiseling words in stone."

How long have these presses been saving electronic files, and what do they save? Although one press (North Carolina) reported having started saving files six years ago, it appears that for most presses systematic saving only began within the past two or three years. At presses with large journals programs, such as MIT and Johns Hopkins, it's evident that the journals department got the ball rolling first, and books came second; at most other presses (some of which publish few or no journals) books are the focus of the effort. Presses that do in-house composition are the ones in the forefront here, and what they save are mainly the files they generate themselves; few of them bother to try to retrieve useable files from outside vendors. The reason is partly financial; vendors may charge as much as $200 to turn over useable files to a press, and the press has to figure out what further benefit might come from having these files to decide whether paying that price can be justified. But the reason is also practical: it can take a lot of time and trouble to convert these files to a form the press can use. The production manager at Illinois reported on two efforts that proved unsuccessful, "one because of the tremendous amount of garbage we had to remove before we could reuse the file for a revised edition and the other because after five or six attempts (that involved much interaction with the compositor) we were ultimately unable to access the text files." Just to give you an idea of how many books are still being typeset by outside vendors, for even these presses that are most advanced technologically, Princeton (which is probably the most advanced of any press) still does only one-third, about sixty, of its books in house.

One of the most interesting parts of the discussion had to do with the format in which the files are being stored and the corresponding need to save software files and even hardware so that at some point down the road the formatted files could still be retrieved and read. North Carolina, for instance, saves its own in-house generated files in XyWrite that contain its own generic coding system for typesetting, while also saving disks from outside vendors in those vendors' own coding systems; Illinois and Tennessee save everything in PageMaker; Colorado and Johns Hopkins use Postscript files for their archiving; Minnesota is planning to store "some kind of plain text, some kind of program useable file, and some kind of Postscript file" and, in addition, will "store a copy of the program that created the program useable file." It's at this point that as a non-"tekkie" I begin to get lost!

The final part of this discussion rose to a more philosophical level, as Chicago's information systems manager, Bruce Barton, posed this challenge: "We have been assuming that we are in the text formatting business.... This is a mistaken but common enough assumption.... Our task in setting type is to represent content clearly, perhaps elegantly. Content, understood as the logical structure of a document (rather than its argument), has always played a critical but unsung role in book design. It has almost been one of those things that goes without saying. And in our electronic document preparation, this has been literally true. Capturing content, not format, is what we should aim for. Saving a QuarkXPress version X document (format) allows us to reproduce pages reliably using QuarkXPress version X. Capturing typesetting tapes from Y typesetting system allows us to use Y typesetting system to reproduce typeset pages. Neither helps us to do anything (easily). If, on the other hand, we were to tag the structural elements of a document in our canonical electronic representation of it and store it in this austere form uncontaminated by format, we could move to formatting on a wide variety of delivery platforms easily, that is to say, programmatically.... These platforms include traditional typesetting; the electronic book on CD, floppy, or whatever; slicing and dicing for course packs; mounting in searchable, on-line databases; the rich hypertextual world of World Wide Web and its successors. I am arguing that we need to make explicit a step that has always been implicit in book production: codifying content. This step occurs immediately prior to mounting the book on a delivery platform. SGML is the best currently available method for accomplishing this." To this challenge Chuck Creesy of Princeton replied that he agreed totally on philosophical grounds but that decisions being made now have to take into account "narrow practicalities and tight budgets." As he put it, "We don't want to pay for SGML-encoding a monograph today that there may be no market for in the electronic tomorrow, but we want to preserve what we have just in case there is a demand for it, so long as it doesn't cost too much to do so. Accordingly, I have long argued for saving the electronic text with the richest possible coding intact--not for future typesetting but for future conversion to something else that I cannot even imagine yet. Don't throw anything away and keep as many options open as possible." Creesy, after surveying storage options again, then concluded: "As we evolve toward thinking of print as only one of a range of products, we will also convert to Bruce's view of coding for content and structure rather than typestyles. In many ways, we are passing through an intermediary stage; the problem is trying to perceive what lies on the other side."

On this forward-looking note I conclude the survey of current university press practices for "preserving electronic writings." It's clear that there are many different approaches being tried, and far from everything is being saved. Where do we go from here? I suggest that it is critically important for us all--libraries, presses, and other members of the university community--to begin the process of consultation that the AAU report recommends. The AAU's focus was mainly on issues involving intellectual property, and that's certainly part of the picture. But it's not all, by any means, and in fact, as someone who has advised the AAUP on copyright matters for over twenty years and been convinced of its importance at times when other people weren't paying any attention to copyright at all, I feel that now copyright is being viewed too much as the central issue. I believe, rather, that the current fixation on copyright is a symptom of the diseased condition of scholarly communication and that a solution focused on copyright alone will not result in a lasting cure. It would take me another whole paper to lay out a full argument for this claim, but let me at least give you a sketch of it.

Librarians have been very vocal in calling for the preservation and even extension of "fair use" principles in the electronic environment. Two of the seven "principles" in the ARL's recently adopted statement on intellectual property, for example, focus attention on "fair use" in this manner.