

---

# DIGITIZATION FOR THE MASSES

Michael Seadle

## Introduction

On 22 October 1997, Clifford Lynch, president of the Coalition for Networked Information (CNI), spoke to the Federal Funders Group on his concerns about the development of digital libraries. One of his points was that scarce resources get burned up in too many small-scale local projects, which then tend to wither away for lack of institutional support. It is a real problem, and there are several ways to address it.

One approach is to concentrate funding on a few large institutions that have the technical skills and institutional resources to manage major long-term projects of national scope. Universities such as Cornell, Yale, and the University of Michigan have demonstrated their ability to do so through the CLASS Project (Cornell), Project Open Book (Yale), and Making of America (Michigan and Cornell). Although their work certainly has had local application, the plans were consciously drawn with national goals and national needs in mind. These three institutions, and a few others like them, have committed scarce resources, including their own, in ways that will benefit researchers everywhere. Favoring them with the funds to undertake new and bigger projects would make sense. The work would get done and get done right.

In fact, however, this approach will never happen. The institutions that give out grant money, both public and private, are conscious of pressure to spread their largess among a diverse population. The Library of Congress/Ameritech American Memory Project has, for example, created two classes of competitors, one

---

*Seadle* is digital services librarian at Michigan State University, East Lansing, and co-editor of *Library Hi Tech*. <seadle@mainlib3.lib.msu.edu> .

for members of the Association for Research Libraries (ARL) and one for non-ARL institutions. And even within the ARL, enormous differences in size and wealth exist. Digitization may not quite be for the masses, but it is certain that a very broad range of institutions will be undertaking digitization projects in the coming years. This leads to the second approach: helping small and inexperienced institutions do it right.

This is precisely what places like Cornell and Yale are doing. Their projects are among the best documented in the library world as part of a conscious effort to share what they have learned. People like Anne Kenney of Cornell, Paul Conway of Yale, and David Seaman of Virginia teach, talk, and travel extensively for the benefit of hundreds of other institutions. Participating in a workshop with experienced people is invaluable for the digitizing novice. Although digitization seems deceptively easy—just a matter of running a scanner and throwing an image on the Web—it is not. Even an intensive week-long workshop barely touches the surface. Extensive reading is a must.

The purpose of this article is two-fold: to convince the reader of the importance of knowing the digitization literature and to introduce some of the most important elements of that literature.

## The Big Choices

In the early stages of a digitization project, it is tempting to focus on purely technological issues, since those are the ones that seem unfamiliar. This is a mistake. The nature of the project should determine the technical decisions. It is easy to spend time and money in ways that are inappropriate, inadequate, or simply unneeded. The first step should be to decide how much emphasis to put on preservation and how much on access.

“Preservation is access and access is preservation,” writes Paul Conway, and the sentiment is widely shared in the digitizing community.<sup>1</sup> His point is that digitization both preserves an intellectual work by making an image of it and improves access because of the ease with which digital images are transmitted. It does not make sense to do one without realizing the benefits of the other. But in practice either preservation or access generally weighs more heavily among the reasons for digitization. Those who emphasize access are easily tempted to put off preservation considerations such as how to store and refresh high-resolution copies, just as those who digitize mainly for preservation may be slow in putting images in easily readable form onto the Internet. The extremes of this behavior are among the reasons for Lynch’s concerns. Although the digitization literature puts strong emphasis on the need for balance, ultimately it favors preservation because

of the high costs involved in rescanning a work that originally was scanned hastily for short-term needs.

Ideally every digitization project is done both for preservation and access. In a less ideal world, however, curriculum deadlines may drive decisions in an access-oriented project. Getting some form of images up may take precedent over perfect alignment, contrast adjustment, or other quality considerations. Institutions that make that choice may well be doing so for legitimate internal reasons, and they may be tempted to ignore parts of the digitization literature that seem to be hectoring them for doing it wrong. Choosing access as the primary goal is not a mistake, but ignoring the literature is. The choice of what corners to cut should be done with full conscious knowledge of the consequences.

A second major choice is whether to do photographically exact images or structurally marked-up, searchable ASCII text—that is, the choice between the Cornell-Yale model, which emphasizes TIFF (and GIF) images, and the SGML-based approach of institutions like Virginia’s Etext Center. Happily it is not an absolute choice, since high-quality TIFF images often are fed into an Optical Character Reader (OCR) to create ASCII text for SGML-markup. Unfortunately not all images will feed into an OCR well. Typing services actually may create a more accurate representation.

The choice between the photographic and the SGML approach should depend on how the final product will be used. If searching the text for words and phrases is important, especially in the context of a particular structural element within a work such as a chapter or a stanza or an act, then SGML-markup matters. If the main use is to read the pages much as they are in the printed work, the photographic approach may suffice.

## Organization

The bibliography portion of this article covers aspects of all four of the “big choices” options. But rather than try to draw the boundaries between these options more sharply, it emphasizes issues that cut across all options.

The bibliography begins with general works designed to give someone completely new to digitization an idea what it is all about. Closely related to that overview are examples from a few landmark digitization projects.

The next five sections borrow from Paul Conway’s framework for action and examine choice (what to select for digitization), quality (how to do it right), longevity (how to ensure that it lasts), integrity (how to prove that a work is genuine), and access (how to

get to it).<sup>2</sup> None of the elements of Conway's framework is unique to digitization, and that is what gives them importance. Digital artifacts are an integral part of the library world and need to be judged in its terms.

The bibliography finishes with a brief look at SGML markup and an overview of copyright—which no digitization project can ignore without peril!

### General Works about Digitizing

Anyone who is new to digitization should have Stuart Lynn's glossary near at hand to translate the field's vocabulary into more familiar terms. Lynn was one of the moving forces behind Cornell's CLASS project. Another person closely involved with digitization at Cornell is Ross Atkinson. Although his approach may be too monolithic for the taste of many in this era of decentralization, no one who wishes to understand current trends and ideas can ignore his widely discussed article and the solid intellectual groundwork it provides for the "virtual library."

If a person were to read only one work on digitization, it should be Paul Conway's *Preservation in the Digital World*. In it, he lays out the full range of elements involved in digital preservation and access based on his experience with Project Open Book. Conway's shorter article on "Digitizing Preservation" captures key issues succinctly, like the high-entry cost of digitization.

The RLG symposium at Cornell, which Nancy Elkinton edited, gives an overview of digitization up to 1994, with a strong emphasis on the pioneering work at Cornell and Yale. The book may appear slightly out-of-date in this fast evolving field, but the basic issues have changed little. Cooperative projects are a key part of the landscape of the digitizing world, as seen in both Deanna Marcum's article and the Waters and Kenney piece.

Is digitization alone sufficient for preservation? A positive answer should not be taken for granted. Willis argues for a hybrid system of preservation that uses both digitization and micrographics. Weber and Dörr agree with him. Their work (which also appeared in *Die Zeitschrift für Bibliothekswesen und Bibliographie*, volume 44, number 1, 1997) is important in part because it represents the first major German statement on the subject.

Douglas Miller's article is virtually a bibliography of digitization with a strong preservation emphasis. And those wanting to see the results of some digitization projects should consult the Library of Congress' "Electronic Texts and Publishing Resources" site, which has hyperlinks to collections.

Atkinson, Ross. "Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone." *Library Quarterly* 66:3 (1996): 239-265.

Atkinson's ideal is "a single, monolithic, international virtual library," which he admits "is probably impractical." (p. 262) His basic idea is that libraries take responsibility for publishing academic material in the new digital age. The "control zone" of the title refers to bringing works across a boundary from an "open zone" where there is no academic peer review and quality judgment. (p. 255) He argues that university "presses and the library are in fundamentally the same business of scholarly information exchange" and that "the sooner the library and the press can be amalgamated into a single administrative unit..." the sooner academic services can move online.

Conway, Paul L. "Digitizing Preservation." *Library Journal* 119 (1 February 1994): 42-45.

This article gives an overview of the state of the art for digital preservation and adds a few important caveats. One is the need to migrate digital images to new media and systems. Another is the need for database management to maintain a work's structural integrity. A third is the high-entry cost, which most libraries cannot afford.

Conway, Paul. *Preservation in the Digital World*. Washington, DC: Commission on Preservation and Access, March 1996.

This key work outlines Conway's five-part structure for looking at digital preservation. The parts include longevity, choice, quality, integrity, and access. He talks also about the importance of resource sharing among institutions, especially in selection, and the organizational changes needed to do the task.

Elkinton, Nancy E., ed. *Digital Imaging Technology for Preservation: Proceedings from an RLG Symposium Held March 17 and 18, 1994, Cornell University, Ithaca, NY*. Mountain View, CA: Research Libraries Group, 1994.

This book offers a digest of the state of knowledge about using digital imaging for preservation. The speakers include M. Stuart Lynn, who explains how digital preservation works in very simple terms; Anne Kenney and Paul Conway, who draw on their experiences at Cornell and Yale to describe the "preservation tool kit;" Pamela Mason, who discusses the scanning and use of an OCR; Peter Graham, who recounts his proposal for digital time stamping; and others. Donald Waters' closing remarks challenge universities and libraries to transform the nature of scholarly communi-

cation by bringing the "analog resources of the past into the electronic future." (p. 11)

Library of Congress. "Electronic Texts and Publishing Resources." Washington, DC: Library of Congress, 22 November 1996. <<http://lcweb.loc.gov/global/etext.html>>.

This listing includes electronic text collections, general resources, works by specific authors, government and legal documents, poetry sites, etext newsletters, commercial electronic booksellers, and electronic publishing and publishers.

Lynn, M. Stuart, et al. "Preservation and Access Technology: The Relationship between Digital and Other Media Conversion Processes: A Structured Glossary of Technical Terms." Washington, DC: Commission on Preservation and Access, 29 January 1997 (last update). <<http://palimpsest.stanford.edu/cpa/reports/lynn/>>.

This glossary offers substantial definitions and explanations of key technical terms in digital preservation, and it is a good reference source for those who encounter unfamiliar terms. Lynn structures the glossary in three parts: the original document, the selection process, and the preserved copy. Some entries are particularly detailed. The one on paper, for example, distinguishes among the effects of xerographic, thermographic, and photographic printing.

Marcum, Deanna B. "Digital Libraries: For Whom? For What?" *Journal of Academic Librarianship* 23:2 (March 1997): 81-84.

"Since it is unlikely that librarians will conclude that they are obsolete... " Marcum expects the number of digital library projects to grow. (p. 82) In 1995, the Commission on Preservation and Access, the Council on Library Resources, and a group of research libraries formed a federation for a "collaboratively managed, physically distributed, not-for-profit repository of digital information in support of instruction and research." (p. 83) The federation is concentrating on three issues: "discovery and retrieval mechanisms, intellectual property rights, and archiving." (p. 83)

Miller, Stephen Douglas. "Manuscripts and Archives in the Digital Age." University of Kentucky, College of Library and Information Science. Final Version—Revised January, 1996. <<http://www.duke.edu/~sdmiller/archives/manuscript.html>>.

This article provides a "snapshot of the current state of digital imaging and archives and how it compares with other preservation methods...." Is it virtually a bibliography on digital preservation with strong coverage of Project Open Book and the Cor-

nell/Xerox joint study. It also gives an unusual amount of attention to the University of Kentucky's Peal Manuscript Project. Among its notable features is a concise summary of the pros and cons of both microfilm and digital imaging.

Waters, Donald J., and Anne Kenney with the assistance of Lynne Personius, M. Stuart Lynn, and Millicent D. Abell. "The Digital Preservation Consortium Mission and Goals." Washington, DC: Commission on Preservation and Access, 24 March 1997 (last update). <[http://palimpsest.stanford.edu/cpa/reports/dpc\\_miss.html](http://palimpsest.stanford.edu/cpa/reports/dpc_miss.html)>.

This mission statement sets out a broad agenda for a consortium of seven libraries with active digital preservation projects. Among the goals are to verify the usefulness of digital imagery (including establishing its convertibility from paper to digital and back to paper for a stacks copy), to promote shared methods and standards (including copyright issues), to enlarge the base of materials, and to develop reliable and affordable access mechanisms.

Weber, Hartmut, and Marianne Dörr. *Digitization as a Method of Preservation*. Trans. by Andrew Medicott. Washington, DC: Commission on Preservation and Access, October 1997. 24p.

The authors discuss digitization as an alternative to microfilm and find it wanting. They emphasize the rate of technological change, which they contrast with the technological stability and simplicity of microfilm. But the authors accept digitization as a way of enhancing access.

Willis, Don. "A Hybrid Systems Approach to Preservation of Printed Materials." Washington, DC: Commission on Preservation and Access, 1992. <<http://www-cpa.stanford.edu/cpa/reports/willis/>>.

The advantages of micrographics are its durability, its relatively low cost, existing standards, and the fact that its readers "are not likely to become obsolete (all that is needed is light and magnification)." Its disadvantages are a vulnerability to scratches, a ten percent loss when making new copies, and trouble photographing certain color combinations. The advantages of digital images are the ease of access and transmission. The disadvantages are the high storage costs, the speed of obsolescence, the lack of standards, and the fact that "digital storage is not considered archival" because of the need for "periodic rewrite." Optical disk is a potential "permanent" storage medium. ASCII text has a limited usefulness because it only can represent characters.

## Sample Projects

If this bibliography included references to all the digitization projects, even all the important ones, it would be unacceptably long. Instead, references to only four projects are given. The CLASS project at Cornell and Project Open Book at Yale really define what we mean today by large-scale digitization projects. Both were done primarily for preservation, though with strong access components. Conway's study of costs is unique and invaluable for anyone thinking of negotiating with a digitizing vendor.

The article by Beavan, et al. on the Aberdeen Bestiary shows an alternative technology and puts a greater emphasis on providing access to rare material than to preservation. The Bestiary itself is a work of art well worth viewing on their Web site.

Almost everyone involved in digitization knows about the Library of Congress/Ameritech contest to provide digital images to the American Memory project. Below is a reference to the guidelines, as well as the crucial technical information, which set a de facto national standard.

Beavan, Iain, Michael Arnott, and Colin McLaren. "The Nature of the Beast; or, the Aberdeen Bestiary on the World Wide Web." *Library Hi Tech* 15:3/4 (1997): 50-55.

This project used 35mm slides and the Kodak PhotoCD process to digitize the Aberdeen Bestiary from c. 1200 AD. A plain English translation of the text accompanies the images. Web access is through 24-bit color JPEG images.

[Conway, Paul.] "Yale Study of Imaging Costs: Some Early Findings." <<http://gopher://arl.cni.org:70/arl/pubs/newsltr/182/yale.cost>> .

The site provides a summary of the per-volume cost at Yale for the digitizing of microfilm. The tables include the costs and some statistics. One important point is the dramatic impact of training. He also notes that most ARL libraries continue to preserve on film.

Conway, Paul. "Yale University Library's Project Open Book: Preliminary Research Findings." *D-Lib Magazine* (February 1996). <<http://www.dlib.org/dlib.february96/yale/02conway.html>> .

This article gives an overview of what Yale learned from Project Open Book after some years of production. Interesting notes include their decision "not to 'de-select' a particular title from a chosen subject cluster" when encountering technical problems and the decision to use binary rather than gray-scale scanning.

Kenney, Anne R. "The Cornell Digital to Microfilm Conversion Project: Final Report to NEH." *RLG DigiNews* 1:2 (16 August 1997). <<http://www.rlg.org/preserv/diginews/diginews2.html>> .

Cornell did a National Endowment for the Humanities-funded project that was complementary to Yale's Project Open Book: it created Computer Output Microfilm (COM) from digital images scanned from the original paper source. The project used 1,270 volumes (450,000 images) from nineteenth- and twentieth-century agricultural history. Among the important conclusions were 1) there was no detectable loss of resolution in recording on COM; 2) the image quality is better when scanning directly from paper than from microfilm (as Yale did); 3) the cost of scanning first and then creating COM appears to be less than filming first and then digitizing. The full text of the report is available at <<http://www.library.cornell.edu/preservation/com/comfin.html>> .

Kenney, Anne R., and Lynne Personius. "The Cornell/Xerox/Commission on Preservation and Access Joint Study in Digital Preservation." Washington, DC: Commission on Preservation and Access, 1991. <<http://www-cpa.stanford.edu/cpa/reports/joint/>> .

Five principal conclusions emerged from the study: digital imaging 1) provides a cost-effective alternative to photocopying; 2) offers a cost-effective alternative to microfilm preservation as long as the concept of "technology refreshing" is institutionalized; 3) has the potential to enhance access; and 4) can facilitate internal access and provide links to the library catalog through document control structures; and 5) that its infrastructure "supports other applications in the electronic dissemination of information."

Library of Congress and Ameritech. "National Digital Library Competition; 1997/98 Guidelines and Application Instructions." Washington, DC: Library of Congress, [1997?].

This site includes application forms plus some discussion of SGML, URN/PURL, page-turning scripts, finding aids, consortium definitions, restrictions, and requirements. A consortium that consists of an ARL and a non-ARL library will be judged in the non-ARL class. Classroom/educational ties are emphasized.

Library of Congress and Ameritech. "Related Technical Information; National Digital Library Competition." Washington, DC: Library of Congress, 26 August 1997. <<http://lcweb2.loc.gov/ammem/award/tech97.html>> .

This collection of useful materials for the contest includes documents prepared to supplement the guidelines; other background information; technical papers

(such as articles and formats); and related background reading.

Waters, Donald, and Shari Weaver. "The Organizational Phase of Project Open Book." Washington, DC: Commission on Preservation and Access, 1992. < <http://palimpsest.stanford.edu/cpa/reports/openbook.html> >.

This article lays out the plans for Project Open Book, and how it is different than Cornell's CLASS project. Key differences are the medium (digitizing microfilm) and the scope (it plans to do 10,000 books). Design principles, cost estimates, and some details from the Xerox proposal are included.

## Choice

The articles in this section focus on choosing materials that are appropriate for both preservation and access. Several struggles are evident. One is local needs versus the desire to build a comprehensive program such as the Great Collections plan, which influenced the Cornell-Yale approach to selection. Another is whether to choose only older materials that are in bad shape physically, or to include newer works that are merely inaccessible. Billings and Child deal with various aspects of a comprehensive program and its strengths and weaknesses. George offers a less centralized approach to cooperative selection planning. Cohen and Demas suggest approaches that are driven by user need or relative importance, rather than preservation.

Billings, Harold. "Library Collections and Distance Information: New Models of Collection Development for the 21<sup>st</sup> Century." *Journal of Library Administration* 24:1/2 (1996): 3-17.

He discusses an information-management organization that coordinates information sharing among consortium members. He also warns that print resources keep growing and that cutbacks are causing a narrowing of access to scholarly information that will make research libraries look "more and more alike over time." (p. 4)

Child, Margaret. "Selection for Preservation." *Advances in Preservation and Access* 1 (1992): 147-158.

Child discusses the Great Collections and other approaches to selecting materials for preservation (such as the vacuum cleaner approach). In 1987, a refinement was introduced (because of the scope of the effort) to consider the condition of individual items on the shelf.

Cohen, David J. "New Books from Old: A Proposal; Electronic Scanning of Out-of-Print Books; Presented at a Joint Meeting of SSP and NASIG, June 1992." *The Serials Librarian* 23:3/4 (1993): 149-155.

This article offers a proposal for a corporation called "IBID," which would scan and reprint out-of-print books and pay the copyright holders any royalties due. (p. 153-154) Cohen bases the technology on 1) Cornell's experience with scanning and printing on the Xerox DocuTech printer and 2) an analysis of Franklin and Marshall College Library's want list, which includes many works published within the last decade. He is particularly impressed with the quality of Cornell's re-created books.

Demas, Samuel. "Setting Preservation Priorities at Mann Library: A Disciplinary Approach." *Library Hi Tech* 12:3 (1994): 81-88.

"Preservation at Mann is based upon a conviction that preservation must systematically address the literature of disciplines, rather than focus on the holdings of specific libraries." (p. 83) Demas developed a disciplinary approach defining a core historical literature.

George, Gerald. *Difficult Choices: How Can Scholars Help Save Endangered Research Resources?* Washington, DC: Commission on Preservation and Access, July 1995.

This report looks at the recommendations of scholarly advisory committees in Renaissance studies, history, philosophy, medieval studies, modern language and literature, and art history, each of which has its own priorities. In looking at what scholarly involvement "should" mean, the report contrasts the CPA's "macro" model of national efforts with a Harvard group's "micro" approach, which emphasizes local needs and local collection policies.

## Quality

Quality is one of the most discussed and most misunderstood issues in digitization. Anyone digitizing primarily for access faces the temptation to do a quick and dirty job that suffices for the kind of size and resolution that is usable today on the Web. This means that for preservation purposes, or even for future higher-quality monitors and faster networks, the scanning could well have to be redone—a major expense. Those who choose a quick-and-dirty approach at least should understand what they are sacrificing.

What is a quality image? Cornell, as the first major digital preservation project, particularly struggled hard with this question. Existing microfilm-based preservation standards were both inappropriate and unattainable. Chapman and Kenney's notion of "full information capture" represents one of those major breakthroughs of understanding that may be hard to appreciate because, in retrospect, it seems so obviously

right and true. It does not mean a rigid 600 DPI standard, but rather a sliding scale that varies with the smallest part of the printed source. Battin's early article was also important in pushing for national acceptance of standards that were appropriate to digital preservation.

Achieving preservation quality is more complex than just buying the right scanner. The articles by Conway, Fleishhauer, Gartner, and Kenny and Chapman discuss quality issues in projects at Yale, Library of Congress, Oxford, and Cornell. The Lane article is important for understanding when to use GIF and when to use JPEG for display images. The Williams piece is an overview of quality issues.

Battin, Patricia. "Image Standards and Implications for Preservation," a talk presented at the Workshop on Electronic Texts, sponsored by the Library of Congress, 9-10 June 1992. <<http://palimpsest.stanford.edu/byauth/battin/imagestd.html>> .

Battin calls for standards that are appropriate for digital preservation and explains how existing preservation standards inhibit preservation activities. Some issues include fidelity of reproduction, differences between types of users, and costs. An important theme is that "it is more difficult to transform than to create," especially in going from brittle books to electronic form.

Chapman, Stephen, and Anne R. Kenney. "Digital Conversion of Research Library Materials." *D-Lib Magazine* (October 1996). <<http://www.dlib.org/dlib/october96/cornell/10chapman.html>> .

This article argues for "full information capture" so that the digital objects will be useful for as-yet undefined future needs. "The objective is not to scan at the highest resolution and bit depth possible, but to match the conversion process to the informational content of the original—no more, no less." It concludes that the costs of creating high-quality images will ultimately be less than the cost of low-quality images that fail to meet long-term needs.

Conway, Paul, and Shari Weaver. "The Setup Phase of Project Open Book." Washington, DC: Commission on Preservation and Access, 1994. <<http://palimpsest.stanford.edu/cpa/reports/conway.html>> .

This article looks at the second phase of Project Open Book where Yale set up and tested the tools. It discusses how to achieve the "highest possible quality of microfilm conversion," and gives specifics on resolution, density, and reduction ratio. Selection, indexing, and workflow management are also examined.

Fleishhauer, Carl, and Ricky L. Erway. "Reproduction-Quality Issues in a Digital Library System: Observations on the Reproduction of Various Library and Archival Material Formats for Access and Preservation," an American Memory White Paper. Washington, DC: Library of Congress. Draft 22 December 1992. <<http://rs7.loc.gov/pub/american.memory/white.papers/reprod.txt>> .

This paper considers the issues in digitizing different formats. "American Memory's experience suggests that preservation/access, high quality/low quality dichotomy arises for \*every\* format...." Some formats where digitization is particularly problematic are graphic arts, movies, and sound.

Fleishhauer, Carl. "Digital Formats for Content Reproductions." Washington, DC: Library of Congress, 20 August 1996. <<http://lcweb2.loc.gov/ammem/formats.html>> .

Fleishhauer gives information about sample images for thumbnail, reference, and archive, including tonal depth, format, compression, and spatial resolution. Digital artifacts include pictorial materials; textual materials reproduced as searchable text and images (using SGML); and textual material reproduced as images. He also talks about PDF, JBIG, "other proprietary" standards, and half-tone illustration problems.

Gartner, Richard. "Conservation by Numbers: Introducing Digital Imaging into Oxford University." *Microform Review* 23 (Spring 1994): 49-52.

This article describes a three-year digital imaging project at Oxford University. Gartner discusses some of the advantages of the digital format, particularly its "100% accuracy" in copying compared to the ten percent loss with microfilm. (p. 49) He also looks at digital storage options, including the JPEG format and Kodak's Photo-CD. Ultimately Oxford decided on two forms of storage, "deep" storage for the original files and "shallow" storage in compressed format for routine user access.

Kenney, Anne R., and Stephen Chapman. "Bitonal Scanning Means for Benchmarking Resolution Requirements." *Digital Imaging for Libraries and Archives*, 7-9. Ithaca, NY: Department of Preservation and Conservation, Cornell University Library, 1996.

This short section in Cornell's scanning workshop book gives formulas for calculating dpi requirements based on analogous formulas from microfilm preservation. There is also a table of heights in millimeters and quality index rankings (essentially marginal, medium, and high), which can be used to find the necessary dpi. This table is invaluable!

Lane, Tom. "Frequently Asked Questions about JPEG Image Compression." (9 June 1996). <<http://sul-server-2.stanford/mirrors/faq/jpeg/part1>> .

This set of questions and answers explains when to use JPEG for display images and when to use GIF. Basically JPEG is better for color and GIF is better for compressing black and white. There is other useful information about how to recognize a file format and lossless JPEG.

Williams, Don R. "Data Conversion: A Tutorial on Electronic Document Imaging." *Digital Imaging Technology for Preservation: Proceedings from an RLG Symposium Held on March 17 and 18, 1994*, 59-79. Mountain View, CA: Research Libraries Group, 1994.

Williams discusses the differences between traditional and electronic imaging, including topics like point versus line scanning, sampling, how many DPI (dots per inch) are adequate, moire problems, and tone.

### Longevity

One of the problems with digitization as a preservation method is that, even if storage media like CD-ROMs last for centuries, the hardware and software systems needed to read the CDs mutate at such a rapid rate that no one can count on them for more than half a decade. Hedstrom's article makes this point forcefully, as does the Rothenberg article. Waters offers the solution that is widely accepted among computing professionals: think in terms of life cycles, not permanence. It is admittedly an expensive solution, and one that takes a different mindset from that of most current preservationists. But a combination of refreshing the technology and reformatting the medium, which Lesk's 1992 report discusses, is the approach that Cornell has taken. Harvey's call to think in terms of short-term storage fits this solution well.

For a full understanding of this issue, several other articles are worth reading. Calmes gives a good overview of the strengths and weaknesses of the media. Lesk's 1990 report is interesting for its survey of storage media, which has not fundamentally changed, and for his mention of ASCII text, which is often ignored.

Calmes, Alan. "To Archive and Preserve: A Media Primer." *Inform* (May 1987): 14-17, 33.

Calmes discusses the preservation problems of paper, film, magnetic tape, magnetic disk, and optical disk.

Commission on Preservation and Access and the Research Libraries Group, Inc. "Preserving Digital Information; Report of the Task Force on Archiving

of Digital Information." Mountain View, CA: Research Libraries Group, [1996]. <<http://www.rlg.org/ArchTF/tfadi.index.htm>> .

The report recommends establishing a system of national digital archives, which will be licensed and responsible for long-term storage of digital documents. John Garrett and Don Waters were the principle authors.

Harvey, Ross. "From Digital Artifact to Digital Object." Canberra, Australia: National Library of Australia, November 1995. <<http://www.nla.gov.au/3/np0/conf/np095rh.html#roth>> .

Harvey concludes "that there are at present too many unknowns to commit digital data to currently available artifacts for anything other than short-term storage." He discusses the problems with tape and other storage media for long-term preservation.

Hedstrom, Margaret. "Digital Preservation: A Time Bomb for Digital Libraries," [1996?]. <<http://www.uky.edu/~kiernan/DL/hedstrom.html>> .

Digital recording media are vulnerable in a number of ways. The fact that they are short-lived (ten to 30 years) is less important than the problems of "obsolescence in retrieval and playback technologies" and "the absence of established standards." Among the standards needed are 1) how to establish authenticity and 2) what types of metadata to preserve.

Lesk, Michael. "Preservation of New Technology: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access." Washington, DC: Commission on Preservation and Access, October 1992 (last update: 13 August 1997). <<http://sul-server-2.stanford.edu/byauth/lesk/lesk2.html>> .

This report discusses the preservation of electronic materials including digitized books. One key point is that reformatting "will be a common way of life in the digital age." There is also a discussion of formats (including page images, where Lesk notes that "300 dpi is good enough for readability down to 5 or 6 point type, assuming that there are no halftones..."). Economies of scale in new media are mentioned in connection with computer-based archives—and an appendix contains an excellent list of existing archives. He also mentions Cornell's estimates for the cost of refreshing digital storage media.

Lesk, Michael. "Image Formats for Preservation and Access: A Report of the Technology Advisory Committee to the Commission on Preservation and Access." Washington, DC: Commission on Preservation and Access, July 1990 (last update: 13 August 1997).

<<http://sul-server-3.stanford.edu/byauth/lesk/lesk.html>> .

This somewhat dated report still has a useful overview of preservation alternatives, including chemical deacidification, microfilm, digital imagery, and ASCII. The ASCII alternative is rarely mentioned in preservation circles and is worth noting. There is also a discussion of storage media, including WORM, tape, CD-ROM, magnetic disk, optical disk, and digital paper.

Rothenberg, Jeff. "Ensuring the Longevity of Digital Documents." *Scientific American* (January 1995): 42-47.

Rothenberg discusses the problem of reading a CD-ROM discovered in 2045 that has no devices that can read it or software that can translate it.

Waters, Donald J. "Electronic Technologies and Preservation: Based on a Presentation to the Annual Meeting of the Research Libraries Group, June 25, 1992." Washington, DC: Commission on Preservation and Access. <<http://www.clir.org/cpa/reports/waters2.html>> .

This paper puts digital preservation in the broader context of the electronic technologies that belong to the "library of the future." Key enabling principles are "to think in terms of life cycles, not permanency," to simplify, to adopt an incremental approach, to formulate hypotheses, to build on standards, and to cooperate. Waters also quotes Hofstadter's law about estimating computer projects: "[i]t always takes longer than you expect, even when you take into account Hofstadter's law."

### Integrity

Integrity for digital artifacts is about whether one can be certain of their authenticity and accuracy. The Duranti article raises this question, and Waters and Garrett look at it from the perspective of an archivist wondering about provenance and fixity. Graham offers one solution through digital time-stamping. But digital time-stamping remains largely unimplemented. This section is short because little real effort has gone into solving the problem. That will change after the first scandal involving faked documents on the Web.

Duranti, Luciana. "Reliability and Authenticity: The Concepts and Their Implications." *Archivaria* (Spring 1995): 5-10.

Duranti discusses the idea of the authenticity of private records and raises questions about judging authenticity in electronic records.

Graham, Peter S. *Intellectual Preservation: Electronic Preservation of the Third Kind*. Washington, DC: Commission on Preservation and Access, [1994].

This pamphlet addresses a concern for preserving the version integrity of digital documents so that future readers can feel assured that they are viewing them unaltered. Graham identifies one potential solution as "digital time-stamping," which "calls upon a cryptographic technique of one-way hashing...." (p. 3) Several problems, however, remain to be worked out, including forms of bibliographic citation and the financial implications.

Waters, Don, and John Garrett. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Washington, DC: Commission on Preservation and Access and RLG, 1 May 1996.

The authors discuss reference, provenance, fixity, and content in relation to digital documents. They also note that their context includes the software and hardware to view them.

### Access

Access is the real reason many institutions undertake digitization projects. The articles selected here assume that the reader already knows that providing access is important. Hildreth frames the issue of access in the largest terms, arguing that the medium is not what is important. Conway makes the point that Web presence alone is not sufficient without the same quality of intellectual access libraries provide for other materials.

On the technical side of access, the Universal Resource Name (URN) and Persistent URL (PURL) are attempts to give permanence (or at least semi-permanence) to the shifting sands of Web locations. At present every time a server is renamed or even a folder changed, the access path is broken just as if someone deliberately misshelved a book in a multi-million-volume library.

Conway, Paul. "Selecting Microfilm for Digital Preservation: A Case Study from Project Open Book." *Library Resources & Technical Services* 40 (January 1996): 67-77.

He concludes: "The bottom line for all of us ... may well be that, without improvements in intellectual access to microfilm collections that support subject-oriented retrieval, digital conversion of these collections may prove to be quite feasible technically and quite untenable intellectually." (p. 75)

Hildreth, Charles R. "Preserving What We Really Want to Access, the Message, Not the Medium: Challenges

and Opportunities in the Digital Age." *Electronic Documents and Information: From Preservation to Access*. Ed. by Ahmed H. Helal and Joachim W. Weiss, 78-95. Veröffentlichungen der Universitätsbibliothek Essen 20. Essen, Germany: Universitätsbibliothek, 1996.

Hildreth argues that the questions librarians ask should be ones like "[h]ow will research and scholarship be conducted in the future?" rather than whether the library is digital or electronic or virtual. (p. 88) Both for print-based and digitally based information, the container—the medium—is not the point. It is the intellectual content—the message—that matters.

The URN Implementors [William Arms, Leslie Daigle, Ron Daniel, Dan LaLiberte, Michael Mealling, Keith Moore, and Stuart Weibel]. "Uniform Resource Names: A Progress Report." *D-Lib Magazine* (February 1996). <<http://www.dlib.org/dlib/february96/02arms.html>> .

The ideal of the URN is to give permanent names to online resources. Internet RFC 1317 defines the "Functional Requirements for Uniform Resource Names." Its requirements include global scope, global uniqueness, scalability, legacy support, extensibility, and independence. In the Khan/Wilensky model, a "handle" resolves to the name of the repository holding the resource. A sample syntax is <urn:hdl:cnri.dlib/august95> or <urn:inet:library.bigstate.edu:aj17mcc> . Other second-node options in addition to hdl are lifn, path, and inet, each of which has a slightly different format.

Library of Congress, National Digital Library Program. "The Relationship between URNs, Handles, and PURLs." Washington, DC: Library of Congress, 15 August 1997. <<http://lcweb2.loc.gov/ammem/award/docs/PURL-Handle.html>> .

A handle system is explained to be one implementation for a Uniform Resource Name, where the name is registered "in an approach comparable to ISBNs." PURL is an OCLC creation. It is a short-term solution that can be implemented now, but is "not fully location independent." The article includes a short bibliography.

## SGML

Although the emphasis in this bibliography is on the Cornell-Yale model's photographic approach to digitization, it is important to know something about the capabilities and requirements of SGML mark-up. Dixon's article is ideal for the SGML novice who wants to understand how it relates to HTML. Cole and Kazmer put SGML in historical context. Painter gives a good sense of the power of SGML and why an institution would go to the (considerable) trouble and

expense of document mark-up. The articles by Seaman of Virginia and Price-Wilkin of the University of Michigan give overviews of the kinds of projects that benefit from SGML and describe their approaches.

Brugger's article addresses one of the biggest problems with digitization: the capture and display of metadata. She shows how to capture metadata in SGML headers and looks at the relationship between the Text Encoding Initiative's (TEI) document-type definition and USMARC. It is a somewhat technical article, but a must-read for anyone actually doing SGML.

Brugger, Judith M. "Cataloging for Digital Libraries." *Cataloging and Classification Quarterly* 22:3/4 (1996): 59-73.

Brugger draws comparisons between TEI and USMARC. She also discusses the Stanford Integrated Digital Library Project and shows how to code SGML headers.

Cole, Timothy W., and Michelle M. Kazmer. "SGML as a Component of the Digital Library." *Library Hi Tech* 13:4 (1995): 75-90.

Cole and Kazmer discuss the printed page as a static artifact. SGML began in 1969 as GML and became SGML with ISO 8879:1986. The article gives a brief history, discusses authoring tools, and presents samples of SGML-marked pages. The list of implementation issues includes structure granularity, tag meaning, and entities not handled in the ASCII character set.

Dixon, Ross. "SGML and HTML: The Merging of Document Management and Electronic Document Publishing." *Information Management and Technology* 29:6 (November 1996): 251-254.

This article gives a brief description of SGML mark-up works with a DTD (Document Type Definition), both to lay out a document and to format it properly. Dixon makes the point that HTML has only one DTD. He also argues that SGML and HTML are aimed at different applications. SGML is "particularly suited to applications where documents have long lifetimes, are likely to be updated, exist in many versions/variants...." (p. 253) HTML "is more appropriate where documents have a short lifetime...and the document is to be published online—particularly over the World Wide Web." (p. 253)

Painter, Derrick. "The Oxford English Dictionary (OED): A Case Study: A Case Study in SGML." *Information Management and Technology* 29:2 (March 1994): 66-68.

Painter describes how the OED uses SGML to create its CD-ROM version. The DTD lays out a subdivided structure that is not apparent in the print

form except through different formatting. The dictionary really has become a database. SGML's value in constructing it was mainly in rendering the text independent of hardware, not in deconstructing the document. (p. 66)

Powell, Christina Kelleher, and Nigel Kerr. "SGML Creation and Delivery: The Humanities Text Initiative." *D-Lib Magazine* (July/August 1997). <<http://www.dlib.org/dlib/july97/humanities.07powell.html>> .

SGML makes it possible to search, for example, for the first line of a poem. The HTI is using a Xerox 620 scanner for batch processing and does not disbind volumes for the American Verse Project. They also are considering Prime OCR, which uses five OCR engines simultaneously to improve accuracy dramatically. Encoding uses the TEILite DTD and SoftQuad's editor. For the Corpus of Middle English, texts have been sent out for contract keying ("1 error in 20,000 characters"). The article also discusses cross-collection searching and multiple representations of the data.

Price-Wilkin, John. "Using the World-Wide Web to Deliver Complex Electronic Documents: Implications for Libraries." *Public Access Computer Systems Review* 5:3 (1994): 5-21. Also available on <[gopher://nfo.lib.uh.edu:70/00/articles/e-journals/uhlibrary/acreview/v5/n3/pricewil.5n3](http://gopher://nfo.lib.uh.edu:70/00/articles/e-journals/uhlibrary/acreview/v5/n3/pricewil.5n3)> .

This article emphasizes the ability of a marked-up transcript to be viewed many ways. He includes a description of major projects using SGML, including an archive of British poetry, an edition of the works of Dante Gabriel Rossetti, and others. He characterizes the "image collections" as "passive" (p. 11) and calls HTML an "impoverished tag set with little ability to reflect the complexities of the documents earlier discussed." (p. 13) On the interface, he admits that Boolean queries that "ask for the intersection of document structures have been challenging," but fill-out forms and menu selections have done better. (p. 18)

Seaman, David. "Of Books and Bytes: Electronic Texts at the University of Virginia Rare Book School." Charlottesville, VA: University of Virginia Library, 1996. <<http://etext.lib.virginia.edu/articles/watermark.html>> .

Seaman's short article describes the rare book school at the University of Virginia, where visiting scholars learn SGML at the E-Text Center. It includes a list of projects. The emphasis is on building a "searchable" online library.

Seaman, David. "Electronic Text Center Introduction to TEI and Guide to Document Preparation." Charlottesville, VA: University of Virginia, [1996?].

<<http://etext.lib.virginia.edu/tei/uvatei.html>> .

This document describes the mark-up procedures at the E-Text Center in detail, with examples of front matter, body, and back matter coding.

Seaman, David. "Special Collections Digital Image Creation." Charlottesville, VA: University of Virginia, June 1996. <<http://etext.lib.virginia.edu/helpsheets/specscan.html>> .

He recommends scanning at 400 to 600 dpi (400 is their default) and 24-bit color, even for grayscale drawings. JPEG copies are determined by size (300 to 500 dpi for better quality, less than 100 dpi for poorer quality). He suggests that a header saying how, why, and by whom the image was created be added into the binary code of the image file.

Seaman, David M. "Selection, Access and Control in a Library of Electronic Texts." *Cataloging and Classification Quarterly* 22:3/4 (1996): 75-84.

Seaman's motto is: "If it's not SGML, it's ephemeral." (p. 84) He sees SGML as a way to maintain long-term viability for digital texts. The TEI header serves as the basis for cataloging records. It contains 1) the file description (a full bibliographic description of the file including the size in KB); 2) the encoding description (i.e., how the text was normalized during transcription); 3) the text profile description (including non-bibliographic aspects, specific languages used, participants, and setting); and 4) the revision history (a record of changes during the development of the electronic text). (p. 80) It is important to have an accurate source of information available. He writes: "I am beginning to think that it is in our interest to scan as a digital image any title page we process." (p. 81)

## Copyright

No one should contemplate a digitization project, especially one where Web access is important, without a working knowledge of copyright law. Oakley lays out the issues in a clear and concise way that matches library concerns, but more importantly makes clear those areas of the law that are ambiguous and potentially perilous. Hersey looks at the copyright issue in contracting for digital information. The ARL publication puts the issue of fair use into a digital context. Be warned, however, that these three works represent only a fraction of the literature on copyright in the digital environment. The law and its interpretation shift with each new wind. This sample of articles suffices only to give a general understanding of the complexity of the problem.

Association of Research Libraries. "Copyright and Fair Use in Digital Environments." *ARL: A Bimonthly Newsletter of Research Library Issues and Actions* no. 192 (June 1997).

This article includes sections by Mary E. Jackson ("CONFU Concludes; ARL Rejects Guidelines"), David Green ("CONFU Continues? Is It Time to Regroup?"), Brian Nielsen ("Northwestern Affirms Fair Use Through Practice; Electronic Reserve Policy, System Developed"), Mary Case ("Educational Community Articulates Principles for Intellectual Property in the Digital Environment"), and Prudence Adler ("WIPO: Summary and Key Accomplishments").

Hersey, Karen. "Coping with Copyright and Beyond: New Challenges as the Library Goes Digital." *Cause/Effect* 18:4 (Winter 1995): 4-6.

Hersey discusses how the failure of copyright law leads to unequal contracts between information publishers and universities. MIT's experience with *Encyclopaedia Britannica* is cited. EB kept the data on their server to keep control during the negotiations.

Oakley, Robert L. "Copyright and Preservation: A Serious Problem in Need of a Thoughtful Solution." Washington, DC: Commission on Preservation and Access, September 1990. <<http://www-cpa.stanford.edu/cpa/reports/oakley/index.htm#toc>>.

Oakley offers a discussion of the old 1909 and the new 1976 law that went into effect in 1978. Among other points, he says that materials that were in the public domain before the new law do not come under its jurisdiction (i.e., materials published without copyright notice). For works under the new law, copyright protection lasts the life of the author plus 50 years. Anonymous works or works for hire are protected 75 years from publication or 100 years after creation, whichever comes first. The old law allowed 28 years with a 28-year renewal (but works not renewed during the first term are in the public domain).

#### NOTES

1. Paul Conway, *Preservation in the Digital World* (Washington, DC: Commission on Preservation and Access, March 1996), 6.
2. Conway, 8-9.

---

(Continued from page 118)

#### Web Sites

Alexander, Jan, and Marsha Tate. "Teaching Critical Evaluation Skills for World Wide Web Resources." <<http://www.science.widener.edu/~withers/webeval.htm>>.

Grassian, Esther. "Thinking Critically about World Wide Web Resources." <<http://www.ucla.edu/campus/computing/bruinonline/rainers/critical.html>>.

Kirk, Elizabeth. "Evaluating Information Found on the Internet." <<http://milton.mse.jhu.edu:8001/research/education/net.html>>.

Ormondroyd, Joan, et al. "How to Critically Analyze Information Sources." <<http://urisref.library.cornell.edu/skill26.htm>>.

Smith, Alastair. "Criteria for Evaluating of Internet Information Resources." <<http://www.vuw.ac.nz/~agsmith/evaln/index.htm>>.

Tillman, Hope. "Evaluating Quality on the Net." <<http://www.tiac.net/users/hhope/findqual.html>>.

#### Examples of Good Web Sites

BI-L is the premier listserv for instruction librarians. Subscribe: [listserv@bingvmbcc.binghamton.edu](mailto:listserv@bingvmbcc.binghamton.edu)

ACRL/Instruction Section Web site: <[http://www2/colgate.edu/instruction/](http://www2.colgate.edu/instruction/)>

LIRT (Library Instruction Roundtable) Web site: <<http://diogenes.baylor.edu/library/LIRT>>