DOCUMENT RESUME

ED 352 040 IR 054 262

AUTHOR	Kenney, Anne R.; Personius, Lynne K.
TITLE	The Cornell/Xerox Commission on Preservation and Access Joint Study in Digital Preservation. Report: Phase 1 (January 1990-December 1991). Digital Capture, Paper Facsimiles, and Network Access.
INSTITUTION	Cornell Univ., Ithaca, N.Y.; Xerox Corp., Rochester, N.Y.
SPONS AGENCY	Commission on Preservation and Access, Washington, DC.
PUB DATE	Sep 92
NOTE	57p.
AVAILABLE FROM	Commission on Preservation and Access, 1400 16th St., N.W., Suite 740, Washington, DC 20036-2117 ($10).
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Access to Information; Books; College Libraries; Comparative Analysis; Cost Effectiveness; Electronic Equipment; Higher Education; *Information Technology; *Information Transfer; Microforms; *Preservation; Reprography; Research Libraries
IDENTIFIERS	*Brittle Books; Cornell University NY; *Digital Scanning; Electronic Text; Xerox Corporation

ABSTRACT

The primary emphasis of this study of the use of digital technology to preserve library materials was the capture of brittle books as digital images and the production of printed paper facsimiles. Of equal interest, however, was the role of digital technology in providing access to library resources, and preliminary work in this area has also been accomplished. Based on extensive experimentation with one digital scanning system, the study reached five principal conclusions: (1) digital image technology provides an alternative--of comparable quality and lower cost--to photocopying for preserving deteriorating library materials; (2) subject to the resolution of certain problems, digital scanning technology offers a cost effective adjunct or alternative to microfilm preservation; (3) digital technology has the potential to enhance access to library materials; (4) through the implementation of document control structures, digital technology offers a means to facilitate access and to provide links between the library catalog and the material itself; and (5) the infrastructure developed for library preservation and access activities supports other applications in the electronic dissemination of information. This report is divided into three major sections: a description of the products developed to reach project goals; a review of the process of applying digital scanning technology to the preservation of and access to library materials; and a discussion of the findings. Four appendixes offer information on the comparative quality of paper and facsimile copies; a description of a cost study; assumptions of the cost study, including equipment, labor and materials costs; and a scanning diagram. (KRN)

ED352040

The Commission on Preservation and Access

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

JOINT STUDY IN DIGITAL PRESERVATION PHASE 1

A report to the Commission on Preservation and Access

by

Anne R. Kenney
and
Lynne K. Personius
Project Managers
Cornell University

September 1992

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Maxine Sitts

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

BEST COPY AVAILABLE

120542462

1400 16th Street, N.W., Suite 740, Washington, D.C. 20036-2217 • (202) 939-3400

A private, nonprofit organization acting on behalf of the nation's libraries, archives, and universities to develop and encourage collaborative strategies for preserving and providing access to the accumulated human record.

Published by
The Commission on Preservation and Access
1400 16th Street, NW, Suite 740
Washington, DC 20036-2117

September 1992

Reports issued by the Commission on Preservation and Access are intended to stimulate thought and discussion. They do not necessarily reflect the views of Commission members.

Additional copies are available from the above address for $10.00. Orders must be prepaid, with checks made payable to "The Commission on Preservation and Access," with payment in U.S. funds.

This paper has been submitted to the ERIC Clearinghouse on Information Resources.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences-Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

COPYRIGHT 1992 by the Commission on Preservation and Access. No part of this publication may be reproduced or transcribed in any form without permission of the publisher. Requests for reproduction for noncommercial purposes, including educational advancement, private study, or research will be granted. Full credit must be given to the author(s) and The Commission on Preservation and Access.

The Cornell/Xerox/Commission on Preservation and Access

JOINT STUDY IN DIGITAL PRESERVATION

REPORT: PHASE I

(January 1990—December 1991)

Digital Capture, Paper Facsimiles, and Network Access

Anne R. Kenney and Lynne K. Personius
Project Managers

The illustration is derived from a photograph by Charles Harrington and depicts Michael Friedman of the Cornell University Library staff scanning a brittle book using the Cornell/Xerox prototype system.

To order supplements, send check for $10.00 payable to Cornell University to:

Cornell University Library
Department of Preservation and Conservation
215 Olin Library
Ithaca, NY 14853
(607) 255-9440

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Cornell University acknowledges the support of the Commission on Preservation and Access and its Technology Assessment Advisory Committee. This support has been critical to the success and recognition of the Cornell/Xerox/CPA Joint Study in Digital Preservation.

The Joint Study involved the work of over sixty individuals from both Cornell University and Xerox Corporation. Within Cornell, collaboration was apparent at the highest level. Both M. Stuart Lynn, Vice President for Information Technologies (CIT), and Alain Seznec, University Librarian, gave the project their fullest support, which was crucial to its success. Stuart Lynn's position on both the Commission's Technology Assessment Advisory Committee and Xerox Corporation's University Advisory Panel was critical to bringing this project to Cornell.

Additional administrative support and advice came from Ross Atkinson, Assistant University Librarian for Collection Development and Preservation, and John F. Dean, Director of the Department of Preservation and Conservation.

The Cornell portion of the study was co-managed by Anne R. Kenney, Associate Director of the Library's Department of Preservation and Conservation, and Lynne K. Personius, Assistant Director of CIT for Scholarly Information Sources. William R. Turner, Senior Project Leader of the Library Information Technologies office, developed and tested the Print Request Server. Michael Friedman and Elizabeth Freedman served as scanning technicians; Elizabeth was replaced by Sue Poucher in July 1991. Their input into the functioning of the scanning workstation directly affected system design and testing. Through the course of this project, they worked with three different versions of hardware and over a dozen software upgrades.

Pete Baker, Computer Operations Supervisor, and Peggy Roberts, Laser Programmer, both of CIT, were responsible for the operation of the Docutech printer. Steve Rockey, Mathematics Librarian, and Michael Woodall, Assistant to the Mathematics Librarian, devoted innumerable hours to the selection of the first 535 volumes scanned. Their selections were reviewed by a faculty advisory committee, consisting of Keith Dennis, Chair of the Mathematics Department, Anil Nerode, Professor of Mathematics, and L. Pierce Williams, John Stambaugh Professor of the History of Science. The advisory committee also assisted in evaluating the quality of the paper facsimile and the utility of the prototype viewstation.

More than a dozen bibliographers and selectors chose the remaining volumes that were scanned during this project, an effort that was managed by Martha Hsu, Collection Development Office Coordinator. The development of cataloging procedures represented the work of four individuals: Ed Weissman, Catalog Librarian; Renee Chapman, Catalog Management and Authorities Librarian, who was succeeded by Judith Brugger; and Betsy Gamble, Head, Original Cataloging Unit.

Finally, a committee of Cornell librarians and computer professionals provided input into the design of the prototype viewstation and offered advice on the network-related aspects of the project.

Xerox Corporation's support group was equally significant, numbering 28 individuals in total. Xerox commitment included Corporate (Chuck Buchheit, VP Marketing, Dennis Andrews, VP Xsoft); Marketing (Hugh Jarrett, Glenn Alexander); Engineering Software Implementation (Eugene Evanitsky, Joe Hoey, Edwin Monkelbaan, William Crocca, Ben Barlow, Elizabeth Paradise, Ann Davidson, Greg Cholmondeley, William Anderson, Barry Gombert, Alan McReynolds, Richard Dimperio, Joseph Filion, Keith Emanual, Kelly Siggens, Martin Millner, Mike Powers); Scanner Engineering (Dick Tuhro, Dan Young, Colin Dodd, Mark Vannicola, John Walsh); and Soleil Engineering (Gerry Muto, Abde Kapadia).

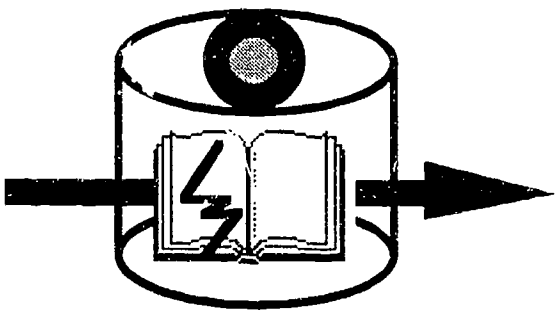
The project was initially brought to Xerox attention by Glenn Alexander. After many

internal conversations and management meetings, Dennis Andrews and Chuck Buchheit agreed to jointly sponsor the activity. The engineering team was recruited from within the former Advanced Systems Concepts group and began work in January of 1990. The marketing organization made the necessary arrangements for the equipment to be acquired, delivered, installed, and supported at Cornell. The scanner design team supplied us with engineering prototypes and supported those devices at the Cornell site.

The engineering software team worked with the Cornell group to understand the library application and to create software which provided the required services. The challenge was to understand what was required in those engineering terms necessary for software specification and implementation. Members of the team took a deep, personal interest in the success of the project at Cornell.

The Xerox Soleil Engineering group provided extensive support to the system development. They developed the image print path and the PostScript extensions now being used at Cornell.

This report represents the work of many of the individuals listed above. The project managers would especially like to thank: Stuart Lynn, Ross Atkinson, John Dean, Jim Harper, and William Turner for their help and encouragement.



I. EXECUTIVE SUMMARY

Cornell University and the Xerox Corporation, with the support of the Commission on Preservation and Access, have collaborated for the past two years in a joint Study to investigate the use of digital technology to preserve library materials. The primary emphasis of this study has been on the capture of brittle books as digital images and the production of printed paper facsimiles.[1] Of equal interest, however, has been the role of digital technology in providing networked access to library resources, and preliminary work in this area has also been accomplished.

The Joint Study has led to a number of conclusions regarding preservation, access, electronic technology, and the role of the library. In particular, participation in this study has convinced Cornell of the value of digital technology to preserve and make available research library materials. Such digital preservation presents a cost effective alternative to photocopying, and — subject to the resolution of certain remaining problems — a potential adjunct or alternative to microfilm preservation. The greatest promise of digital technology as a preservation option is to improve access to materials. Cornell expects to work with others to find ways to resolve the remaining issues surrounding the use of digital technology.

1 Digital image technology, for the purposes of this report, is defined as the electronic copying of scanned documents in image form. The text contained in these images in not converted to alphanumeric representation at the time of scanning, although the potential exists for such conversion, in whole or in part, from the digital files at some later time. The present capabilities of optical character recognition are inadequate for capturing both the information and the presentation of the original page, which is critical when replacing rapidly self-destructing books, especially when one considers the vast number of languages, illustrations, type faces, and printing techniques present in the collections of modern research libraries. The creation of digital images does not preclude the use of OCR capabilities. In fact it represents the first step in that direction--the scanning of paper copies to which character recognition can then be applied. See for instance: Stephen Smith and Craig Stanfill, "An Analysis of the Effects of Data Corruption on Text Retrieval Performance," (Thinking Machines Corporation, Cambridge, MA: December 14, 1988).

PRINCIPAL CONCLUSIONS

1. **Digital image technology provides an alternative—of comparable quality and lower cost—to photocopying for preserving deteriorating library materials.**

   The primary preservation benefits of the current state of this technology are image quality, duplication capabilities, paper output, and cost effectiveness. The intellectual content of a brittle book may be captured in a highly acceptable manner, and a hardcopy version produced on permanent, durable paper that replicates the presentation and format of the original, provided the original scanning is performed with sufficiently high resolution. The Joint Study indicates that the quality of scanning is competitive with photocopying and the costs are projected to be about 20 percent lower in a production environment (see Section V—Cost Study).[2]

   The high-quality paper facsimiles produced to replace the deteriorating originals have proven attractive to members of the Cornell research community and increased their support for the program. For the same reasons, this approach could generate support from the research community at large. At a fraction of the original scanning cost, additional paper copies can be "printed on demand" from the digitally-stored images at any time in the future.[3] The Joint Study suggests that digital images can also be duplicated without loss of fidelity and distributed widely across the nation's networks, providing remote access to other institutions either by local printing of facsimiles or by viewing at desktop workstations. There may also be opportunities to underwrite some of the costs of preservation through the sale of facsimile editions.

2. **Subject to the resolution of certain problems, digital scanning technology offers a cost effective adjunct or alternative to microfilm preservation.**

   In spite of the advantages of digital methods in providing high quality print facsimiles, improved access, and an economic alternative, some problems remain to be resolved before digital technology can fully compete with the preservation advantages of microfilm as an archival medium. In particular, the obsolescence associated with rapidly changing technologies causes concern that scanned digital images may not remain accessible over time. There is a need to "institutionalize" the concept of "technology refreshing," that is, the periodic copying of the digital images to new

---

2 The Joint Study compared the quality and costs associated with monochromatic scanning and photocopying only.

3 Conceivably, this may at some point allow librarians to propose other service alternatives as a substitute for traditional shelf storage.

formats and new storage media.[4] Furthermore, although this study's experiments suggest that digital images can, with no further loss of fidelity, be written to microfilm for archival backup purposes, the resolution used in the process does not match current preservation microfilm standards for material printed in typefaces below four point.[5] Because of the potential advantages of digital formats for access, however, it is important to pursue approaches that can resolve these issues.

3. Digital technology has the potential to enhance access to library materials.

Digital images may be transmitted over communications networks for use at remote locations. The Joint Study demonstrated the feasibility of such remote access through the delivery of digital images over the Cornell network for printing and for viewing on a prototype workstation. With network access, researchers will ultimately be able to use the resources of a library at any time from across the campus or across the country. As has been noted by many observers, access will no longer be defined geographically or temporally. It has also been noted that digital technology could spur the library to shift further along the continuum from physical ownership of materials to providing access to information regardless of its location. Much remains to be done, however, to define the architecture and develop the systems needed to support such remote access. This will be the focus of the next phase of Cornell's investigation in this area.

4. Through the implementation of document control structures, digital technology offers a means to facilitate access and to provide links between the library catalog and the material itself.

The Joint Study has recognized the need for a document control structure to facilitate navigation through the scanned digital book. Xerox has designed a flexible file format and indexing structure to facilitate direct access to the contents of individual books in the digital library. The architectural definition and early testing of this document control structure has been completed. So far, only

4 Contrary to the frequently expresses concern about the longevity of the physical storage medium itself, it is the obsolescence of standards, formats, and access software tools that is of greatest concern. The physical medium will normally long outlive these considerations.

5 This Report covers the period of the project ending December 31, 1991. Subsequent to this date, Cornell project staff have verified that digitally-produced microfilm produced by this project does not match microfilm preservation standards. This is not surprising given the scanning resolution. However, such microfilm may nevertheless be adequate for preserving texts produced at 4 point type and larger. In addition, early experiments suggest that halftone images can be scanned with resultant quality superior to that normally obtained with most production microfilming processes. Quality issues will be discussed in subsequent reports.

references to whole books are stored in the file. The details about each book have not yet been included. Once this information is stored in the document structure file, researchers at desktop workstations will be able to move from the on-line catalog record to a description of the content and structure of the book so as to assist them in determining whether the book meets their needs. This, in turn, will allow direct retrieval or printing of all or parts of the book as identified in the document structure file.

5. **The infrastructure developed for library preservation and access activities supports other applications in the electronic dissemination of information.**

   The library application can be broadened beyond an exclusive focus on preservation for use in storing, accessing , and distributing other information. For example, the Cornell Campus Store has started to use the system to produce customized course packs and class notes, and Cornell University Press plans to test its application to a reprint series. Electronic journal production and dissemination are also being investigated. Application to other areas helps to ensure that the substantial investments required in technology development can be amortized over a variety of projects.

These general conclusions of the Joint Study derive from extensive experimentation with one digital scanning system. This report describes the findings, as well as the assumptions underlying the study, and the products and processes adopted. While many issues surrounding preservation, access, the library's role, and electronic technology remain to be resolved, this study represents a solid beginning. It also represents the initial phase of Cornell's continuing investigation in the use of digital technology. The second phase, again funded by the Commission on Preservation and Access, establishes a testbed to further the exploration and use of digital technology to meet library preservation needs.

II. GUIDING PRINCIPLES

The Joint Study was guided from the outset by four principles: a preservation program using digital technology has certain requirements; access is as important as preservation; the use of an emerging technology requires collaborative expertise; and only practical, widely applicable processes that include a recognition of the importance of standards and product availability should be used.

National preservation requirements were considered in the Joint Study.[6] The study assumed that the use of digital technology must be both cost effective and result in products of sufficiently high quality to be considered viable for the preservation of deteriorating library materials. A major focus of the study was the development of a system that would optimize the time spent scanning a book while producing a suitable paper replacement for it. The scanning and printing system exploited in this study enabled technicians to scan at a level of quality and a rate comparable to photocopying and to create a digital print master for producing high-quality paper facsimiles at low cost. The study evaluated the long-term storage implications for the digital "preservation master," including refreshing and periodic recopying, and the feasibility of creating microfilm from the digital files.

The second guiding principle was that convenient access to preserved library materials is essential. For many disciplines a paper version of the work remains the medium of choice,[7] and for those disciplines, print-on-demand is a strategy that will enhance access. From the beginning, this project was designed as a networked application for the creation, storage, and transmission of digital images. The effect of enhanced access on research use and methods was not studied. Such evaluation may require the development of a critical mass of material in digital image form that is easily available to researchers.

6 The current national preservation program to preserve brittle material is based on the replacement of originals with copies that faithfully capture their intellectual content, including text, illustrations, and presentation. In order to preserve the largest number of items possible, the time spent in copying material should occur just once and should result in the production of a print master that can be used to make subsequent copies at lower costs. Information about the availability of copies should be widely publicized and included in the national on-line bibliographic databases. Finally, a preservation master of the original should be stored and maintained in a manner that will guarantee its long-term availability.

7 For instance, in the field of mathematics from which over half of the materials were selected, users "object to the inconvenience of microfilm, especially for monographs...Hardcopy reformatting (through photocopying) of older monographs is the preferred way to provide access in many libraries." Constance C. Gould and Karla Pearce, Information Needs in the Sciences: An Assessment, (Mountain View, CA: Research Libraries Group, Inc., 1991), pp. 65-68.

The third guiding principle was that the investigation into the use of digital technology in a library setting could only be undertaken with a spirit of collaboration. All participants understood that no one organization possessed the experience and expertise to establish digital technology as a preservation option. Cornell University relied heavily on its development partner, the Xerox Corporation,[8] and on the support of the Commission on Preservation and Access. The project also relied heavily on internal collaboration at Cornell between the University Library and Cornell Information Technologies. This study has convinced Cornell that extensive collaboration is critical to further investigation of digital technology, and should include the participation of: other research libraries and university technology organizations, groups that facilitate national programs, standard-setting organizations, technology vendors, funding bodies, and service bureaus that provide scanning, filming, or other services.

In keeping with the spirit of collaboration and cooperation, every attempt was made to use standards that promote future exchange of digital material among libraries and scholars. For example, proprietary image file formats, data compression standards, and network protocols were avoided in the design and architecture of the system. Image files from this project can be made available for use by commercial software packages running on standard hardware systems.

An important guiding principle was to rely to the extent possible only on technology that was readily available as product or near product and that was developed for broader marketplaces. This decision should ensure that the technology will be widely accessible to other institutions, and that its continued existence does not depend solely upon library applications. It also helps to ensure the continuing support of the system developed at Cornell.

8 For a Xerox Corporation perspective on the importance of co-development, see William Anderson, William Crocca, and Steven Barley, "Customer Co-Development: The Cornell/Xerox Joint Study Project Interim Report," PARC Technical Report SSL-91-139.

III. PRODUCTS

The first phase of the Joint Study was successfully concluded in December 1991 and substantially achieved the original goals of the Workplan of May 10, 1990. Products of the Joint Study that support these goals include the following:

A. the development, implementation, and testing of a **Networked Scanning System** for creating, storing, printing, and accessing digital images;

B. the creation of an **Electronic Library** consisting of the digital files for 950 deteriorating volumes;

C. the production of acceptable **Paper Facsimiles** to replace each of the original volumes by remote printing across the network;

D. the **Cataloging** of both the digital files and the paper facsimiles in the national and local on-line bibliographic databases;

E. the initial definition of the **Document Control Structure** to provide access points beyond basic bibliographic information;

F. the design of a **Print Request Server** that will enable researchers to obtain a print on demand version of any scanned volume;

G. the prototype implementation of **Electronic Access** to portions of the digital library from a distant workstation attached to the network;

H. an investigation of the feasibility of producing **Microfilm** directly from the digital files.

These products have contributed to the development of a scanning application suited to the preservation of research library materials. They are summarized in what follows. More details of the scanning and other processes are given in the following section on Process.

A. NETWORKED SCANNING SYSTEM

A networked system for creating, storing, printing, and accessing an electronic library was developed and tested. This system allows for the distribution of various functions to a number of locations served by a

high speed network. Among other things, the system represents the first step in providing user access at remote locations to library materials.

Xerox Corporation and Cornell University have developed the College Library Access and Storage System (CLASS) to meet preservation reformatting needs. The CLASS system is designed as a network compatible, distributed system composed of scanning workstations, that are based on IBM PC or PC-compatible workstations running DOS, high speed/high resolution scanners, and application software; an optical storage server composed of a UNIX based server with locator database pointing to an optical jukebox; a print server consisting of a Xerox Docutech printer and network conversion server; a print request server running as an X-Windows application on a SUN Sparcstation; and a prototype viewing workstation running on an IBM PC under DOS and Microsoft Windows (clients for other environments are under development).

The transition from a standalone scanning workstation to a fully networked system has required a significant investment of time, money, and expertise. The system in place at the conclusion of this study meets the design goals of this architecture, although it is still in an early state, with ongoing changes being made to improve reliability. Xerox Corporation has prepared a project time-line highlighting the development of the various components of the CLASS System (Supplement I).

The architectural design of the networked system is predicated on a client/server environment in which the geographical proximity of system components, information, and reader is not important. The development of high bandwidth networks provides an opportunity to transfer large amounts of information quickly, meeting the needs of this application. The components of the system are distributed across Cornell's campuswide TCP/IP network that forms part of the worldwide Internet. A representation of the system architecture is shown in Figure 1.

The system developed in this study creates files of bitmapped images that represent pages of books. These files are being stored on a large capacity optical jukebox. Until recently, the cost of storage for such files would have been prohibitive, but new information storage devices and larger capacity media make the use of digital image files possible, with every indication that technology costs will follow their historic decline well into the future.

An integral component of the study has been the use of the networked Xerox Docutech printer, a high-speed graphic printer that uses digital images to format printed pages. The network server that connects

Docutech to the Cornell TCP/IP network accepts print requests for compressed images or encoded documents (e.g., ASCII, PostScript, Interpress). On-line Docutech finishing hardware offers binding and stitching options that support on-demand printing.

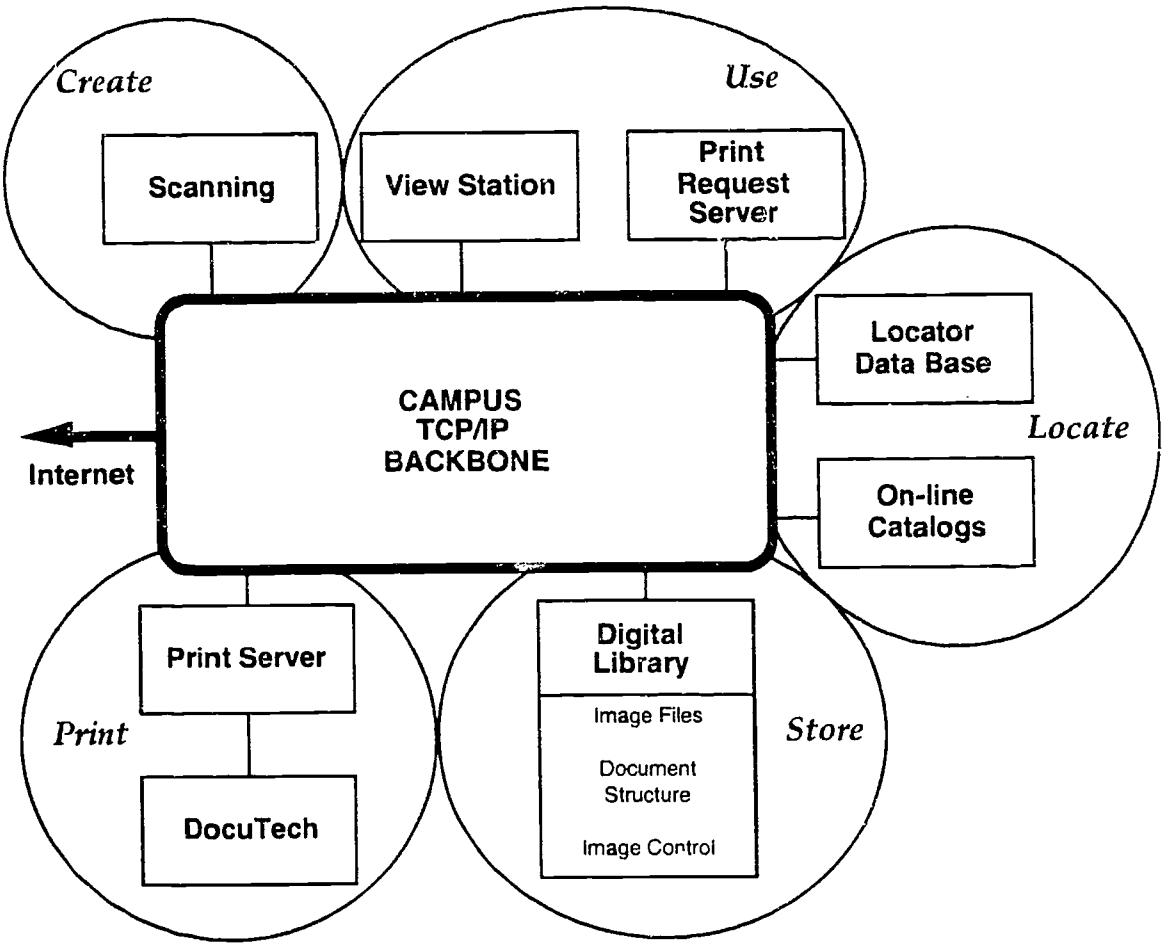


FIGURE 1. 1991 SYSTEM ARCHITECTURE

B. THE ELECTRONIC LIBRARY: IMAGE CAPTURE AND FILE FORMAT

Digital versions of 950 books were created The University intends to maintain and expand this embryonic electronic library and make it accessible to the broad national and international research community. The electronic library will be periodically copied to conform with newer technologies and standards. It will also serve as the basis for an experimental testbed by which further studies of storage, distribution, and access technologies may be evaluated.

At the end of this phase of the project, the electronic library contained approximately 285,000 digital files, each file representing a page of a book.[9] Each page was scanned as a bitmapped image and stored at a resolution of 600 dpi (dots per inch). The prototype scanner captures images using a complex scanning and interpolation scheme. The nominal scanning resolution of the scanner used in the Joint Study was 600 dpi. However, the definition of scanning resolution is not straightforward, and an explanation is required.

In fact, a 400 x 400 dpi aperture is used. A single scanline in the fast direction (across the platen) is sampled at 400 dpi, with 256 levels of grey (8 bits). This grey-level scanline is in turn sampled at 600 dpi. Thus, three 600 dpi samples are derived from two 400 dpi samples. The two end samples are directly converted to bits (a 1 or a 0) according to a thresholding algorithm; the two equal parts of the middle sample are averaged before thresholding. Thus, a single scanline 1/400th of an inch wide contains 600 bits (1's or 0's).

In the long direction (along the platen), this process is repeated 600 times per inch. Thus, information from overlapping scanlines, each 1/400th of an inch wide, is obtained. The result is a 600x600 dpi bitmapped image. (See diagram, Appendix IV.)

The scanner used is under development by Xerox and is not yet generally available in the marketplace. It represents an effective compromise among speed, resolution, and quality. Although higher resolution scanners are available, they are too slow with today's technology to be competitive in a production environment. No doubt, this will change with time.

The files were encoded in Aldus/Microsoft TIFF Version 5.0, which meets the new Internet Engineering Task Force standard definition for exchange of black and white images within the Internet.[10] TIFF (tagged image file format) provides a means for labeling a file so that it can be deciphered by application software, thus making it possible to exchange files among applications.[11] The files were compressed prior to storage

---

9 One thousand books were chosen for scanning. Fifty of the most heavily illustrated ones have been reserved for scanning using the windowing capabilities recently developed by Xerox.

10 Katz, A. & Cohen, D. Network FAX Working Group of the Internet Engineering Task Force, A File Format for the Exchange of Images in the Internet, Request for Comments number 1314, April 1992.

11 Digital files must be created in a manner that provides users with instructions on how to gain access to the information contained in them. It is one thing to store information on a disk, and another to gain access to it. Material can not be considered preserved if one can not "read" it. Thus a file must contain documentation on its format. Though there are many competing file formats, TIFF is in wide use. Unfortunately there are multiple TIFF formats, but a committee currently exists to address this issue. Today TIFF comes close to

and transmission using facsimile compatible CCITT Group 4 compression.[12] Because the images are binary representations, the compression algorithms resulted in considerable storage and transmission economy as well as a lossless means for compressing and decompressing the files. In this study, image compression has resulted in an approximate 40:1 reduction in file size. Even compressed, however, the digital files are large. File size varies depending on the size of the page and its content. An average 6" X 9" page composed of black and white text requires approximately 50 kilobytes of storage when compressed. A page of similar size that contains a halftone image can yield a file as much as ten times as great.

The files were stored on removable optical disks pending the development of software to store the images on the optical jukebox. At the end of Phase 1 covered by this Report, the images had not been fully transferred to the jukebox. Testing was in process.

C. PAPER FACSIMILES

Digital images were used to create hardcopy facsimiles for each of the volumes scanned in this project. The paper output is considered of sufficient quality and durability that the facsimiles serve as replacements for the deteriorating originals.

A primary goal of the Joint Study was to evaluate the paper output from the Xerox Docutech printer. The quality of the paper copy is very high: there is less than 1% variation in print size from the original; skew results only when the page trim is not parallel to text; front to back registration is reproduced within 1/100th of an inch of the original; the contrast between text and background is sharp; and the 600 dpi resolution compares favorably with the capture capabilities of photocopy. Illustrated material is exceptionally well rendered. As the copies are printed on paper that meets the ANSI standards for permanence, and the Docutech printer meets the machine and toner requirements for proper adhesion of print to page, the product is considered to be the archival equivalent of preservation photocopy.[13]

representing an industry standard. Aldus Corporation and Microsoft Corporation, "Tag Image File Specification Revision 5.0" (Aldus/Microsoft Technical memorandum, August 1988).

12 The International Telegraph and Telephone Consultative Committee (CCITT) has originated two algorithms, Group 3 and Group 4, that are in wide use for black and white images.

13 Norvell M.M. Jones, Archival Copies of Thermofax, Verifax, and Other Unstable Records. National Archives Technical Information Paper No. 5 (Washington: National Archives and Records Administration, 1990). ANSI Standard Z39.48-1984, currently being revised, covers the requirements for permanent/durable paper. See also RLG Preservation Manual (1986) and the Reproduction of Library Materials (ALA) draft photocopy guidelines of the Subcommittee on Preservation Photocopying Guidelines. The

Library staff and faculty advisors evaluated the quality of the paper product. Their subjective approval was a critical factor in the decision to replace the rapidly self-destructing books with the paper facsimiles. In most cases the original volumes were discarded after being scanned. A discussion of the quality achieved in this project and a comparison with light-lens processes is found in Section V, Findings.

D. BIBLIOGRAPHIC ACCESS

The digital books were cataloged in Cornell's local on-line catalog and in the Research Library Group's national database (RLIN). Although existing cataloging conventions were followed, some modifications will be necessary if digital technology is to become an accepted preservation format. Issues still to be resolved include what additional technical information is required to facilitate access, how preservation information should be conveyed, and what links can be drawn between the catalog records and other forms of indexing to the digital book.

In determining how to represent the digital files, catalogers started from the premise that the digitized book is a preservation product analogous to a preservation microform, and its treatment in the catalog record should be parallel. Sample records and a report on cataloging considerations and instructions by Judith Brugger, Catalog Management and Authorities Librarian, are located in Supplement II. The use of digital technology as a preservation medium, however, is a new concept in cataloging. Unlike microfilm reproductions, computer files are not accorded a "preservation reproduction" status. For instance, Chapter 9, the computer file chapter of Anglo-American Cataloging Rules, 2nd Edition, assumes that all items being cataloged are originally produced in machine-readable format. Computer files for digital books, therefore, are generally considered new editions, rather than versions of the original volume.

This interpretation precludes the parallel treatment of computer files and microforms. In 1980, the Research Libraries Group, Inc., with funding from the National Endowment for the Humanities, developed modifications to RLIN that resulted in the current system capabilities to highlight microfilm generational information and to display institutional decisions to reformat particular titles (the so-called "queuing" function). The latter capability was designed to assist institutions in avoiding duplicative filming efforts. Unfortunately in this study, Cornell was unable to "queue" records for titles to be scanned, nor were other institutions aware that a title had been preserved by Cornell when they searched for preservation

guidelines currently available for preservation photocopying place greater emphasis on image stability and paper permanence than image quality.

replacements. Moreover, if paper replacements had not been prepared and catalogued, there would have been no RLIN record in the book file indicating the availability of paper copies and digital versions on demand. Records for the digital book appear only in the computer file of RLIN, which is not normally searched by an institution looking for a replacement for a brittle book.[14]

In the future, Cornell would like to see enhancements to the Research Library Group's RLIN (Research Library Information Network) to record preservation information for material that has been reformatted using digital technology. These enhancements may be addressed by the reorganization and redefinition of certain data elements in the MARC record and by the movement toward format integration and/or a multiple versions approach.[15]

A cataloging issue still to be resolved involves the links to be drawn between the basic bibliographic record and other forms of indexing. The catalog record must carry information regarding the document structure file that accompanies the image files (see next section). Currently, both the call number field and the local notes field (USMARC 590) are reserved to record information on how to request a printed copy and ultimately to view the digital files directly. The means to assure a smooth transition from bibliographic record, to indexes, and finally to the electronic library has yet to be developed and tested.

E. DOCUMENT CONTROL STRUCTURE

A document structure is used to organize the individual images captured during the scanning process. It will also be used to provide direct access to components of the book. The arrangement of a physical book provides information to readers. For instance, the table of contents and the index are placed so that they can be easily found and used by any reader. The document structure file is designed to assist a reader in using an electronic version of the book.

Requirements for the document control structure have been defined and a prototype created for a number of books. Cornell recommends a collaborative process involving other libraries and consortia to define further the document control structure and to establish it in a standardized form for use in digital libraries of multiple institutions.

---

14 Cornell did prepare a Preservation Scope Note for the mathematics material which appears in the RLIN Conspectus. Preservation Scope Notes provide RLG and individual institutions with information about large preservation projects, both in progress and completed, to assist in the planning and coordination of preservation activities.

15 Format Integration and Its Effect on the USMARC Bibliographic Format, Library of Congress, 1988. Prepared by Network Development and MARC Standards Office.

Because the digital images comprising a book are not text-searchable, there is a need to find easy ways for users to search and reference the major parts of each book to facilitate access. For example, the page numbers printed on the originals have to be incorporated into the document structure and correlated with the image file numbers so that a request to retrieve a particular page number recalls the image with that number printed on it.

The digital book as currently configured consists of two parts. First, individual pages are stored as a collection of discrete bitmapped images. Second, the document structure links the images into a single document. A database entry for each document also contains descriptive information such as the author, title, and document identification number. Further enhancements to the document structure will allow references to the major parts of the book, such as table of contents, chapters, indexes, and so forth.

The creation and storage of the document structure are critical to the system design. Xerox Corporation has produced detailed specifications for the database and the software to implement the document structure architecture. At this time, these are described only in internal Xerox project reports.

Although a file exists with the elements defined to hold structure data, only the most basic structure information (the order of the files) has been collected. For the purposes of testing the print request server and the view station described below, complete document structure records for a small number of books were created. Cornell is continuing the process of software testing and development.

F. PRINT ON DEMAND ACCESS

The distributed design of the CLASS system has allowed Cornell to separate physically the scanning from the printing and storage functions. Cornell staff members are developing a print request server that will enable researchers to request from their offices printed copies of documents stored in the digital library. At the close of this phase of the study, individual requests are being handled by the scanning technicians, and the Print Request Server is undergoing initial testing.

A prototype print request server has been developed and tested that simulates the process of identifying relevant material and initiating a print request. Functional specifications for the print request server are located in Supplement III. For the print request server to be an integrated part of the distributed digital library, it must interact with several components of the CLASS system. Since Xerox is still in the process of developing some of these, Cornell decided to implement

temporary alternatives for some database information. For instance, the request server is designed to communicate with the image server to retrieve the document structure. Only page-linking information has been recorded in the Document Control Structure. A fully functional request server also depends on the development of a complete Docutech job ticket to record information on requests for printing. As of this writing, not all job ticket information had been specified for use by the request server. Thus the request server is still a prototype, and work will continue on its development next year.

G. ELECTRONIC ACCESS

A prototype view station which can be used from anywhere on the network provides electronic access to the digital library. The view station retrieves and displays electronic books stored in the digital library. Further work must be done to develop view station software that is suitable for use by library patrons. However, the feasibility of viewing books remotely has been established.

The prototype view station used in the study was developed by Xerox with design advice provided by a committee of Cornell librarians and computer professionals. The view station software represents the first step in providing a level of electronic browsing and retrieval at the desktop. The view station offers all the search, retrieval, and printing functions present on the scanning workstation, but without any updating or editing capabilities. Some enhancements were added to facilitate navigating through multiple documents.

Using the view station, a library patron can search the digital library by author and title, with the results being displayed in a window. Images of pages are then displayed, with readability a function of the page size of the original, the window size, and the screen resolution of the monitor. The images were found to be readable by most users. The option of enlarging the page to fill the screen and a zoom feature make it easier to read small text. Books originally consisting of pages no larger than 6"X 9" are easily read on screen. Larger texts, which must be scaled down to fit the screen, are more difficult to read without enlarging. In general, the quality of the on-screen image proved acceptable if screen viewing is used primarily for rapid browsing and retrieval. For extended reading, a print-on-demand request of the 600 dpi digital images provides a workable use copy. [16]

16 Performance issues associated with reading material from the network will be addressed in the Testbed Project, begun in January 1992.

H. DIGITAL-TO-MICROFILM FEASIBILITY

Microfilm can be produced directly from the digital files. The advantage in producing film is that it can serve as the preservation backup for an emerging technology. In the unlikely event the digital files were to become unreadable, the microfilm could be scanned and digitized at a fraction of the costs of initial capture. While preliminary experiments in this area were performed with promising results, the complementary roles of digital technology and microfilm require further examination.

Cornell has conducted preliminary tests to establish the feasibility of producing microfilm from the high resolution digital files. The first test, conducted in September 1991, was relatively modest. The digital images for several pages were transferred to magnetic tape and sent to Image Graphics, Inc of Shelton, Connecticut for output onto microfilm using Image Graphics' MICROGRAPHICS EBR SYSTEM 3000, an electron beam recorder.

Image Graphics successfully recorded the digital images at 10X reduction on 70mm, non-perforated KODAK Direct Electron Recording Film, SO-219, a film designed expressly for use in recorders that expose film by means of an electronic beam brought to bear directly on the emulsion.[17] The company produced both negative and positive versions of the film. Density readings on the negative version averaged .95. The positive film was inspected at Cornell on a light box and a microfilm reader. The images contained in the test strip were crisp, with sharp contrast between text and background. More significantly, the quality of the resulting images was faithful to the quality obtained in the digital files as represented in the paper copies. While issues of quality control that center on film base, processing, and resolution were not evaluated, the results appear promising, especially for illustrated material where the potential to create a high quality reproduction favors digital technology.

In late fall 1991, the digital files for a 70 page volume that contains halftones and other illustrations was sent to Image Graphics to produce microfilm. The film was not completed in time for evaluation under this phase of the project[18], but it will be subject to full technical and bibliographic inspection. A standard microfilm version for the same volume has been prepared for comparison purposes. Both copies will

17 The film emulsion layer is unusually thin and characterized by extremely fine grains and a relatively high silver to gel ratio; the support is ESTAR base, a clear 4-mil polyester film. Based on discussions with technical experts at Kodak and University Microfilms, it appears that the archival properties of the SO-219 are questionable. Image Graphics is investigating the use of Image Link film for subsequent tests.

18 Subsequent to the close of Phase 1, the microfilm was indeed produced. The quality will be discussed in subsequent reports.

be evaluated, principally to determine how the digital film compares in quality and technical specifications to the light lens microfilm. Yale University's proposed project to convert large quantities of preserved library materials from microfilm to digital images will provide valuable comparative data on the means, costs, and benefits involved.[19] The issue of image quality should also be studied carefully. Cornell will continue to investigate the process and costs of creating microfilm from the digital files.

19 Donald J. Waters, From Microfilm to Digital Imagery. On the feasibility of a project to study means, costs, and benefits of converting large quantities of preserved library materials from microfilm to digital images (Washington: The Commission on Preservation and Access, 1991).

IV. PROCESS

In addition to the development of a scanning system, the study resulted in the adoption of a process that applies digital scanning technology to preservation and access of library materials. The process parallels in many respects that used in preservation microfilming or photocopying projects.

A. SELECTION

Material representing a wide range of subjects was selected for this study. The first 535 volumes came from Cornell's Mathematics Library, representing the period 1850-1916.[20] These materials were chosen for a number of reasons: the Cornell mathematics collection is especially strong; mathematics monographs from 1850 on are considered of current scholarly interest; the material is in poor physical condition and had been identified as one of the library's highest preservation priorities; potential users are technically sophisticated (2/3 of Cornell's Math faculty have Sun or equivalent workstations); the material falls outside of copyright restrictions; and very few libraries nationwide have strong retrospective holdings in this subject area. Further, the mathematics faculty had determined that these books had to be replaced in paper form, so that many of the volumes had been scheduled for preservation photocopy.

The mathematics monographs chosen included the works of significant authors and those individual titles that have contributed substantially to the development of the discipline. All were in need of preservation, and at the time of selection, had not been reprinted or microfilmed. A faculty advisory committee reviewed the selections made by the Mathematics Librarian and his assistant. The advisory committee also assisted in the evaluation of the quality of the paper copy produced from the digital files.

Cornell bibliographers, representing the sciences, social sciences, humanities, and various area collections, selected the remaining 415 volumes. They chose items primarily on the basis of their deteriorating condition that were representative of a cross section of materials typically found in research libraries. Items covered by copyright or which had been microfilmed or recently reprinted were excluded.

---

20 The selection process is described by Steven Rockey in "The Cornell-Xerox-CPA Project to Digitally Reformat Books," paper presented to the AMS/MAA Joint Mathematics Meetings, Baltimore, MD, January 8-11, 1992. A bibliography of the mathematics books preserved in this project is included as Appendix VII. A bibliography of all volumes scanned in this project can be prepared by conducting a search on RLIN using the Series Note ("CXJSP"), and downloading the on-line records.

Selection decisions were also guided by the limitations of the prototype scanning system, and adherence to current compression standards. For instance, all items had to be disbound and trimmed and the page size could not exceed 8 1/2 X 11 inches, including the dimensions of foldouts.

B. PREPARATION

In addition to the actual scanning, two project staff technicians performed all pre- and post-scanning functions. They collated each item to assure completeness, repaired torn pages, and ordered replacements for missing pages through interlibrary loan. Annotations and marginalia that did not obscure text were left intact. Bibliographers decided whether the technicians should attempt to capture or delete this information during scanning.

The volumes were then disbound[21] and the binder's edge trimmed parallel to the text. The scanning technicians prepared a worksheet for each volume, recording bibliographic information, physical description, document control structure information, the scanning settings used, and basic workflow. For the last three months of the project, they also recorded time spent on the various scanning functions.

C. SET UP

Prior to actual scanning, the technicians performed a variety of setup functions, using the CLASS software in quality control mode. These included keying in primary bibliographic data, defining the page size and page trim, establishing front to back registration, and scanning sample pages for on-screen review to identify a default range of settings for the entire volume. Scanning settings included choosing an image type (line art, photo, or halftone); setting the brightness level(density); adjusting the background setting (for paper that is yellowed or colored); and selecting filters, screens, and Tone Reproduction Curves (TRCs).

The image display window enabled technicians to preview each scanned page on the screen at 600 dpi resolution, although in production this was normally only done for a few pages for each book to determine standard production settings for the entire book. Highly-illustrated texts containing halftone images, however, required more manual intervention to adjust the settings.

21 Disbinding books with minimal artifactual value met little faculty resistance when high-quality replacement facsimiles were produced, and additional copies can be printed on demand.

The final step in setup involved the scanning of the production note that is reproduced in every book. The production note describes the scanning process, the paper used for printing, and serves as notice of Cornell's copyright of the digital files.

D. PRODUCTION SCANNING

Once setup was complete, the technicians moved into production and scanned rapidly, performing little on-screen inspection for the rest of the book. The quality control windows on the scanning workstation were closed to improve response time, and the technicians concentrated on scanning--raising the platen, positioning each page, lowering the platen, and pressing the scan button. Technicians occasionally confirmed image placement (right-handed or left-handed page) by checking the position of the page in the book icon on the monitor. Very little quality control was required during production scanning, especially for books that were largely textual and printed in a consistent manner. If technicians came upon unusual material (an illustration or a very faint page), they returned to the quality control mode to check the on-screen image and to make any necessary adjustments in the settings.

In production mode, technicians could scan at a rate of about 5 pages per minute. Total production times, however, had to allow for setup and other factors. Since the pages of the scanned books are brittle, automatic document feeders are not used; however, Cornell plans to experiment with automatic feeders on books that have already been scanned to assess the degree of brittleness that can be tolerated. The use of such feeders may be realistic for certain classes of books that are (a) minimally brittle, and (b) available from other libraries, so that replacement pages may be obtained in the event a page from the original book is destroyed by a document feeder.

E. PRINTING

To print a volume, the scanning technicians initiate a command to transmit the digital files over the Cornell TCP/IP network for printing. Transmission time averaged 6 to 10 seconds per page depending on the file size. The delay was not caused by the network, but the time taken to transfer the files from the local disk—a result of the particular disk technology used. About mid-way through the study, Xerox provided software that made the printing command a background task so that digital books could be queued for printing after working hours. This enhancement led to an increase in scanning productivity.

Since rescans for quality control reasons were few, printing was done directly on paper meeting ANSI permanence standards.

F. QUALITY CONTROL AND RESCANS

For quality control, the scanning technicians found it easier and more reliable to inspect the paper version, rather than to view the on-screen images. The paper copy was inspected page by page for completeness, order, legibility, and, by direct comparison, fidelity to the original. Any missing pages or those deemed unacceptable were rescanned from the original copy. The rescan rate was under one percent of all copies made. Until the volume was proofed and a final version accepted, the digital files were stored on the local hard disk of the scanning workstation.

G. BINDING, CATALOGING, AND SHELVING THE PAPER REPLACEMENT

All printing was done on Gilbert Neu Tech twenty-five percent rag, alkaline paper. The final paper version was bound with a one and a half inch binder's margin, using the double fan adhesive method of leaf attachment and a full cloth binding by a local book bindery. Catalogers referred to the bound volume and the project worksheet that contained information on the digital files in creating bibliographic records. The completed volume was then sent to the stacks to replace the original book, which in most cases was withdrawn from the library.

H. STORING AND ACCESSING THE DIGITAL FILES

Once a satisfactory facsimile had been produced to replace the original, the final versions of the digital files were transferred to optical disks. During this phase of the project when the jukebox was not available, technicians transferred the digital files to local, removable 5.25" optical disks using a disk drive attached to the scanning workstation. The scanning technicians manually reloaded these disks to retrieve the digital files for subsequent printing and viewing. As in after-hours printing, copying to optical disk became a background function made possible by a software upgrade.

Once it becomes available, the final versions will be transferred directly to the optical jukebox.

I. TECHNOLOGY REFRESHING

The networked document imaging system used as the foundation of this project relies on emerging technologies. The obsolescence of formats and software access tools associated with a rapidly changing technology is a concern not only to those prepared to use the technology but to the preservation community as well. Preservation using this medium will therefore entail the recopying on a regular schedule set well below the medium's expected longevity, a process that has become known as "refreshing." Software programs are also

becoming available that check the stability of a disk every time it is used.

Our cost study suggests that the costs of refreshing are likely to be offset by space savings as compression and storage capabilities improve. Cornell is committed to a process that will continually "refresh" our digital library so that each volume is copied every four years.[22] Technology refreshing will be done at the University by the Information Technology department rather than the library. To rely on a unit outside the library to maintain actual collections is something of a departure in library practice, although not completely without precedent in that libraries often rely on their information technologies' colleagues to maintain and refresh the on-line catalogs. Costs for this process are included in our cost study. Procedures to assure that this work is done are being developed. The responsibility for ensuring that the digital library will serve the interests of library users in the future depends on an institutional commitment in the present.

A program of maintenance also includes regular backup of the digital files, processes familiar to central university information technology organizations. Optical disks can be duplicated on additional optical disks or other electronic media, which can be stored in a separate, climate controlled location. But the backup costs can be high, especially if one must also refresh the additional copies on a regular schedule. Of course, the paper facsimile provides a form of backup. As previously noted, Cornell is also investigating the use for backup of microfilm produced from the digital files.

22 It is anticipated that as data exchange standards are developed and implemented, the time between refreshing will increase from four years to ten years and beyond. See for instance, Charles M. Dollar, "The Impact of Information Technologies on Archival Principles and Practices: Some Considerations," Draft Version 16, November 15, 1990, pg. 63.

V. FINDINGS

A. NEW PRESERVATION METHOD

Digital image technology provides an alternative—of comparable quality and lower cost—to photocopying for preserving deteriorating library materials. Subject to the resolution of certain problems, digital scanning technology also offers a cost effective adjunct or alternative to microfilm preservation.

Digital technology represents a new preservation technique that can be used in the place of or in combination with analog processes, such as photocopy or microfilm. This study has demonstrated that 600 dpi scanning is comparable or superior to the quality achieved in preservation photocopy, a standard preservation option in most libraries. The digitally produced paper product is also superior to the paper printout from microfilm produced on a standard reader/printer. An evaluation of quality is described below. The production of microfilm from the digital files and its comparison to the quality achieved via standard light lens film will be considered in the next phase of Cornell's continuing investigation into the use of digital technology.

The cost study, also presented below, indicates that the economics of production, maintenance, and duplication are competitive with photocopying and microfilming, and, over the course of the next decade, will become significantly cheaper. The print duplication capability and the advantages of network access associated with the use of digital technology should enhance the national and international preservation effort.

Digital technology's practical utility, however, is dependent on the successful resolution of a number of issues, including: the development and application of standards and protocols for creation, storage, preservation, and use; the development of standards for technology refreshing; the growth of service bureaus or regional centers that can provide preservation scanning services for libraries unable to establish in-house programs; and the recognition of digital technology as a legitimate preservation technique by federal, state, and private funding agencies.

1. Quality Evaluation

Six hundred dpi binary scanning represents a viable preservation alternative to light lens processes for creating paper replacements of deteriorating originals. In reaching this conclusion the study compared the 600 dpi paper output to photocopies and to paper

printouts produced from microfilm, as well as to paper copies produced via lower resolution scanning.[23]

There are several advantages to copying a page digitally. Because the resulting image is digitally encoded, it can be reproduced and transmitted with no loss of quality. With analog processes, there is a discernible difference between the second and subsequent generations of an image.[24] This means that preservation in the digital world will be based on maintaining "information" in an accessible form. In the analog world, preservation is based on maintaining the physical medium (e.g., paper, film).

Digital image scanning can also lead to improved image capture. For instance, Xerox has developed a windowing application that segments a page containing both text and illustrations in a manner that enables different settings to be used and optimizes the reproduction of both.[25]

A digital image can also be edited and density levels adjusted to remove underlining, foxing, and stains or to increase legibility,

---

23 This study investigated the quality achieved with binary scanning only. Depending on the object being scanned, grey scale or color scanning may be superior, and the advantages/disadvantages of the various approaches need to be examined. Scanning resolutions and file formats can represent a complex tradeoff between time, file size, fidelity, on-screen display, printing, and equipment availability. The study had as a primary emphasis the production of printed facsimiles that were largely black and white text in a timely and cost-effective manner. With binary scanning, large files may be compressed efficiently and in a lossless manner using CCITT Group IV Facsimile compression algorithms. Grey scale compression, using JPEG, is much less economical and is "lossy," which may make it inappropriate as a preservation method. It appears that while binary files produce a high quality printed version, other combinations of spatial resolution with grey and/or color will also be adequate. Grey scale can offer an advantage for on-screen viewing. For instance, on a low resolution screen display, two bits of grey at 100 dpi may be more readable than 600 dpi or 300 dpi binary. The advantage is lost, however, when the on-screen image is enlarged. The quality associated with binary or grey scale is also dependent on the equipment used, for instance binary scanning produces a better paper copy when it is printed on a binary printer. See Michael Ester, "Image Quality and User Perception," LEONARDO Digital Image, Digital Cinema Supplemental Issue, (1990) pg. 51-63.

24 Generational loss is acknowledged in the draft photocopying guidelines of the Subcommittee on Preservation Photocopying Guidelines, of the Reproduction of Library Materials Section of ALA. The August 1991 version emphasizes that acceptable copy image quality should consider reproducibility (i.e., can the text be copied again). The generational loss with microfilm is not as great, but does represent about a 10% reduction in resolution with each generation. As such the technical specifications for microfilm vary from one generation to the next. See, for example Research Libraries Group, Inc., RLG Preservation Microfilming Handbook, edited by Nancy E. Elkington, (Mountain View, CA: The Research Libraries Group, Inc., 1992), Appendix 18. See also, Don Willis, A Hybrid Systems Approach to Preserving Printed Materials using Microfilm and Digital Imaging, presentation at the AIIM conference, April 1991.

25 A process of auto-segmentation, which incorporates the windowing function automatically as a page is scanned, is being refined by Xerox. When available, it will increase the speed of capture for illustrated text.

options which are especially valuable when paper containing high levels of lignin has darkened considerably. A page may be cropped so that black borders common in photocopying are eliminated. Obviously, in producing a replacement copy, decisions must be made as to how much enhancement is desirable or affordable, but the capabilities exist for producing a copy that meets or exceeds the quality of one produced with photocopy technology.

The Joint Study compared the image quality obtained using light lens processes to that of the prototype scanning system. Copies of a standard facsimile test chart were produced on a Canon 8580 photocopier, a Minolta RP605Z reader/printer, and the Xerox scanner, which was used to produce 200, 300, and 600 dpi versions. The IEEE Facsimile Test Chart provides a means for evaluating the capture of text, gray scale, line art, photography, and resolution. The results of this comparison are included in Appendix I. The sample pages located there offer one illustration of the advantage of 600 dpi scanning and printing over lower resolution image capture and also demonstrate the current system's superiority to light lens print processes in capturing illustrated text.[26]

2. Cost Study

Digital technology represents an affordable alternative to light lens processes for reformatting brittle books. In evaluating the various reformatting options, it is important to consider not only the costs of the initial copying, but also the costs (and value) over time of providing subsequent copies, access, storage, and maintenance.

Since each preservation process has different objectives or advantages, comparison of equivalent costs is difficult. To the extent possible, our cost comparison attempts to relate similar objectives, but this is not always feasible. Our most general finding is that when similar objectives are compared, the costs of using digital methodologies are competitive with the costs of using traditional preservation reformatting techniques. Moreover, there is a greater likelihood that costs of digital processes will decrease over time than is the case with other reformatting options. Thus, cost alone should not be the determining factor in the choice of a preservation format.

26 An excellent discussion of relating photographic quality indexes with digital scanning is presented in AIIM Technical Report (TR 26), "A Tutorial on Photographic and Electronic Imaging Resolution," draft, 2/5/92. See also Tom Bagg, "Image Quality," paper presented to the Digital Image Applications Group, Sept. 25, 1986; and Don Willis, "A Hybrid Systems Approach to Preserving Printed Materials using Microfilm and Digital Imaging," draft paper, 1991, unnumbered.

TABLE A. COMPONENT COSTS OF USING SCANNING TECHNOLOGY TO CREATE PAPER REPRODUCTION OF PREDOMINANTLY TEXT OR LIGHTLY ILLUSTRATED 300 PAGE BOOK AND TO MAINTAIN THE DIGITAL MASTER
(see Appendix III for Assumptions)

Unit Costs in 1992 Dollars (unit = 1 book)
5% rate of change represents normal inflation

	Rate of Change	Year 1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	10 Year Average Cost
Scanning Costs												
1. Labor	5%	$20.64	$21.67	$22.76	$23.89	$25.09	$26.34	$27.66	$29.04	$30.49	$32.02	$25.96
2. Equipment	-10%	$4.92	$4.43	$3.99	$3.59	$3.23	$2.91	$2.61	$2.35	$2.12	$1.91	$3.20
3. Total Scanning Cost		$25.56	$26.10	$26.74	$27.48	$28.32	$29.25	$30.27	$31.40	$32.61	$33.93	$29.17
Optical Storage Costs												
4. Optical Jukebox	-15%	$1.60	$1.36	$1.16	$0.98	$0.84	$0.71	$0.60	$0.51	$0.44	$0.37	$0.86
5. Optical Disk	-30%	$5.36	$3.75	$2.63	$1.84	$1.29	$0.90	$0.63	$0.44	$0.31	$0.22	$1.74
6. Technology Refreshing*						$2.12	$1.61	$1.23	$0.95	$0.75	$0.59	
Printing Costs												
7. Printing Equipment	-5%	$6.75	$6.41	$6.09	$5.78	$5.50	$5.22	$4.96	$4.71	$4.48	$4.25	$5.41
8. Acid Free Paper	5%	$1.50	$1.58	$1.65	$1.74	$1.82	$1.91	$2.01	$2.11	$2.22	$2.33	$1.89
9. Total Printing Cost		$8.25	$7.98	$7.74	$7.52	$7.32	$7.13	$6.97	$6.82	$6.69	$6.58	$7.30
Binding Costs												
10. Library Binding	5%	$7.00	$7.35	$7.72	$8.10	$8.51	$8.93	$9.38	$9.85	$10.34	$10.86	$8.80
11. In-line Finish	5%	$1.00	$1.05	$1.10	$1.16	$1.22	$1.28	$1.34	$1.41	$1.48	$1.55	$1.26
12. Unbound/Stapled	0%	$0.25	$0.25	$0.25	$0.25	$0.25	$0.25	$0.25	$0.25	$0.25	$0.25	$0.25
13. Weighted Binding Cost **		$1.90	$1.99	$2.08	$2.18	$2.29	$2.40	$2.51	$2.63	$2.76	$2.89	$2.36
Access Costs												
14. Access Cost /Book ***	-10%	$1.60	$1.44	$1.30	$1.17	$1.05	$0.94	$0.85	$0.77	$0.69	$0.62	$1.04

* On average a book will be refreshed twice in a decade. For instance, a book created in 1992 will be refreshed in 1996 and 2000.
** This weighted average binding cost assumes 20% library binding, 40% In-line, and 40% unbound/stapled.
*** This is highly dependent on the choice of technology.

Table A enumerates the component costs associated with the use of digital technology to create paper facsimiles, to maintain a master in the digital library, and to produce subsequent printed copies. These figures are based on a time and cost study conducted in the last three months of 1991. They are also based on a number of assumptions and projections. For example, rates of change of component costs (increases or declines) are assumed and projected into the future, as indicated in Table A. A description of the cost study is contained in Appendix II. Details of the measurements and assumptions underlying Table A are contained in Appendix III.

It must be noted that the scanner used in this study is not yet available in the market. A cost has been imputed to this scanner based on comparable costs and capabilities of other scanners.[27] Within reasonable limits, the findings of this study are not sensitive to this figure, since scanning costs are dominated by labor rather than equipment costs.

These component costs can be combined in a number of ways to facilitate various comparisons with other methods. We illustrate, as examples, comparisons with photocopying for the production of one paper facsimile; with microfilming as a potential means of long-term storage, including the costs of "technology refreshment"; and with photocopy for the production of subsequent printed copies of the book. These are summarized in the form of "findings."

Finding 1: Digital technology offers an economic alternative to photocopy to produce a paper replacement.

Traditionally, photocopy is chosen as a preservation technique when a paper copy of a book must be produced and returned to the shelf. Table B compares the cost of producing a paper facsimile via photocopy and digital imaging. To ensure valid comparisons, in each case it is assumed that no copy other than the paper replacement is retained.

The costs indicate that producing one copy of a book using digital technology has economic advantages, even at this early stage in the development of the technology. The investment of labor to handle the deteriorating book is the largest component of cost in each case. Labor will increase with inflation, and therefore the costs of each option will increase over ten years. With digital technology, however, costs will rise more modestly as the costs of technology

27 Nonetheless, Xerox has concurred with the figure used in the cost study.

decline to partially offset increases in labor and finishing. Because of its lower costs and other advantages, digital scanning and printing could replace photocopying once widespread production capabilities can be established.

Table B	1992 Cost	10 Year Average
Digital Technology		
Scanning: Labor and Equipment [3]	$25.56	$29.17
Printing [9]	$8.25	$7.30
Overhead - 30%	$10.14	$10.94
Library Binding [10]	$7.00	$8.80
Total Printed Copy from Digital	**$50.95**	**$56.21**
Photocopy		
Total Photocopy	**$65.00**	**$74.52**

TABLE B. PHYSICAL REPLACEMENT OF BOOKS (NO STORAGE OF FILES)[28]

Finding 2: The production and long-term storage costs for digital technologies are competitive with those of microfilm. Subject to the resolution of certain problems, digital scanning technology will offer a cost effective adjunct or alternative to microfilm preservation.

If a paper replacement is not required, microfilm is currently the preferred format due to a high degree of permanence when properly processed and stored. The high bandwidth of film also enables the capture of the finest details of the text, although production microfilming may not always capture halftone images effectively. With today's technology, production digital scanning, while adequate for most practical user purposes, is of lower capture resolution than microfilm.

These differences must be taken into account. However, digital technology is competitive today with microfilming for creation, storage, and maintenance of a duplication master when only *costs* are compared, including the costs of technology refreshing (see Section IV-I, Technology Refreshing). Table C compares the relative costs of a duplication master over ten years for both microfilm and

28 Costs associated with digital technology are derived from Table A. The numbers in [brackets] refer to line numbers in Table A. Overhead reflects the general and administrative costs and profit margin that would be included by an outside vendor. The 1992 cost of photocopying is based on two quotes for photocopying and binding a 300 page book (Library Bindery Service and Ridley's Book Bindery). The average annual inflation rate is calculated at 5%.

digital technology. The duplication master in the context of digital technology is the file maintained by the creating institution. The cost for both one-up (one page per frame) and two-up (two pages per frame) microfilming are included. Although most current microfilming projects use the latter, it may well prove that one-up microfilming will provide better results if and when the microfilm is converted to digital files. The Yale study, mentioned earlier, should provide information on the best method for creating microfilm that will subsequently be scanned. In Table C, we only show the comparative costs of capture in 1992 and maintenance for ten years. However, the digital costs for capture in subsequent years will rise more modestly than the microfilm costs because of the declining costs of digital technology. Thus, although the costs are more expensive in 1992 than two-up microfilming, digital technology will have a steadily increasing cost advantage. Digital technology already shows significant cost savings over one-up microfilming.

Table C	1992 Capture and 10 Year Maintenance	
Digital Technology		
Scanning: Labor and Equipment [3]	$25.56	
Storing: Optical Jukebox & Media [4&5]	$6.96	
Refreshing - 10 years [6]	$2.87	
Overhead - 30%	$10.62	
Total Digital	**$46.01**	
Microfilm	**One-up**	**Two-up**
Creating Archival Master	$58.50	$29.25
Creating Print Master	$5.00	$2.50
Storing 2 Generations for 10 Years	$6.66	$3.33
Total Microfilm	**$70.16**	**$35.08**

TABLE C. CREATION, STORAGE, AND MAINTENANCE OF DUPLICATING MASTER FOR 10 YEARS[29]

29 The numbers in [brackets] for digital technology refer to line numbers in Table A. A book scanned in 1992 will be refreshed twice in the next decade, in 1992 and 2000. Overhead reflects the general and administrative costs and profit margin that would be included by an outside vendor. Microfilm figures are based on 1992 prices quoted by MicrogrAphics Preservation Service (MAPS). Cost of archival master is based on $.195/frame for one-up and two-up filming. Cost of print master is $15. For two-up filming, assume six books can be stored on each roll; for one-up filming, assume three books. The cost of one book on the print master will be $5.00 (one-up) or $2.50 (two-up). Storage costs are based on $1/year to store one roll of film. The cost of book storage/year will equal $1 divided by 3 (one-up) or by 6 (two-up). Since two generations are being stored, the cost equals $.66 (one-up) and $.33 (two-up) per year times 10 years, or $6.66 and $3.33 respectively.

Finding 3: Digital technology represents an economic means for producing subsequent printed copies.

Subsequent copies of a book can be printed on demand from the stored digital files and at a fraction of the cost of the first copy, because labor, the dominant cost, is required only once in the initial capture. Table D presents the costs of printing a book in this way compared with photocopy. Photocopying of course suffers from the disadvantage that the original or, if the loss of quality can be tolerated, a photocopy, has to be recopied each time a another copy is required. Subsequent copies will be at least as expensive as the first. This Table demonstrates that the cost to create subsequent copies using digital technology is extremely competitive. Of course, the storage costs are already assumed to been combined with the capture costs (see Table C). A paper copy can also be produced directly from microfilm, however, the cost would be more and the quality would suffer. Due to the complexity of comparing user preferences between microfilm readers, desktop workstations, and paper copies, we have not compared the costs and benefits associated with these access technologies.

Table D	1992 Cost	10 Year Average
Digital Technology		
Access [14]	$1.60	$1.04
Printing [9]	$8.25	$7.30
Overhead - 30%	$2.96	$2.50
Binding [13]	$1.90	$2.36
Total Digital	**$14.71**	**$13.20**
Photocopy Technology		
Subsequent Copy	**$65.00**	**$74.52**

TABLE D. COST TO CREATE A SUBSEQUENT PRINTED COPY OF A BOOK[30]

30 The numbers in [brackets] for digital technology refer to line numbers in Table A. Overhead reflects the general and administrative costs and profit margin that would be included by an outside vendor. The binding cost included here assumes that 20% of all requests for subsequent copies will be bound with a full cloth library binding , 40% will be bound using Docutech in-line tape binding, and 40% will be unbound stapled. If we assumed that all subsequent copies were bound in a full cloth binding, the total digital cost would rise to $19.81 in 1992 dollars.

Summary of Findings

These findings indicate that when the need is to replace paper with paper, the use of digital technology is economically preferable. In addition, it is considerably less expensive to produce subsequent copies from a digital file. In the future, digital technology should replace photocopy as a preservation format, once production facilities are established.

These studies also indicate that the costs of digital technologies are competitive with microfilm, including the cost of technology refreshing. For digital technology to compete with microfilm technology, however, refreshing must become institutionalized. Furthermore, the resolution of the stored image using today's technologies is not as high as that of microfilm, although it offers advantages for capturing illustrated material and may be adequate for many purposes.

Paper copies can be produced more cheaply with digital technologies, and the quality is superior to that produced by most microfilm reader-printers. Of course, the primary means of access to microfilm is the microfilm reader normally located in the library. Further work needs to be done to compare the added value of providing desktop access at the researcher's workstation to stored digital books.

B. NEW ACCESS METHOD

The network-connected digital library offers a new access method. In the future this technology will permit viewing books on workstations, browsing collections from several institutions at the same time, and producing print-on-demand facsimiles for use. New forms of indexing will be needed to navigate this information resource.

1. Network-Connected Digital Library

Data communication networks provide instant connection to information resources. The growth of the Internet in the past five years is an indication that an increasing number of individuals are using networks for communication and information delivery. In 1991, the United States Congress passed the High Performance Computing Act, intended to support the creation of the National Research and Education Network (NREN), a national network system that will support the bandwidth needed for the rapid transmission of digital image files.

The digital library can be viewed by researchers from home or office workstations connected to the network. Browsing virtual library "shelves" from the workstation introduces a new kind of access to library collections. Digital technology now provides the ability to generate print-on-demand facsimiles of library books. The use copy from the digital library can be printed in response to a request submitted at the workstation. The network infrastructure connects all the components of the system: the request for a printed copy travels to the library image storage system, then to the printer, producing a cost-effective paper copy for use. Network protocols regulate the transmission of requests, and the transfer of data files to satisfy the request.

2. Navigating the Digital Library

With new and more complex sources of information on-line, improved indexing is needed so that the researcher using a workstation can find the resources that are available. New indexes in no way obviate the need for traditional catalogs. In fact, the catalog needs to be updated with records to locate library material stored in digital format, and new links between the catalog and this material need to be developed. The most important reason to add information to the catalog is so that these sources become a part of the total collection of the library. The library catalog brings sources together. But we also need to change the concept of the catalog (i.e., records of items in a particular collection) as the concept of the "collection" changes.

New indexes will link the on-line catalog to the digital library, and one digital library to others. Further indexing is required to represent such detail in the digital library. Traditional bibliographic records bring the user to the whole book. In the digital world, the user will want to be able to access parts of a document, such as a chapter or index of a book.

A very preliminary experiment took place in June 1990 using the table of contents and the index for one of the 19th century volumes in this project. The pages were scanned and run through a Kurtzweil Optical Character Recognition program with an error rate of 3%. It was concluded that this error rate was too great for representing those parts of the book to the reader, particularly given the nature of the material in this project, which included a heavy preponderance of mathematics texts and volumes in non-Roman languages. A second pilot project involves the keying in of the table of contents for some of the math books using TeX software. This is a very time consuming process, and only a few titles had been completed by project's end.

C. APPLICATIONS BEYOND PRESERVATION: ELECTRONIC PUBLISHING AT CORNELL

Library preservation operations have only limited resources to devote to system development. If a digital solution to the preservation crisis is to be achieved, that solution needs to have commercial application beyond preservation alone. The library can then leverage other applications that have commercial viability, and benefit from the development that they fund. In the case of the CLASS system, electronic publishing applications are emerging that will use the same technological infrastructure as CLASS, and preservation can reap the benefits of developments that were created for other purposes.

Cornell University is among a number of universities engaged in projects that will define the boundaries of electronic publishing for the future. Cornell expects to participate in a multi-institutional project with Elsevier, the largest commercial publisher of technical journals, to experiment with the electronic distribution of material science journals. At Cornell, these journals will become a part of the digital library, and the view stations that support browsing of digitally preserved books will also support the browsing of Elsevier journals.

The Synthesis Coalition (a coalition of engineering colleges from several universities, headquartered at Cornell) is developing network based tools for teaching engineering. As a part of this work, the Coalition is engaged in a joint study with John Wiley and Sons, Publishers to experiment with the electronic viewing and use of engineering textbooks. Approximately forty Wiley texts have been included in the experiment. These books will also become part of the digital library. Engineering students will use the view station software from the CLASS project to read and select from these volumes, which will be integrated with the navigator tools provided by the Synthesis Coalition.

Customized publishing that combines material from various sources to meet the needs of a particular course is already an important Cornell program. The Cornell Campus Store, under the direction of Rich McDaniel, is pioneering the use of electronic publishing to produce customized coursepacks composed of selected published material combined with faculty prepared selections, and to print them on demand. The publishing system used for this application is an extension of the CLASS system. Customized publishing depends on efficient procedures and systems to manage the clearance of copyright. In response to the new pressure implicit in the electronic arena, new copyright clearance services are being offered by such organizations as the National Association of College Stores (NACS) to meet the needs of academic organizations.

An application of electronic publishing currently being explored at Cornell involves the potential collaboration between the Cornell University Press and the Campus Store. Cornell Classics on Demand is a proposed experiment where out of print books from the Press could be scanned and offered for sale on a print-on-demand basis through the Campus Store. These books may never again be out of print. Short print runs, even one or two copies, can be done to meet customer demand. The system that could run Cornell Classics on Demand is essentially the same CLASS system used for library preservation.

These are by no means all of the projects in electronic publishing that are now being conducted at Cornell. Library preservation has some requirements that are special and will not be met by any of these applications, but it also has much in common with them. The common aspects of each project should result in significant progress that can be of benefit to preservation.

VI. CONCLUSION

The Joint Study concludes that digital image technology represents a new method for the preservation reformatting of library materials that in the future will replace or complement microfilming and photocopying. The use of digital technology is currently cost-effective as a reformatting option, and the quality is comparable to light lens processes. The technology offers a means for replacing paper with paper, while simultaneously providing new access opportunities. In the future, researchers will be able to access not only catalog records but also the full text to which those records refer.

Given these strong conclusions, why in the future and not today? Digital technology is a new preservation technique and, as such, standards for its application are not in place. Cornell has identified some areas that warrant further investigation, and that need to be resolved cooperatively among several institutions. The document structure definition is an example. In order for digital technology to realize its potential for scholarly use, the document structure must be the same for each library. Like bibliographic records, document structures facilitate access only when consistent and easily understood by the library patron. Cornell concludes that cooperation among institutions is essential during this period of transition.

A second example involves the role of service bureaus. Ultimately, preservation scanning is likely to be contracted out to service bureaus. Over the course of the past five to seven years, preservation microfilming service bureaus have developed, and most institutions have shifted from in-house microfilming to using such services. While quality has been an issue in this transition, the major reason for the shift centers on economics and scale. It is anticipated that similar economies will drive the digital scanning process. Moreover a service bureau would be in a better position to absorb the costs and risks associated with using a developing technology.

Requirements for institutionalizing the storing, backup, and refreshing of digital files will be issues in the use of this technology as a preservation medium. The refreshing requirements for digital technology, including frequency of upgrades, costs, and administration, are not clearly defined and represent a significantly larger commitment on the part of an institution than does providing a proper storage environment. Service bureaus should be involved in assisting in the development of standards and procedures for the creation, storage, and use of digital masters. However, research libraries must be active partners in the development of such requirements to ensure that the needs for preservation of their collections will be met.

Today is a period of transition from the established analog technologies to the newer, more flexible, and rapidly changing digital technologies. Cornell believes that digital technology offers benefits to librarians and scholars that justify continued study into its use for preservation and access.

VII. APPENDIXES

Appendix I—Paper Facsimile Comparisons

Appendix II—Cost Study Description

Appendix III—Cost Study Assumptions for Table A

Appendix IV—Scanning Diagram

APPENDIX I—PAPER FACSIMILE COMPARISONS

The sample pages located in Supplement V graphically represent the quality obtained through light lens and digital processes. The samples are reproductions of a standard Facsimile Test Chart that provides a means for evaluating the capture of text, gray scale, line art, photography, and resolution.[1] The Test Chart includes two samples of the Microcopy Resolution Test pattern for evaluating microrecording systems. They are used to measure resolution, which is defined in terms of the number of line pairs per millimeter that can be seen or differentiated on these test patterns. The first measures resolution on both the vertical and horizontal axis; the second measures resolution at right angles. While primarily used to determine the quality of microfilm, the test charts provide a convenient means to compare the paper output of light lens processes to digital scanning.

The results and observations of the text pages were as follows:

1. Canon 8580 photocopy (regular setting)
   - — text: 2 pt. text legible, some characters partially represented
   - — resolution patterns read: 6.3 (first pattern); 5.0 (second)
   - — photograph: poor definition
   - — gray scale: 5 discernible levels
   - — other: second set of fine lines at top not reproduced

2. Canon 8580 photocopy (photo setting)
   - — text: lower case 2 pt. text partially represented
   - — resolution patterns read: 6.3 (first pattern); 4.0 (second)
   - — photograph: marginal quality
   - — gray scale: 6-8 levels
   - — other: second set of fine lines at top not reproduced

3. Minolta RP605Z reader/printer (printout from microfilm)
   - — text: 2 pt. text not rendered, 4 pt. and 6 pt. partially represented
   - — resolution patterns read: 2.5 (first pattern); 2.2 (second)
   - — photograph: unsatisfactory
   - — gray scale: 5 levels
   - — other: second set of fine lines at top partially rendered

1 Chart IEEE Std 167A-1987. Prepared by the IEEE Facsimile Subcommittee and printed by Eastman Kodak Company. For use in accordance with IEEE Std 167-1966, Test Procedure for Facsimile. Copyright 1987, Institute of Electrical and Electronics Engineers.

4. 600 dpi Digital File
   - — text: upper case 2 pt. text barely legible, lower case 2 pt. text partially rep.
   - — resolution patterns read: 5.6 (first pattern); 5.0 (second)
   - — photograph: good definition
   - — gray scale: 9-11 levels
   - — other: second set of fine lines at top fully rendered

5. 300 dpi Digital File
   - — text: 2 pt. text illegible, 4 pt. text breaking up
   - — resolution patterns read: 4.0 (first pattern); 3.2 (second)
   - — photograph: dotted, like pointillism
   - — gray scale: 10 levels, but dotted
   - — other: fine line sets at top partially rendered

6. 200 dpi Digital File
   - — text: 2 pt text illegible, 4 pt. text breaking up; block representation to 12 pt. text
   - — resolution patterns read: 2.8 (first pattern); 2.8 (second)
   - — photograph: very grainy, eye-legible dots
   - — gray scale: 9 levels, eye-legible dots
   - — other: fine line sets at top poorly rendered

The first two sample pages were produced on a Canon 8580 which is used at Cornell to make preservation photocopies of brittle books. The first photocopy was produced on the regular setting and the second was produced using the "photo" setting. This latter shows an improvement in the capture of the gray scale and photograph, but there has been a marked decline in the capture of text and fine lines.

The third sample is a printout from a microfilmed version of the scanner test target. The copy was produced on a Minolta RP605Z reader/printer that uses a dry printing process. The film from which this printout was produced has a high resolution, and the grayscale and photograph are acceptably rendered. The paper printout, however, represents the poorest quality of the six samples. Text, photograph, and gray scale have all clearly suffered. The results indicate that, although the film copy may meet national standards, the quality of the paper use copy does not. More attention should be paid to the quality requirements of the use copy.

The fourth sample is a 600 dpi representation. The gray scale at the top of the test chart and the photograph were windowed during scanning and captured using the photo mode with a selected filter, screen, and Tonal Reproduction Curve (TRC) that rendered them exceptionally well. Eleven shades of gray are discernible and the photograph possesses a depth not present in the photocopied versions. The text and fine lines were captured in the mixed mode, again using a variety

of settings, and fall somewhere between the two photocopy versions in terms of text capture. The overall presentation of the 600 dpi scanned version is superior to either of the photocopies and the microfilm printout, and dramatically illustrates digital technology's advantage over light lens processes for recording illustrated material.

The final two examples are versions scanned at 300 and 200 dpi resolution. Each was scanned using the identical windows and settings that were applied to the 600 dpi copy. The text, gray scale, and the photograph have declined visibly in overall presentation. The quality achieved with 300 dpi resolution is comparable to that of a laser printer; the 200 dpi version resembles the quality achieved with a dot matrix printer. The gray scale and the photograph are rendered as eye legible dots in both copies. While lower resolution scanning can produce satisfactory copies from crisp, high contrast modern documents of 6 point type and larger, these examples graphically illustrate the limited utility of lower resolutions to capture the variety of printing techniques and illustrations found in older research materials.

APPENDIX II—COST STUDY DESCRIPTION

From October - December 1991, Cornell conducted a time and cost study to determine the costs associated with using digital technology to reformat brittle books. In addition to tracking technician time, the study calculated the costs of equipment (amortized over four years), of storing and refreshing the digital files (every four years), the cost of printing and binding the paper facsimile, and a 30% additional amount for overhead.

The Joint Study focused on developing, testing, and evaluating a prototype scanning system for preservation. Because much time was devoted to product development, a stable production environment was not achieved for most of the study. Nonetheless an average production rate of 5,000 images/week was sustained over the course of the last year. This figure represents the total number of images scanned in a week by two technicians who also performed all of the non-scanning activities (collation, disbinding, inspection, etc.) normally associated with a preservation reformatting project. They also reflect staff time out for sick leave, vacation, training, and trips to Rochester for Project Development Team meetings, and time spent in demonstrations that resulted from the high visibility of this study.[1]

Due to the difficulties of obtaining reliable measurements in a production environment, staff recorded actual production figures in the last three months of 1991. Scanning technicians logged on a worksheet for each volume the time they spent on set up, production scanning, and rescanning. They did not record time spent in other tasks associated with selecting, preparing, and inspecting material. These tasks are common to other reformatting methods such as microfilm and photocopy and were considered the same in this project. The worksheets were then used to calculate average times for each task and the number of pages per hour scanned.

Although the size of the books varied from 100 pages to well over 700, for comparison purposes, the time spent in the various functions was

1 The research and development flavor of the study was reflected in fluctuations in scanning productivity. Between April 5 and May 24, 1991--an eight week period--the average weekly scan rate was 6,795 pages, which represents 22.65 books/week. This highly productive period was followed by a week in which only 7.5 books were scanned. System upgrades occurred at regular intervals throughout the year and a reduction in scanning production invariably accompanied software installation. Installation itself usually took a day for testing and debugging. Technicians had to prepare for the installation by clearing the hard disk of work in progress. They then had to learn the new system. Difficulties associated with installing new software on a networked system also were common. For instance, during the week that the P1 software was installed, 3,883 images were scanned; the week the P2.0 software was installed only 3,245 images were scanned; and the week the P2.1 software was installed only 2,778 images were scanned.

adjusted to represent a 300 page book. The total time spent averaged 1.72 hours/book (103 minutes), which represented a scanning rate of 175 pages per hour. Actual scanning rates varied considerably from this figure, with a low of 92 pages/hour to a high of 264 pages/hour, depending on the size of the book, the frequency and type of illustrations, the consistency of the printing, and a number of other factors. Worksheets for volumes that involved major system difficulties (e.g. system crashes, network rollovers) were excluded, although sheets for volumes scanned immediately after software upgrades were included.

A similar time and cost study for a preservation photocopy project of Cornell's Entomology Library materials was conducted during this same period. The average time spent in photocopying a 300 page book was 2.25 hours (135 minutes), an increase of 31% in time over the scanning process.[2]

The time spent in scanning was divided among the following tasks:

SET UP. Average Time: .4 hours/book (24 minutes)

The average time for set up varied from one technician to another. The comparable statistic for preservation photocopy is the time it takes to set up the template, which averaged 19 minutes in the Cornell Entomology project. Considerable time in set up is required to establish page size, determine front to back registration, and to scan the production note. It is estimated that set up time would be halved if these functions did not have to be performed manually. Xerox has a "wish list" of technical improvements for the system that will decrease the time necessary for set up.

PRODUCTION SCANNING. Average Time: 1.13 hours/book (67.8 minutes)

The speed of straight production scanning averaged 270-300 pages/hour, although this varied widely with the text density, the size of the book, the quality of the printed material, and the number of illustrations. Scanning time for pages that were densely packed with text or which contained illustrations increased as the file size increased. The length of a book also affected scanning time: for a 700 page book the time to scan one leaf (front and back) increased from 20.23 seconds at page 200 to 23.91 at page 400, to 26.95 at page 700. The delay was caused principally by an increase in the time it took to save a leaf and build the document structure for the book. The occasional need for quality control scanning of faint text or

2 Statistics prepared by Dorothy Wright, Preservation Librarian, Mann Library, Cornell University, December 1991.

illustrations slowed production scanning down significantly. Finally, fatigue from scanning more than two hours at one setting led to a noticeable reduction in production. Technicians were encouraged to alternate scanning with other non-scanning functions, such as quality control or the preparation of material.

RESCANS. Average Time: 0.1 hours/book (6 minutes)

The final step in scanning involved rescanning of images that were either missing or found to be of unacceptable quality during the inspection of the paper copy. Fortunately, the number of rescans was low, averaging less than 1% of all pages scanned. The rate of rescanning also dropped off as technicians became familiar with the system's image capture capabilities.

APPENDIX III—COST STUDY ASSUMPTIONS FOR TABLE A

Some of the cost data used in this analysis are based on confirmed measurements and quotations. Others are assumptions or projections. These measurements and quotations, along with the rationales for the assumptions and projections, are presented in this Appendix.

All assumptions are based on costs associated with preserving a 300 page book. The unit costs are calculated in 1992 dollars, and the average annual rate of inflation is figured at 5%. For all equipment, the annual costs are calculated at 47.2% of the capital cost. This figure is calculated to reflect the initial cost of the equipment, a four-year amortization, and a 10% interest rate; 50% is added for space, utilities, and maintenance. This figure represents a very conservative estimate of costs: the interest rates have fallen and the 50% load may not apply to all equipment, but for consistency it is used throughout. The rates of change in the cost of computer equipment and storage are based on historic price declines of the last decade and industry projections. For instance, optical disk storage costs are assumed to decline by 30% per year, scanning equipment by 10%, and printing equipment, which includes a number of mechanical features, to decline by only 5% per year. Obviously, any one of these rates of change could be debated: they should be considered indicative of trends currently underway rather than definitive figures.

Costs were projected for a ten year period. As Chart A indicates, the cost of digital technology will rise slightly over the decade, since the declining costs of technology are dominated by increases in labor and finishing. The labor figure for scanning assumes that there will be no increase in production, which almost certainly will occur as institutions move from prototype to production operations, as service bureaus begin to offer this service, as improvements in automatic and semi-automatic feed mechanisms reduce the risk of a paper jam, as automatic skew correction and page definition become standard, and as bound volume scanners are developed to enable the scanning of two pages at a time.[1]

[1] For instance, subsequent iterations of system software will increase the speed of scanning. Xerox has developed a fast scan capability which delays the document structure building until after the actual scanning has been completed. This upgrade has been tested on a scanning workstation located in Cornell's book store and its use at 300 dpi scanning led to a doubling of the production rate. Cornell did experiment with using a feed mechanism. It was determined that pages that were only marginally brittle (i.e., it took five double corner folds before the paper broke) could survive most paper jams. Libraries may be willing to risk a paper jam to achieve faster production rates for material held by a number of libraries. Before feed mechanisms can be used with this system, however, registration and deskewing must become software functions.

The numbers below refer to the line numbers in Table A.

Scanning Costs:

1. Labor. It takes a scanning technician 1.72 hours to scan an average 300 page book. The 1992 hourly rate, including benefits, is $12.00/hour. ($12 X 1.72 = $20.64) The ten year outlook does not include an increased production rate, which almost certainly will occur. The costs of labor are projected to increase by 5%/year.

2. Equipment. Equipment includes scanner, personal computer, high resolution monitor, application software and network connections. Annual scanning equipment cost is $9,440 ($20,000 X .472). The equipment is assumed to be operational for 2 shifts. Assuming a standard shift equals 37.5 hour work week, the total number of hours/year that the equipment is in use is 3,300. The hourly cost for equipment ($2.86) is computed by dividing the total number of hours per year by the annual scanning equipment cost ($9,440/3300). Since the average scanning time per book is 1.72 hours, the scanning equipment cost per book is therefore $4.92. This figure declines by 10%/year.

Transfer to Optical Storage:

4. Optical Jukebox refers to the cost of the time it takes to write a disk on the jukebox (20 minutes for 300 page book). This figure is calculated by taking the annual cost of jukebox ($75,000 X .472) and dividing it by 21,900 (which equals 20 minutes of a year: 365 days/20 hours per day,[2] divided by 3 for the 20 minutes). This figure declines by 15%/year, which is very conservative, given declines in technology cost and time taken to write the disk.

5. Optical Disk Cost. This figure represents the cost of the 12" disk ($500) amortized over 4 years, and divided by the number of books that will fit on a disk.
$500 X .472 / 44 = $5.36. This figure declines by 30%/year.

6. Technological Refreshing. The cost of refreshing, which begins in year 4 and is repeated every four years, is calculated by adding the cost of equipment (#4 above), which is the cost of the time taken to transfer to a new disk, and the cost of that portion of a new disk that the volume occupies (#5 above). Certainly the time between refreshing will increase as data exchange standards are developed.

Printing and Binding:

7. Print Equipment. This figure includes the cost of the Docutech printer, maintenance, set-up for printing, and associated costs. Cornell has set a price of printing at $.0225/side (excluding paper), which is based on full costing of the Docutech, including allowances for space and overhead. A 300 page book would cost $6.75. This figure will decline by 10%/year.

8. Acid Free Paper. The cost of acid free paper is calculated at $.01/sheet (a 300 page book would be printed on 150 sheets). This figure will increase by 5%/year.

10. Library Binding. The binding figure includes the cost of full cloth binding (Ridley's Book Binder) and preparation for binding. It will increase by 5%/year.

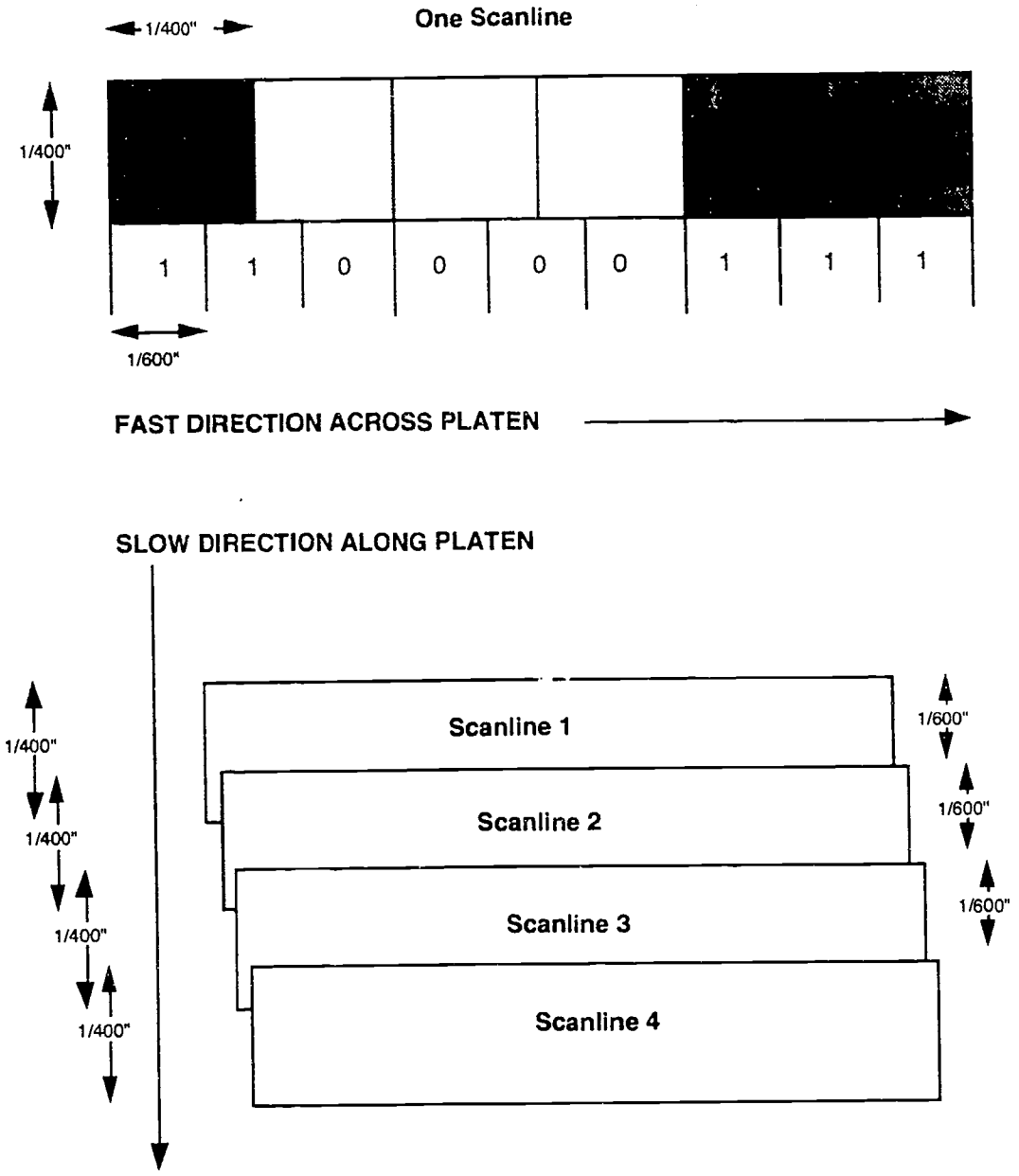
11. In-line Finish. The Docutech offers a number of finishing options, including a heat-set tape binding, the price of which is listed here. This figure will increase by 5%/year.

12. Unbound / Stapled. A nominal figure is included here for stapling.

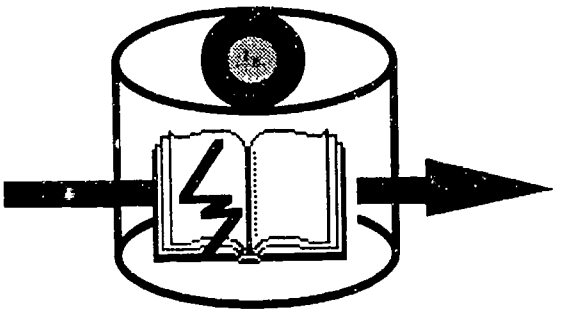
13. Weighted Binding Cost. Binding costs for subsequent copies are based on the assumption that not all requests will result in a library binding. It is assumed that a full cost library binding will be done for 20% of the print requests, an inexpensiv in-line finishing will be used for 40%, and that the remaining 40% will be stapled or left unbound.

14. Access Cost / Book. The cost of accessing the book assumes that the optical disk containing the book will be mounted on the jukebox when needed. This cost is calculated as the cost of one tenth of an hour ($1.20) and 5 minutes of optical jukebox time ($.40) (See #4 above). This figure will decline by 10%/year.

APPENDIX IV—SCANNING DIAGRAM



SCHEMA INDICATING HOW SCANNER INTERPOLATES TO ACHIEVE 600 DPI BITMAP SCANNING (SEE SECTION III B).



The Cornell/Xerox/Commission on Preservation and Access

JOINT STUDY IN DIGITAL PRESERVATION