DOCUMENT RESUME

ED 314 087 IR 053 038

AUTHOR Blamberg, Donald L., Ed.; And Others
TITLE Proceedings of the Conference on Application of Scanning Methodologies in Libraries (Beltsville, Maryland, November 17-18, 1988).
INSTITUTION National Agricultural Library (DOA), Washington, D.C.
PUB DATE 89
NOTE 142p.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Character Recognition; Digital Computers; *Information Systems; *Library Automation; *Library Technical Processes; Machine Readable Cataloging; National Libraries; *Optical Data Disks; *Optical Scanners; *Technological Advancement
IDENTIFIERS Digital Data; *Digitizing; Retrospective Conversion (Library Catalogs)

ABSTRACT

Planned and organized by the National Agricultural Library (NAL), the conference was designed to bring together people involved in scanning projects and to evaluate the feasibility of using hand-held scanning devices for transcribing bibliographic information during the cataloging and indexing processes. Ten previously unpublished papers and a concluding address from the conference are presented in this report: (1) "Automated Document Architecture Processing and Tagging" (Stuart Weibel, John Handley, and Charles Huff); (2) "The Library of Congress Pilot Project with Optiram, Ltd." (Leo H. Settler, Jr.); (3) "Comparison of Scanning Methodologies for Conversion of Typed, Printed, Handwritten, and Microfilmed Materials" (William M. Holmes, Jr.); (4) "Digital Imaging at the Library of Congress" (Audrey Fischer); (5) "Issues in Document Conversion" (Frank L. Walker); (6) "Scanning and Digitizing Technology Employed in the National Agricultural Text Digitizing Project" (Pamela Q. J. Andre, Nancy L. Eaton, and Judith A. Zidar); (7) "Developing an Optical Disk System for Adult Education Manuscripts: The Kellogg Project at Syracuse University" (Terrance Keenan and Elizabeth Carley Oddy); (8) "Access to Little Magazines: An Index of Optically Scanned Contents Pages" (Stephen M. Roberts and Robert J. Bertholf); (9) "Experience with an Optical Disk System at FDA" (Kenneth Krell); (10) "Desktop-Digitization: Prospects and Problems" (Bradford S. Miller); and (11) "Concluding Address" (Robert M. Hayes). (SD)

United States
Department of
Agriculture

Proceedings of the Conference on Application of Scanning Methodologies in Libraries

November 17-18, 1988
National Agricultural Library
Beltsville, Maryland

Compiled and Edited by:
Donald L. Blamberg
Carol L. Dowling
Claudia V. Weston

Photography by Joseph N. Swab

National Agricultural Library • Beltsville, Maryland • 1989

National Agricultural Library Cataloging Record:

Conference on Application of Scanning Methodologies in Libraries (1988 : National Agricultural Library)

Proceedings of the Conference on Application of Scanning Methodologies in Libraries

1. Optical character recognition devices--Library applications--Congresses. 2. Optical scanners--Congresses. I. Blamberg, Donald L. II. Dowling, Carol L. III. Weston, Claudia V. IV. National Agricultural Library (U.S.)

Contents

Preface

This volume of proceedings contains previously unpublished papers presented during the *Conference on Application of Scanning Methodologies in Libraries*. The conference, held on November 17 and 18, 1988, at the National Agricultural Library in Beltsville, Maryland, focused on current projects and techniques using scanning and optical character recognition technologies in the library and information community. The aim of the conference was to provide a forum through which librarians and information professionals could learn the advantages and disadvantages of this emerging technology, gaining knowledge which could open the door to potential applications in their own institutions.

The success of this conference can be attributed to a large extent to the timeliness of the topics presented and to the many talents of those who participated. On behalf of the National Agricultural Library, I would like to extend my appreciation to the conference participants, particularly to the speakers whose papers comprise this proceedings; the moderators, Susan K. Martin, Executive Director, National Commission of Libraries and Information Services; George Thoma, Chief of Communications Engineering, National Library of Medicine; Anne Mathews, Director, Library Programs Office of Educational Research and Improvement, U.S. Dept. of Education; and William Hooton, Director of the Optical Digital Image Storage System, National Archives and Records Administration; and the conference organizers, Robert Hayes, Dean, Graduate School of Library and Information Science, University of California at Los Angeles; Sarah Thomas, Chief, Technical Services Division, National Agricultural Library; and Mary Kahn, Vice President, Education & Training, Washington Consulting Group, whose energies and skills made this conference a reality.

This conference exemplifies the National Agricultural Library's commitment to testing the practical applications of new technologies within the library and information community and disseminating the information gained from its experiences. It is our hope that this conference will be the first of many of this nature hosted by the NAL.

Joseph H. Howard
Director,
National Agricultural Library

Introduction

For decades, librarians have been exploring new technologies as a means of providing better and faster access to the vast array of information resources housed in their collections. In the 1930's, Watson Davis of Science Service and Atherton Seidell, a chemist at the National Institutes of Health, began promoting microfilm as an inexpensive, portable method of reproduction that would provide the scholar and researcher with improved access to realms of information. Late in 1934, Claribel R. Barnett, the librarian of the U.S. Department of Agriculture Library (now known as the National Agricultural Library), allowed Seidell to install a Draeger camera in the library. With this installation, the USDA Library became the first library in the Nation to regularly use photoduplication in lieu of interlibrary loan. In the 1940's, the then librarian Ralph Shaw led the library in the exploration of the use of mechanized indexes. The 1970's saw the development of online databases offering millions of citations and locations, enabling the library user to identify and obtain literature hitherto only retrievable through painstaking manual efforts. With the advent of each new technology, librarians and information specialists have tested its application to the information field. One of the most promising recent developments has been that of the optical digital scanner, which, when combined with optical character recognition capabilities, promises to break new ground in providing access to printed information. The digitized product obtained as a result of scanning expands the reproductive quality of microfilm and offers the possibility of depth of access that far exceeds that of the bibliographic databases now so commonplace.

The *Conference on Application of Scanning Methodologies in Libraries*, planned and organized by the National Agricultural Library, arose out of an evaluation study funded by the USDA. The purpose of the study was to evaluate the feasibility of using handheld scanning devices for transcribing bibliographic information during the cataloging and indexing processes. Cooperating with NAL in the study was the Graduate School of Library and Information Science at the University of California at Los Angeles, which reviewed the state of the art in hand-scanning and reported on its results in testing the effectiveness of three devices with bibliographic materials. Although none of the scanners could perform as desired in an operational setting, the NAL remained interested in scanning technology. The creation of the National Agricultural Text Digitizing Project, in which text is scanned into bit-mapped images and the resulting images are converted into ASCII for enhanced retrieval, reflects NAL's ongoing commitment to evaluate this new technology.

In the process of its investigation into the feasibility of scanning, NAL became aware of several other library projects utilizing scanning technology. Because interest in the topic was pervasive, NAL decided to bring together a group of people with experience in this new technology who would share their findings with other information professionals. Our hope was that a conference on the use of scanning technology in libraries would save others from repeating costly mistakes or taking unwanted detours, while at the same time stimulating others to conceive new applications for this promising technology.

The *Conference on Application of Scanning Methodologies in Libraries* was held in Beltsville, Maryland at the National Agricultural Library on November 17 and 18, 1988. About 125 people attended the conference. Registrants heard 11 presentations on various aspects of scanning, visited a small but highly targeted number of exhibits, and saw demonstrations of several new technologies employed at NAL. Evaluations of the conference were extremely positive, and several conference participants suggested a revisitation of the theme in 12 to 18 months, when the technology has matured and users have gained more familiarity with it.

A reading of the papers which follow will reveal that scanning and optical character recognition are still emerging technologies as far as their applications in libraries are concerned. Nonetheless, they point to the future, a time when redundant keying will be eliminated, and when material formerly known to only a few will be made readily accesible to many. The real promise of scanning technology for libraries is that it will permit further access to information for those who need it to advance the frontiers of knowledge. The proceedings of this conference provide a glimpse into how this might be accomplished.

Sarah E. Thomas
Chief, Technical Services Division
National Agricultural Library

SESSION 1

SUSAN KATHERINE (OROWAN) MARTIN, MODERATOR

Executive Director,
National Commission on Libraries
and Information Science

Dr. Martin received her B.A. from Tufts University in 1963, her M.L.S. from Simmons College in 1965, and her Ph.D. from the University of California, Berkeley, School of Library and Information Studies in 1983. She is presently Executive Director of the U.S. National Commission on Libraries and Information Science and, in that role, is responsible for providing leadership, direction, and planning of national library and information science programs in cooperation with State and local governments and public and private agencies. She is also responsible for developing legislation and advising the executive and legislative branches of Government on the problems and needs of the library and information science community. From 1979 to 1988, she served as the Director of the Milton S. Eisenhower Library of the Johns Hopkins University where she, her staff, the faculty, and the administration planned and implemented policies for the Library to fulfill the research and instructional needs of the University. Dr. Martin also represented the University and the Library in regional and national groups and engaged in development and fund-raising activities on behalf of the Library. From 1973 to 1979, she served as the Head of the Library Systems Office, University of California, Berkeley, where she coordinated the development of automated systems throughout the library and supervised systems analyses in all areas. She participated in librarywide administration, worked with libraries within and outside the University of California system, and participated in a number of specialized task forces. In 1977, she also served as the coordinator of the UCB/Stanford Research Library Cooperative Program, for which she coordinated the UC Berkeley activities of a 3-year, externally funded program to stimulate resource sharing between the 2 libraries. Her activities included fund accounting, management reporting, and liaison with campus ad-

ministration. From 1965 to 1973, she was a Systems Librarian at Harvard University Library, for which she provided systems analysis and design for various units and programming and maintenance of the acquisitions and circulation systems. From 1963 to 1965, she served as an intern in the Harvard College Library, where she was exposed to cataloging, reference, and data processing activities. In addition to these professional experiences, Dr. Martin has served the library community in a variety of teaching, consulting, and advisory capacities.

Dr. Martin is presently a member of the American Library Association (ALA) Council and has served the ALA in various other capacities such as being the Chair, President, Chairman, and member or resource person for a variety of ALA associations, committees, boards, subcommittees, and delegations. She has also had a number of roles in the American Society for Information Science, the Association of Research Libraries, the Universal Serials and Book Exchange, the Philadelphia Area Library Network, and the Research Libraries Group, Inc. She is presently serving on the Editorial Board of the "Journal of Library Administration" and on the Board of Consultants of the "Advanced Technology/Libraries." She has also been on the Board of Contributors for "Library Issues: Briefings for Faculty and Administrators" and served as editor for the "Journal of Library Automation."

Automated Document Architecture Processing and Tagging

Stuart Weibel, John Handley, and Charles Huff
OCLC Office of Research

Abstract

Optical character recognition will assume increasing importance as a tool for building databases from printed documents. This paper describes efforts in the OCLC Office of Research to quantitatively and qualitatively characterize some of the limitations of OCR and identify means to minimize these limitations. These efforts include (1) the development of test procedures to improve our understanding of the variety of errors encountered with OCR technology and (2) the prototyping of systems designed to improve OCR processing and to incorporate OCR output in structured databases suitable for electronic document retrieval and delivery systems.

OCR and OCLC: Overview

The paper document remains the dominant information vehicle in business and academics, in spite of increasing emphasis on electronic media. Processing such documents will remain an important consideration in information delivery for the foreseeable future. Converting paper-form information to electronic media constitutes a costly, but essential, aspect of bridging the paper-bound present to the electronic future.

The options for crossing this bridge are few and costly: discard the data, rekey the characters, or employ optical character recognition (OCR) technology to automate the capture of the information. The choice for a given body of information will depend on the value of the information and the resources available for conversion.

Although still a developing technology, OCR is emerging as an important method for capturing data from paper. This paper describes efforts currently under way in the OCLC Office of Research to evaluate OCR capabilities and identify methods for improving the quality of systems based on this technology.

OCLC and CMU: Project Mercury

OCLC and Carnegie Mellon University (CMU) are jointly pursuing a broad range of research activities aimed at prototyping an "electronic library." Dubbed **Project Mercury**, the goal of this project is to provide seamless access to an on-line electronic library. The operational objective will be met when a scholar can access the majority of relevant information resources directly from his or her workstation, irrespective of the physical location of such resources.

Inclusion of existing paper documents in an electronic library environment requires more than simple character recognition. Documents can be made more accessible and useful by structuring and tagging their contents. Identification and tagging of document structure (markup) by humans is a costly and time-consuming process. Our re-

search addresses automation of certain aspects of the markup as well as the character recognition of paper-form documents.

In an ideal world, all electronically published materials would begin life in a descriptively specified markup format, readily exchanged among publishers, database vendors, and authors. The reality of our environment is such that many conflicting systems will need to be reconciled to provide a coherent retrieval and delivery environment. The conversion of paper documents to electronic format is among the most problematic of these challenges. Our current research on enhancement of OCR output addresses this need for economic digitization of paper material.

Evaluation of OCR Performance

Benchmarking the performance of OCR systems is fraught with difficulties which include rapidly changing products, variation in source document quality, style and size of font, and resolution of the imaging technology. Measuring the progress of our own efforts, as well as technological advances of the devices themselves, requires a systematic and replicable approach to performance evaluation. We have undertaken to establish a series of test documents and a suite of test procedures that will enable us to accurately gauge the performance of OCR devices and to understand the variety of errors these devices make.

We have preliminary procedures and data that provide some insight into the operational characteristics of OCR devices. The data presented here represent a preliminary comparison of the performance of two of the more powerful OCR devices currently available; they should not be construed as a definitive benchmark of the capabilities of these machines.

The measure of accuracy used here is based on "string edit distance," i.e,. the minimum number of edit operations (insertion, deletion, or substitution) needed to make the actual recognized document identical to a normalized, correct version. The recognized version was compared with the reference version on a page-by-page basis. Both the recognized and corrected versions were normalized by removing multiple contiguous white spaces (i.e., blanks and newlines). The comparisons were done using an algorithm by Wagner and Fischer [5].

Preliminary Results:

Table 1 summarizes our preliminary results with a small sample of test documents. Per cent error rates refer to the edit distance divided by the number of characters in the reference document.

Table 1. Comparison of edit distance statistics for a series of test documents.

			Calera			Kurzweil		
Document Name	pages	correct chars.	recog. chars.	edit dist.	% err.	recog. chars.	edit dist.	% err.
Quote	3	6,045	6,048	6	0.1	6,024	34	0.6
Fonts1	6	3,822	3,875	345	9.0	3,505	731	19.1
Fonts2	10	26,166	26,786	1,263	4.8	26,214	1648	6.3
LISP	3	2,687	2,756	99	3.7	2,608	157	5.8
Test2	1	2,072	2,092	48	2.3	2,078	39	1.9
Total	24	40,792	41,556	1,761	4.3	40,429	2609	6.4

Test Document Descriptions:

Quote: A brief (2 paragraph) book excerpt. Each page of the document consisted of the same quotation set in a different typeface, using the TEX computer typesetting program. The three 10-point typefaces used were TEX's default Computer Modern Roman, a Computer Helvetica, and "typewriter."

Fonts1: This document contained short samples of various typefonts, generated using TEX and troff. Each sample included all printable ASCII characters, as well as ligatures such as 'fi', 'fl', etc. Fonts used included two kinds of Roman, Computer Helvetica, and several kinds of "typewriter." Type styles included plain, slanted and unslanted italic, and boldface. Two pages were set at 12-point, the others at 10-point.

Fonts2: Ten font samples, photocopied from a published type-specimen book (by V & M Typographical, Inc.). Figure 1 illustrates the differences in recognition performance among the fonts. Figure 2 illustrates samples of some of the font styles.

LISP: Fragments of LISP code, originally printed on a Xerox x2700ii laser printer. The originals were printed in landscape format, so it was necessary for both scanners to rotate the image 90° before performing recognition.

Test2: This was another short quotation, set in various typefaces using TEX, which featured changes in typeface within the text and extensive use of ligatures, such as 'fi.'

Mixed Font Performance

These data illustrate the high degree of variability in the recognition performance among various fonts and between the two devices we tested. Percent edit distance error ranged from a low of 0.1 percent (1 error per 1,000 characters) to a high of 19.1 percent (almost 1 error per 5 characters).

Comparison of Font-Specific Performance

The *Fonts2* sample in table 1 consisted of 10 font styles in a variety of sizes. Figure 1 details the OCR performance of these fonts. Figure 2 illustrates several of the font styles in 10-point faces. Unfortunately, the comparisons are not completely equivalent because of the differences in number of words in each sample and differences in the number of sizes for each font. Nonetheless, this figure illustrates that font styles have significant impact on OCR performance. There is a distinct bias in favor of simpler, sans serif fonts.

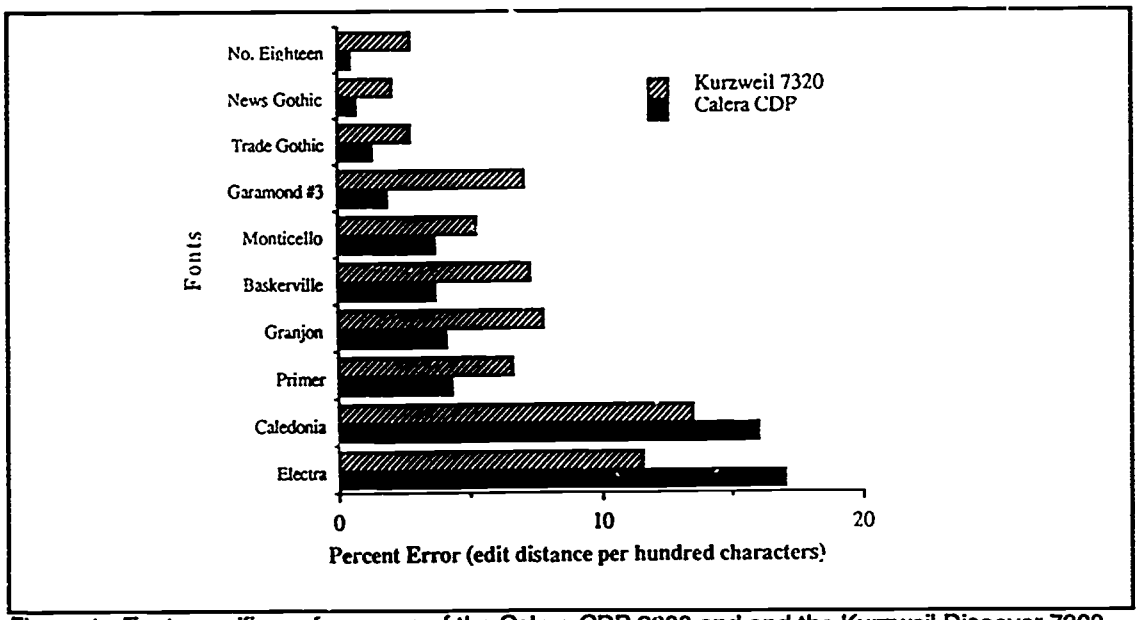


Figure 1. Font-specific performance of the Calera CDP 9000 and and the Kurzweil Discover 7320 OCR devices.

11 point Number Eighteen (Linotype)
The basic character in a type design is determined by the uniform design characteristics of all letters in the alphabet. However, this alone does not determine the standard of the
ABCDEFGHIJKLMNOPQRSTUVWXYZ&
abcdefghijklmnopqrstuvwxyz 1234567890$

11 point News Gothic Condensed (Linotype)
The basic character in a type design is determined by the uniform design characteristics of all letters in the alphabet. However, this alone does not determine the standard of the type face and the quality of composition set with it. The
ABCDEFGHIJKLMNOPQRSTUVWXYZ&
abcdefghijklmnopqrstuvwxyz 1234567890$

11 point Caledonia (Linotype)
The basic character in a type design is determined by the uniform design characteristics of all letters in the
ABCDEFGHIJKLMNOPQRSTUVWXYZ&
abcdefghijklmnopqrstuvwxyz 12345678905
ABCDEFGHIJKLMNOPQRSTUVWXYZ

11 point Electra (Linotype)
The basic character in a type design is determined by the uniform design characteristics of all letters in the alphabet. However, this alone does not determine the
ABCDEFGHIJKLMNOPQRSTUVWXYZ&
abcdefghijklmnopqrstuvwxyz 1234567890$
ABCDEFGHIJKLMNOPQRSTUVWXYZ

Figure 2. Examples of typefaces from the font-specific comparisons in figure 1.

The data in figure 1 reflect OCR performance on actual samples of printed fonts, but the sample is small in size and not precisely equivalent. To evaluate the performance of the Calera CDP 9000 on a realistic conversion project, an entire book was set in six typefaces using TEX. The book, *Joseph Heller,* is part of the Twayne Series on American Authors. The text of this volume was available to us in machine-readable form, making it relatively simple for us to generate the alternative forms of the book typography.

Raster images of the book pages were rendered at 300 dots per inch in each typeface. These noiseless page images were sent directly to the OCR device with no optical scanning. In addition, the six different versions of the book were printed on standard copier paper using a laser printer at 300 dots per inch. These were processed by the Calera CDP 9000 in the usual way using the native scanner. The results are summarized in table 2.

Table 2. Summary of edit distance statistics for rendered and scanned versions of *Joseph Heller* in each of six fonts.

	Times Roman	Century Schoolbook	Computer Helvetica	Elite	Pica	Computer Modern
pages	144	148	154	186	218	146
characters	371,326	372,240	372,564	375,965	377,664	371,812
rendered sum of edit distances	2,153	9,349	2,223	3,362	5,114	1,806
% error	0.58	2.25	0.60	0.89	1.35	0.49
scanned sum of edit distances	2,206	4,478	7,106	1,682	2,078	4,481
% error	0.59	1.21	1.91	0.45	0.55	1.21

As with earlier samples, the font style exerted a major influence on the recognition error rate. It is curious that noise actually decreases the error rate in some fonts. This may be because some characteristics of the letter forms in the noiseless versions consistently confuse the recognition algorithms. For example, in the Century Schoolbook font, the dot of the lower case 'i' and 'j' is just a bit too high above the body. Depending on the word context, the dot is taken for an apostrophe or a comma in the line above. In the scanned version which contains noise introduced by the optical scanning process, the dots of the 'i' and 'j' are slightly bigger and the gap between the dot and the body will be smaller much of the time. It appears that noise can help overcome biases in the recognition algorithms.

The best recognition rate achieved was in the neighborhood of 99.5 percent, whether the image was rendered electronically or scanned in the usual manner. This may represent an ideal maximum for today's technology, especially considering that these books are of very simple layout with few tables, line drawings, or halftones.

The MARC-UP Catalog Card Conversion Prototype

Goals

Automating the analysis of document architecture will be a long and complex process of incremental development. Our first attempt at this problem is a relatively simple instantiation of the larger problem, but one that targets an immediately useful application: the conversion of catalog cards to machine-readable (MARC) format. This labor-intensive process represents an important component of the services that OCLC offers to members. Automating significant parts of this task will result in substantial economic savings.

The process currently begins with an operator at an M300 workstation searching the on-line union catalog (OLUC) for a matching record. If the record is identified on the OLUC, the appropriate holding symbol is added to the record and the operator proceeds to the next card. Such a transaction is referred to as an *update*. If no match for the card is found on the OLUC, the card must be keyed in manually and is referred to as an *input*.

The current rate of entry for updates is in the range of 30-40 per hour, while input records are entered at a rate of only 3-6 per hour. Our **MARC-UP** prototype system is intended to demonstrate a model for automating such processing and to provide a foundation for more advanced document processing systems.

The basic assumption of this project is that handling of the actual catalog card should occur only when it is scanned. All subsequent processing should take place either in batch mode or on a workstation designed to facilitate every aspect of searching, data capture, correction, and validation of card information.

The processing model is segmented into 6 components:

1. Image capture,
2. Image preprocessing,
3. Optical character recognition,
4. Postprocessing output filtering,
5. Rule-based inferencing, and
6. Operator validation.

Image capture

Scanning of a card image can be done with any high-quality scanning device. We are currently using 400-dpi images captured on a flatbed scanner; optimum image resolution will be determined experimentally.

Image preprocessing

There are several possible image-enhancement steps that are likely to result in greater accuracy and efficiency in the conversion process. These include de-skewing, segmentation of card images, and classification of resulting segments.

De-skewing of images would be desirable to increase the accuracy of subsequent OCR processing. A scanned image may deviate from orthogonal orientation resulting from misalignment in the scanning process or poorly aligned original typesetting. Current recognition technology will accept approximately 2 degrees of skew; further deviation leads to higher error rates. Although we have not implemented de-skewing in this prototype, it could be done as part of initial batch image processing of scanned cards.

Segmentation of card images is a key step in our processing model. This step identifies the bounding box coordinates for all text and noise regions in an image. The segmentation method used in this demonstration is based upon the paper by Wong et al. [6].

The catalog card image is segmented in a five-step process:

1. The image is smoothed horizontally by converting all contiguous runs of fewer than 123 white pixels to black pixels. This has the effect of smearing the image in the horizontal direction.

2. A similar procedure is performed in the vertical direction on the original image, except that runs of white pixels of length 833 or less are converted to runs of black pixels.

3. The two smeared images are combined using a logical *AND* operation, pixel by pixel.

4. Next, another horizontal smooth is performed; this time white run lengths of less than or equal to 50 are converted to black. The result is an image consisting mostly of black boxes.

5. The final step is to find bounding boxes for all of the contiguous black areas.

Feature Extraction and Classification

A set of 14 features for each object in the original document image is extracted. The aggregate of these features is used to classify each box as either noise or text. Two classification techniques have been applied to discriminate between noise and text. In the first, a classification tree is built using the Classification and Regression Trees (CART) technique [1]. To classify an object, the classification tree examines one of the features at a time and either classifies the object as noise or text or chooses another feature to look at. Eventually, the object falls into one of the two categories.

Optical Character Recognition

Optical character recognition processing of the enhanced image is accomplished with a Calera CDP 9000 OCR recognition server. This device is among the best OCR machines currently available.

Recognition accuracy is partially determined by the quality and nature of the input material. Catalog cards are neither the best nor the worst source material in this regard.

One advantage of our segmentation process is that small, "clean" text fragments can be sent to the Calera for recognition, a process that we believe will enhance recognition as compared to full-image processing. In a test set of 50 cards, edit distance per character was reduced from 0.14 to 0.09 (a drop of 36 percent).

Although the current development system takes advantage of the particular characteristics of the Calera CDP, the overall approach is independent of the specific nature of the OCR component; we anticipate that advances in this area may dictate that other OCR systems will merit consideration, and the system design incorporates that philosophy.

Postprocessing

Postprocessing filters can include anything from a conventional spelling checker to application-specific authority checking, context analysis, or key word trapping. Here again, the modular nature of the system will allow for incremental improvement or modification of the system. Many OCR errors fall into predictable patterns; *u's* and *n's* can often show up as *ii's*. One of the goals of this research is to achieve a better understanding of the error patterns of OCR devices that will support the building of effective postprocessing filters to improve OCR output.

Rule-Based Inferencing

In most applications, the layout attributes of text blocks will convey information concerning the functional role of the block. A rule processor will allow the construction of modular rules that can facilitate the automated selection of role assignments for text objects.

The current implementation includes simple examples for the identification of call number and LC card number. In a test sample of 50 cards, the call number was recognized correctly only 42 percent of the time, but the LC card number was captured correctly 96 percent of the time. More elaborate rules will provide automated identification for other fields on the card, further streamlining the capture process.

Fields extracted in this manner can be used to formulate search keys that will retrieve potential record matches from the Online Union Catalog. These records can then be ranked by similarity and presented to an operator as possible matches, thereby reducing the time required for an operator to identify a matching record already in the database.

Operator Validation

Few conversion projects will find current OCR error levels satisfactory for totally automated conversion. An operator will be required to correct and validate the records created with this system. It is likely that a smooth and facile workflow model will ultimately have greater impact on the success of this system than the error rates of the OCR device. Figure 3 illustrates what the user-interface for such a conversion tool might look like.

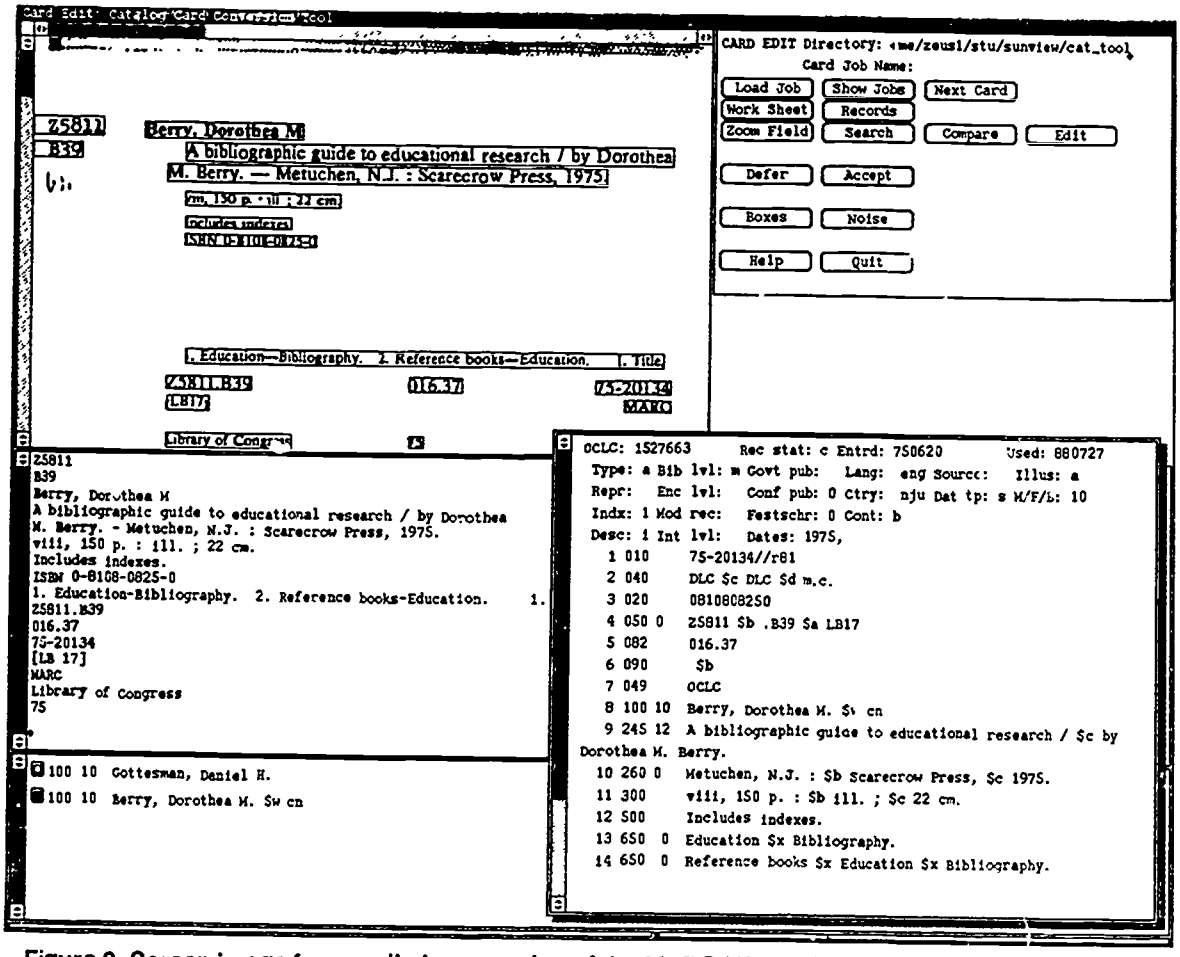


Figure 3. Screen image for a preliminary version of the MARC-UP catalog card conversion prototype.

The general goals of this phase are to:

- automatically formulate search keys for identification of potential matches,
- provide high contrast between differences in compared fields or records,
- automate as much error detection as possible, and
- provide an optimized editing and correction facility that will minimize correction keystrokes.

Project ADAPT: Automated Document Architecture Processing and Tagging Goals

The overall goal of this project is to develop an integrated system of document processing that will digitize, markup, index, and structure for retrieval paper-form journal articles, all with a minimum of human intervention. The processing model is similar to that for the MARC-UP prototype, with several significant additions, figure 4.

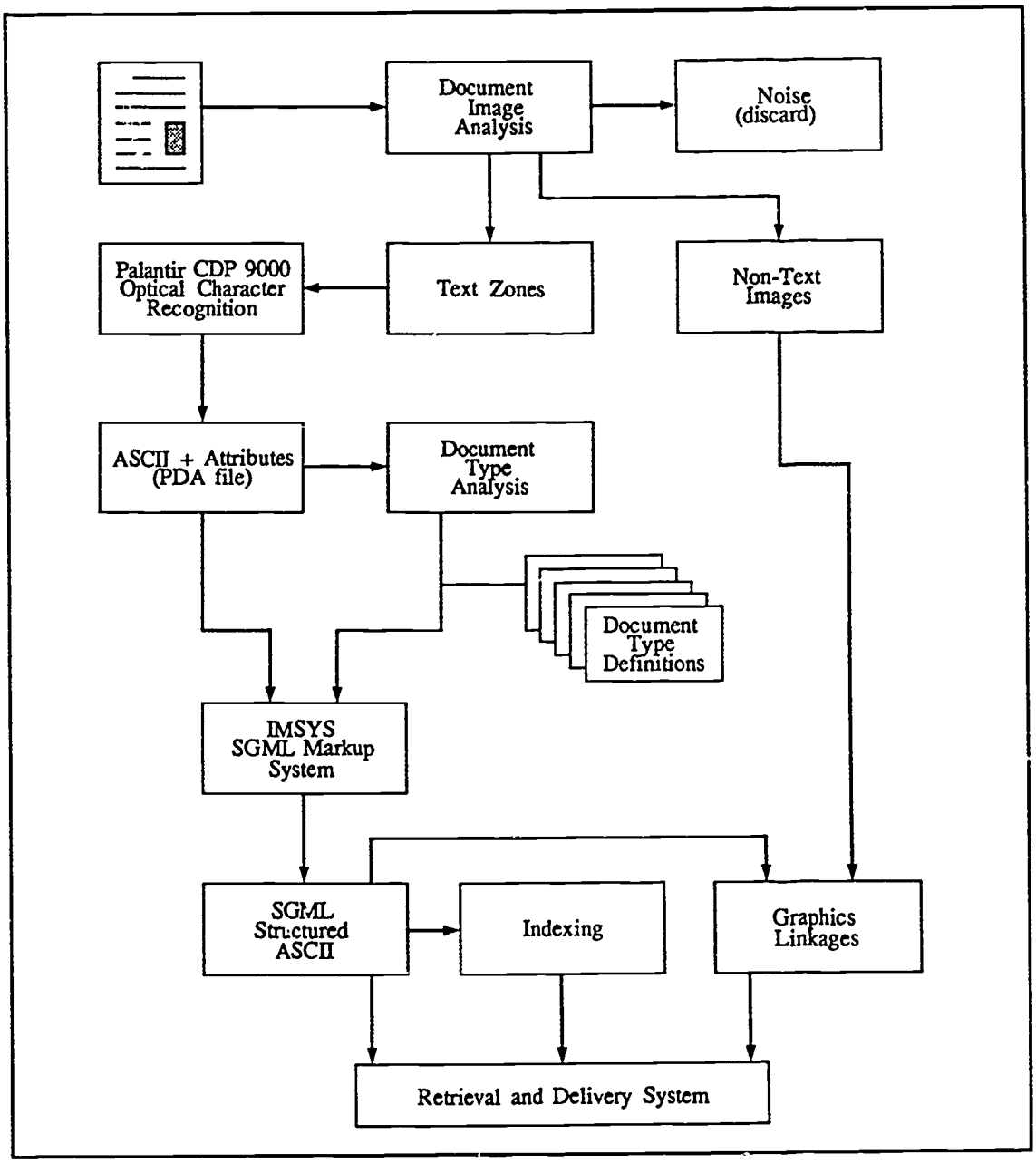


Figure 4. System block diagram for the ADAPT project.

A major point of departure between the systems has to do with markup. Catalog cards have substantial variety of layout, but the diversity of journal article layout, for example, is far greater and will require a more elaborate approach to process automatically.

Similarly, building automated document-processing systems that build databases of retrievable objects that can be displayed on a variety of display devices presents challenges far beyond that of the catalog card conversion problem. It is our belief that automated markup analysis can be expected to provide answers for some of these challenges.

Automated Markup

Markup serves two major functions in electronic document-processing systems; the first is the conventional notion of formatting a display, whether a sheet of paper or a video display screen. This role is often referred to as *presentational markup*, and is incorporated in most conventional text-processing systems (indeed, typing on a conventional typewriter implicitly incorporates many notions of presentational markup in the form of margins, spacing, pagination, and the like).

The second major role of electronic markup is oriented toward the structure and content of the document. By identifying the hierarchical organization of the document (headings, subheadings, paragraphs, etc.) and structural elements of the document (titles, authors, abstracts, figures), descriptive markup provides the capability to generate multiple views of a document and provide access to the substructure of that document in ways that substantively enhance its value to an author or scholar (see Coombs et al. [2] for an illuminating explanation of the advantages and power of descriptive markup).

Standard Generalized Markup Language (SGML) predominates as the standard descriptive markup language. Only a small number of commercially available systems now implement SGML, but the economic benefits of its use to authors, editors, publishers, and database providers will promote more widespread use as developers move to support its implementation. The building of translators and parsers for interconverting between various standard markup systems and SGML is now an active research area and is likely to grow (e.g., [4]).

SGML provides useful hooks for document components that can be used to implement advanced document retrieval, display, and delivery systems. For example, OCLC's Graph-Text system uses SGML tagging to identify document elements in this manner [3].

The IMSYS (Intelligent Markup SYStem, Avalanche Development Corporation) generates validated SGML markup by analyzing layout objects in a document and using a *Document Type Definition* to assign roles to these objects.

Issues to be researched here are (a) the extent to which this system can successfully tag materials from an arbitrary collection of journal articles and (b) how the document type definition can be selected automatically from among probable document types. The latter problem might well be addressed by abstracting features from the document layout and discriminating among patterns with some combination of a rule-based system and pattern-recognition approaches.

Conclusions

We expect that substantial gains in productivity will result from the availability of tools that facilitate OCR-based capture of printed information. Offshore keying of data is more economical for many applications, but suffers from lack of local control, lack of quality control, and turnaround time. It will never be an acceptable option for build-

ing personal databases, for example; every scholar wants his or her own personal system to convert stacks of paper to structured databases.

There are no easy or quick fixes for making OCR work better, but with increasingly powerful processors and software systems occupying the scholar's desk, the role of OCR-based systems will inevitably increase in importance. Our work is directed toward prototyping such systems with the expectation that genuinely useful information management tools will result.

References

[1] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J.Stone. *Classification and Regression Trees. The Wadsworth Statistics/Probability Series,* Wadsworth International Group, Belmont, California, 1984. ISBN 0-534-98054-6.

[2] James H. Coombs, Allen H. Renear, and Steven J. DeRose. Markup systems and the future of scholarly text processing. *Communications of the Association for Computing Machinery*, 30(11), November 1987.

[3] Thomas B. Hickey and John C. Handley. Interactive display of text and graphics on an IBM-PC, *Impact of New Information Technology on International Library Cooperation*, Essen Symposium, Essen, German Federal Republic, 1986.

[4] Sandra A. Mamrack, M.J. Kaelbling, C.K. Nicholas, and M. Share. A software architecture for supporting the exchange of electronic documents, *Communications of the Association for Computing Machinery*, 30(5):408-414, 1987.

[5] Robert A. Wagner and Michael J. Fischer. The string to string correction problem, *Journal of the Association for Computing Machinery*, 21(1):168-173, January 1974.

[6] K.Y. Wong, R.G. Casey, and F.M. Wahl. Document analysis system, *IBM Journal of Research and Development*, 26(6):647-656, November 1982.

STUART WEIBEL

Research Scientist,
Office of Research, OCLC

Stuart Weibel is a research scientist in the OCLC Office of Research. His research interests include artificial intelligence, automated cataloging, automated document structure analysis, and connectionist learning algorithms. Dr. Weibel's current projects include the development of a system prototype to automate aspects of retrospective conversion of catalog cards and a system prototype to investigate the feasibility of automated document structure analysis to facilitate the building of structured databases for arbitrary collections of paper documents.

The Library of Congress Pilot Project with Optiram, Ltd.

Leo H. Settler, Jr.
Library of Congress

Abstract

The Library of Congress (LC) became aware of Optiram's capabilities in the fall of 1984. After some initial discussions, a contract with Optiram was let to work with LC to convert printed cards from the Cataloging Distribution Service's Record Set (the non-MARC card set) into machine-readable form. The early stages of the project concentrated on Optiram's development of format recognition techniques to enable them to output scanned bibliographic data in the USMARC format. Later in the project it was decided, for technical reasons, to concentrate the investigation on the potential for converting NUC reports and serial holding records rather than cards from the CDS Record Set. Several sets of records were converted with generally good results. LC is currently investigating additional similar projects with Optiram.

Initial Contact

Optiram is a London-based automation firm that has developed a highly sophisticated optical scanning capability. They are currently able to scan printed, typed, or handwritten documents in all roman-alphabet languages and convert them into machine-readable form. The first contact that the Library of Congress (LC) had with this company was in October of 1984 when Mary Price, LC's Director for Bibliographic Products and Services, was in London attending a meeting of the IFLA Newspaper Working Group. Someone at the meeting told her about the optical scanning capabilities of Optiram and she then went to their London office for a demonstration. A page of her hand-written notes from the meeting was scanned by their equipment, digitized, and printed out with a high degree of accuracy. After this impressive demonstration, she felt that this technology was worthy of further investigation for possible applications at the Library of Congress.

In November, Frederick K. Ibbotson, Optiram's International Sales Director, came to LC to discuss how their scanning capabilities might be used by the Library. He met with several Processing Services' managers, including Henriette Avram, the Assistant Librarian for Processing Services; Gerald Lowell, the Chief of the Cataloging Distribution Service (CDS); and Mary Price. Later that month, Gerald Lowell sent Optiram 200 printed catalog cards and a copy of the *MARC Formats for Bibliographic Data* to examine. After Optiram had reviewed these materials, it was decided to go forward with a small pilot project to see what capabilities Optiram could develop for using artificial intelligence to format scanned data into the MARC communication format. A contract was let in April of the next year for the conversion of 30,000 roman-alphabet records into machine-readable form. The objective was to determine if machine-readable records created by automated means could replace the CDS card files and could also be used as part of the various MARC distribution services. An important

requirement, however, was that immediately upon receipt, these records must be ready for internal use or distribution without human intervention or review.

Development of Format Recognition Capability

In May, Optiram contracted with another British firm, LIBPAC Computer Services, to develop the capability to format data scanned and converted into digital form by Optiram into the MARC communication format by means of artificial intelligence. LIBPAC's president, Martin Harrison, spent several days at LC discussing the Optiram contract and the Library's experience in developing format recognition software. As a result of these discussions, everyone involved with the project began to consider seriously some "less-than-ideal" options, including limitation of the project to English language monographs, and concluded that the project would have to be extended beyond the originally established deadline of September 30, 1985.

Harrison prepared a "discussion" conversion document outlining some possible solutions to the problem of formatting scanned data into the MARC communication format. In November, he and Ibbotson came to LC to discuss the current status of the project. By this time, it had been determined that the development of a fully functional format recognition capability within the allotted time frame was a much more formidable task than had been originally anticipated. There were frank discussions about the advisability of cancelling the contract, but, after a careful review of the options, it was decided to proceed with the project, with the proviso that additional flexibility in terms of the output requirements would be acceptable. The completed conversion document was received by March 1986.

Decision Point in CDS

By the fall of 1986, Gerald Lowell had left CDS and Susan Tarr was now the Acting Chief. She contacted Optiram to determine the current status of the contract. Although Optiram had clearly demonstrated their ability to convert data into machine-readable form, they were continuing to encounter problems with the formatting of these data into the MARC communication format. Optiram advised Tarr that they had engaged a new subcontractor to find a solution to this problem and to produce the tapes for shipment to LC This company was RHM Computing, the firm that creates MARC tapes for the British Library.

In January 1987, Optiram's president, John Jenkins, came to Washington to talk with staff of the National Archives about a possible project there. Staff from the Library's Serial Record Division also attended that meeting in order to determine if any of their data conversion needs might be candidates for scanning by Optiram. They gave Jenkins samples of two kinds of records found in the serial record. The first were records from the Visible File which are recorded on 4- by 6-inch sheets and represent most of the live titles being received at LC. The other records were from the so-called "3- by 5-inch Serial Check-in File." This is an older file that is housed in standard card drawers and contains various kinds of printed and hand-written cards for mostly dead serial titles. Later at a meeting with Ibbotson at LC, the processing of contributed card reports

to the *National Union Catalog* was discussed as another possible application for scanning by Optiram.

New Focus to Contract

By February, it had become quite clear that applications where records would be reviewed and edited before being used in an on-line file would be better candidates for this type of optical scanning than those such as the conversion of cards from CDS. It appeared that the problems associated with developing the artificial intelligence necessary to format data into the MARC communication format could only partially be solved. Although it seemed that some basic level of formatting could be achieved, it would probably not be sufficient, at least at this stage of its development, to allow records to be used in an on-line file or distributed as part of the MARC distribution service without record-by-record review and editing. The contract with Optiram was revised to specify contributed records to the *National Union Catalog* and cards from the Serial Record "3- by 5-inch Serial Check-in File" as the source documents for Optiram's conversion effort rather than cards from the CDS Record Set.

In March, meetings were held with the Catalog Management and Publication (CM&P) and Serial Record Divisions to develop specifications for converting their data. The total number of cards to be converted would be fewer than the originally anticipated 30,000. CM&P would select appropriate card reports and group them into sets having common characteristics. Serial Record would also have a microfilm copy made of one drawer of cards from the "3- by 5-inch Serial Check-in File."

The source documents were sent to Optiram during April and May 1987. In April, John Jenkins was again at LC and met with members of the Library's technical processing staff to discuss some of the details involved in using the MARC communication format. In May, Ibbotson returned to LC to clarify some final details of the specifications from the CM&P and Serial Record Divisions.

On August 17, 1987, LC received the first tape of converted records--a sample of unedited NUC reports. We created a hexidecimal printout of the tape for analysis and loaded the records successfully into a test partition of the on-line NUC file. Everything worked well and it was very exciting to see on-line displays of these machine-converted records. I immediately contacted Optiram and told them to proceed with the production of all deliverables with the software as then in production. The last tapes were received in March 1988. After review of the records, payment for the contract was made in June, concluding the project.

Description of the Source File

The source records selected for the project consisted of 3- by 5-inch cards selected by the Catalog Management and Publication (CM&P) and Serial Record Divisions. The CM&P Division compiled three groups of cards representative of the kinds of records used in the creation of the *National Union Catalog (NUC)*. Group 1 consisted of about 2,000 unedited reports to the NUC. Group 2 consisted of about 2,000 reports containing various kinds of editorial markings in pencil. Both Groups 1 and 2 were cards ex-

hibiting a very consistent layout. Group 3 consisted of about 4,000 cards which had been typed in CM&P for reports published in the 1958-62 cumulation of the NUC. The records from the Serial Record Division consisted of a drawer of about 1,200 cards selected from the Division's old "3- by 5-inch Serial Check-in File." The Serial Record cards were filmed by the Photoduplication Service and a 35-mm negative film was sent to Optiram.

Description of the Requirements

Optiram was to scan the source documents and provide the Library with tape files in the USMARC communication format along with printouts for these documents. The specifications for USMARC as given in the *MARC Formats for Bibliographic Data* were to be followed as closely as possible, but we acknowledged that it was probably not possible to accomplish this completely within the time frame of the contract. We therefore suggested some default values which would be acceptable at the time. Some of these defaults were the following:

- Subfield Coding -- The development of the ability to identify the location of subfields progressed quite smoothly, but the development of the logic necessary to provide the correct alphabetic code for each subfield proved to be a very difficult task. We decided that whenever the proper subfield code could not be identified, the code "a" should be used.
- Indicator Values -- Determining the correct setting for the indicators was also a problematic area. Here we advised that a default setting could be established for each variable tag which has defined indicators. The default chosen would be the value that is most common for that tag, e.g., 100 10 or 245 10.
- 008 Field -- It was obvious that several of the values in the 008 field, such as the language of the publication, might be difficult to record correctly without developing special logic routines to determine them. Defaults were suggested for these values.

Both CM&P and Serial Record also had some specific requirements for processing their individual records. For the NUC reports, there was the requirement for recording in an 040 field the NUC symbol of the reporting library in subfield "a" and the symbol OtR (Optiram) in subfield "c." The NUC symbol also was to be recorded in field 850. For the edited reports, there were various kinds of editorial markings such as checks and circled data that were not to be converted. Editorial additions that were "ballooned in" were to be placed in the indicated locations of the appropriate data fields. Some nonbibliographic information given at the bottom of the cards was not to be recorded in the converted record.

The records from the Serial Record Division were sent purely as an experiment. The conversion capability developed so far by Optiram was specifically for monographs and these were serials with added notations indicating holdings, routings, selection decisions, etc. We asked that these records be scanned and formatted using the same software being used for the NUC records and that any data that could not be identified for inclusion in any of the standard variable fields for monographs be placed in multi-

ple 500 notes. It was expected that holdings data, for instance, would be recorded as 500 notes.

Evaluation of the Results

NUC Records

In order to evaluate the results of the conversion of these approximately 8,000 records, a sample was taken and analyzed. We selected 130 records from each of the three groups and evaluated the hexidecimal printouts for each. There were two objectives to the review. The first was to determine the accuracy of the scanning and the second was to determine how well the scanned data were formatted into the MARC communicatic n format.

The results were quite impressive. The analysis revealed a scanning accuracy of well over 99 percent. In fact, there was only one character identified in the sample records that was clearly missing because of a scanning error. Dates in personal name main entries were often not recorded, but this seemed to be caused by a programming error in the formatting software rather than a scanning error.

As expected, various inconsistencies were encountered in the tagging of the scanned data. The control fields were tagged quite consistently and the suggested default values were used in the 008 field except for the publication date, which was picked up from the imprint.

Most of the problems seen in the tagging of the variable fields seemed to have resulted either from not having the full range of variable tags available from the software or perhaps from not having the logic in place to identify the appropriate data that should be placed in these fields. All main entries were coded 100, the 250 field was lacking, subjects were all coded 650 or 651, and all added entries were coded 700 or 740. All variable fields had the indicators set to default settings and had all subfields coded "a." The proper identification of some subfields seemed to be linked to the presence of ISBD punctuation.

Headings for corporate bodies posed quite a problem. Short corporate bodies used as main entries were coded 100, but longer ones were placed at the beginning of the 245 field. As added entries, some corporate bodies were coded 700, but most were coded 740.

All things considered, this was not a bad accomplishment for a first attempt at creating MARC records by means of optical scanning and artificial intelligence. We found that the records were structurally sound and passed our error-checking programs and were successfully converted into our internal on-line format. We were able to load a few of the records into a test partition of the on-line NUC file and were able to use a utility print program to print some of them in a card image format.

Serial Record Records

The records from the Serial Record "3- by 5-inch Serial Check-in File" might best be characterized as a "challenge." In the first place, the records in the file comprise a wide variety of formats and styles of data ranging from printed LC cards to hand-written entries. Also, the source document that Optiram scanned was a paper print which they made from the microfilm we had sent. A representative group of 18 records was analyzed to try to determine if this might be an avenue worth pursuing for the conversion of these data into machine-readable form.

As was the case with the NUC reports, the actual scanning was very accurate for those data which could reasonably be expected to be recognized on the paper printout. The major problem that was seen with this approach, however, is that we would expect to load these records into one of our on-line files, the SERLOC file, without any kind of human review. Since the software used to create these records was actually developed for monographs, it had some predictable problems in formatting serial records. However, the major problem is the question of how to record holdings. All holdings were, as required in the specifications given to Optiram, recorded in 500 notes. It is not certain, however, that this convention would produce usable results if these records were loaded into the SERLOC file without some modifications. More analysis will be required before we can ascertain if this is, in fact, a productive path to follow.

Plans

The Library has recently let another small contract with Optiram, again to gain further information about their capabilities. For this project, Optiram will convert photocopies of 6,000 cards from the Library's official catalog used in technical processing by Processing Services. The records will consist of two groups. The first will consist of records which the software should have little problem formatting. The second group will consist of a small set of records containing a variety of what would seem to be scanning and/or formatting problems. The objective of this is to see just what the current version of the software can do with these various kinds of problems. An analysis of the converted records will help us to decide how Optiram might fit into any future retrospective conversion efforts at the Library of Congress.

Conclusion

Optiram clearly has a very highly developed ability to convert printed and hand-written data into machine-readable form. Its ability to format such data into a USMARC record is still at a basic level. Whether or not records created with this software, at its current stage of development, would be adequate would depend on the individual requirements of any specific application. In the workflow of the Library of Congress' Catalog Management and Publication Division, where reports are reviewed field-by-field and edited in order to bring them into conformity with national bibliographic standards, there seems to be a very good possibility for success. However, with other applications, this may not be a feasible approach.

We have recently begun discussing with Optiram the characteristics of additional development that would be required to upgrade the artificial intelligence in order to

provide a more complete USMARC record. Although both feel that enhancements would be possible, there is perhaps a point beyond which the benefits of such enhancements would not justify their cost. It is not expected that Optiram would invest more of their own resources into software development at this time unless it were clear that they would be able to recoup that investment from substantial future ongoing contracts.

Scanning technology is viewed by many as perhaps the one means by which the international library community has a chance to convert its still-huge files of manual cataloging records into machine-readable form within a reasonable time frame. Perhaps there is a basic level of content designation that would be feasible for Optiram to develop that could provide MARC records for use in on-line systems, at least for basic functions. The formulation of the specifications for this basic level of content designation would require input from various interested members of the library automation community. Much additional analysis and assessment will be required before we can be sure how this technology will actually fit into our future automation planning efforts.

LEO H. SETTLER, JR.

Assistant to the Director,
Bibliographic Products and Services,
Library of Congress

Leo H. Settler, Jr. attended the University of Michigan, completing his Ph.D. in Musical Arts, and the Catholic University of America, completing his M.L.S. He is currently the Assistant to the Director for Bibliographic Products and Services in the Processing Services Department of the Library of Congress. Prior positions at the Library of Congress were in the Labor Relations Office, the Cataloging in Publication Division, the Cataloging Distribution Service, and the Processing Services department office. Before coming to the Library of Congress, he held the position of Assistant Professor of Music (oboe) at Shenandoah Conservatory in Winchester, Virginia. His professional affiliations include membership in the American Library Association, the Music Library Association, the American Society for Information Science, and the International Double Reed Society.

Comparison of Scanning Methodologies for Conversion of Typed, Printed, Handwritten, and Microfilmed Materials

William M. Holmes, Jr.
National Archives and Records Administration

Abstract

The Archival Research and Evaluation Staff of the National Archives has been engaged in a multiyear project to evaluate several methodologies and technologies for retrospective conversion of archival documents and finding aids to machine-readable form. The speaker, who heads the Archives' technology assessment unit, will give a progress report on their findings, explain some of the different conversion methodologies and their suitability for different types of documents, and display samples of original materials and the products of their conversions.

At the halfway point in the 1980's, virtually no finding aids to the 3 1/2 billion documents comprising the National Archives of the United States existed on media other than paper or microfilm. The National Archives and Records Administration (NARA) recognized that computer database and text retrieval software offered considerable promise for creation of machine-readable finding aids that would facilitate reference to archival documents through computer-based search and retrieval.

In 1985, NARA's Archival Research and Evaluation Staff, the Agency's technology assessment unit, launched a project to study and experiment with different methodologies for retrospective conversion of manual finding aids to machine-readable form. The project began with the acquisition of a desktop optical character recognition (OCR) reader, a Dest Corporation Workless Station. This simple reader was programmed in its hardware to recognize about a dozen contemporary typefonts. It was connected to the serial port of an IBM personal computer and operated under control of Crosstalk communications program software. Thus, to the PC, the Dest reader emulated a very fast typist.

The project staff experimented with the conversion of various paper-based finding aids. The procedure was to 1) feed the finding aid document through the reader and capture the equivalent text file in the PC; 2) read the captured text file into WordPerfect, a word-processing program; and 3) use the facilities of WordPerfect to detect and correct OCR scanning errors and to format the machine-readable text into a usable layout. To complete the project, the formatted text was then loaded into Textbank, a software program which provides the capability to perform rapid, flexible free-text search and retrieval using boolean arguments. A turnkey, computer-based finding aid application was thus developed to provide the capability to search record group descriptions for NARA's still pictures collection.

During the course of the development of the finding aid application, a number of observations were made. Only paper-based textual materials created using fairly contemporary electric typewriters could be successfully read and converted by the Dest reader. Considering its programmed font recognition set, this had been expected. When trying to work with low-contrast, electrostatic copies where the type was faded or broken, the speed with which the Dest reader processed the paper and the accuracy rate for its conversion dropped rapidly. For the poorer quality materials, it became a cumbersome and time-consuming process to have a typist find and correct the scanning errors once the text had been loaded into the word-processing software. Records kept of time spent tended to establ sh a rule of thumb that, unless the accuracy rate of the OCR conversion exceeded 90 percent, the time required to detect and correct scanning errors and properly format the document would exceed the time necessary to rekey the whole document from scratch.

The methodologies used by OCR readers and software vary, but they all begin with a capture of a basic raster map of the document. The document is scanned as a matrix of an arbitrary number of horizontal and vertical lines per inch. Each crosspoint in the matrix is considered a picture element or pixel. A raster map captured at 200 lines per inch would consist of 40,000 pixels per square inch or 3,740,000 pixels for an 8.5- by 11-inch sheet of paper. Each pixel may vary from pure white to pure black, depending upon whether or not a part of a typeface is present on the paper at that point. In reality, the pixel will probably be a shade of gray between the two extremes, but the scanner may be designed to intentionally "binarize" the gray scale at an arbitrary threshold to reduce each pixel to a single binary digit (bit) of information representing pure black or white. It is then up to the OCR program to interpret the raster map in order to partition out, isolate, recognize, and identify groups of neighboring pixels which form characters. The program may also attempt to determine the boundaries of character strings and validate them against a dictionary of known words. The Dest OCR reader is typical of the simpler types which use a matrix-matching technique to recognize a specific programmed set of fonts. In pure matrix matching, pixel maps of known typefaces in a particular font and point size are matched against isolated [black] pixel groupings captured from the input raster map. If a close match is detected, the character has been identified. The reader or software may also use symmetry analysis to recognize type point sizes for the font other than the primary point size for which it was programmed.

Other OCR readers or OCR software use character feature analysis or a combination of that and matrix matching. In feature analysis, the OCR recognizes different characters by their makeup. For example, an "A" may have two diagonal lines, one horizontal line, and two line endings. A "B" may have one vertical line, two curved lines, and no line endings. A "C" may have one curved line and two line endings. The most sophisticated programs will also employ algorithmic software to perform a numerical analysis of the raster map.

Commercially available OCR systems generally fall into three classes. First, there are the simple first-generation desktop readers. These usually employ matrix matching to

read a limited set of programmed fonts and point sizes. There are few provisions, if any, for flexibility; graphics on the page will cause problems. At the second level, there are the "omnifont" systems that comprise most of the current generation. These systems use a combination of recognition methodologies that provide the capability to deal with a wide variety of typed and typeset (printed) fonts and point sizes, even when they are comingled in the same document. They may also have the capability to steer around or through any graphics that may be present on the page.

The most sophisticated of the currently commercially available OCR systems are "trainable." These are omnifont systems that permit operator intervention and assistance. If the system cannot identify a particular typeface, the operator is shown the pixel map on a monitor screen and is asked to identify the character or set of characters. For example, if a 10-pitch font was printed at 12-pitch, two adjacent characters may touch or slightly overlap each other. Although the OCR system may not be able to initially make sense of this, once the operator identifies it as a character pair, the system will generally be able to handle any further occurrences. Similarly, it would be possible for the operator to train the system to read any new font that the system could not initially handle.

Commercially available OCR systems may consist of hardware, software, or a combination of the two. The Dest reader is an example of a strictly hardware system. It is a self-contained unit in which the OCR capability is programmed into a microchip (firmware). The information that leaves the port in its backplane is, in effect, a "finished product," a stream of ASCII character codes representing the captured text. Although the NARA project staff connected the Dest reader to a PC, it could equally as well have been interfaced to an electric typewriter, a printer, or any other device with an RS-232 serial communications port. The advantage of hardware systems is that they tend to be fast, but often at the expense of functional flexibility.

Some of the newer OCR products consist of software only. They are designed to be used on a PC and generally rely upon some other manufacturer's raster scanner to deliver a pixel map of the document into the PC's disk storage. The OCR software then analyzes the raster data and outputs character data to another disk file. Purely software products are usually the most flexible because the developer was not tied to a hardware design and manufacturing process. As the product is improved, it is a relatively simple process to release new versions of software as opposed to retooling new hardware. Software OCR systems have become more common today, paralleling the evolution of faster PC's. Five years ago, it was generally necessary to put OCR capability into specialized hardware because to drive the equivalent in software on first generation PC's using 8088-class microchips running at 4.77 MHz was tantamount to "watching paint dry." Today's newer PC's with 80386-class chips and cache memories yielding effective speeds on the order of 25 MHz provide the processor power necessary to run complex OCR software at an acceptable rate of speed. Yet, a small number of vendors of OCR systems choose to use a hybrid approach by providing solutions which are a combination of hardware and software. The hardware portion of these systems usually consists of a microprocessor board designed to be placed in an expansion slot in a PC. The

board will usually have a backplane port designed to be connected to any of a number of commercially available raster scanners. The board hardware is specifically designed to do most of the brute-force preprocessing on the raster data. The data are then passed to the vendor's software, which does the more complex numerical analysis (and operator interaction, if provided), to yield the final product.

Considering the limitations of its Dest OCR reader, the NARA project staff has continued to try its archival materials on the newer generations of OCR systems. The results have not been encouraging. It was found that while vendors had made significant progress in being able to handle mixed format, multifont, typed and printed materials of good quality, little progress had been made in being able to handle poorer quality materials with low contrast, broken, or incomplete typefaces, or more than minor blemishes. For archival materials created on manual typewriters, often with erasures and corrections, it was rare for the OCR accuracy rate to exceed 60 to 70 percent.

The project staff could not find any commercially available OCR system to handle microfilm, although it is not inconceivable that one may exist. There do exist scanners that can produce raster maps from microfilm, so it would only be necessary to couple that capability with OCR software. If these systems are not readily available commercially, it is because there has not been a market niche large enough to warrant their development.

In late 1985, a member of the NARA project staff was speaking at an international conference and described the successes and failures that had been experienced with NARA's textual materials. In particular, she described the difficulties in handling poorer quality materials and the lack of any capability whatsoever to process handwritten materials. In the audience was the sales representative for Optiram, Ltd., a British firm. Optiram claimed to have a "text conversion system" which not only could process NARA's poorer quality typed and printed materials, but handwritten documents as well.

Since validation of Optiram's claims would represent a technological breakthrough of dramatic proportions, between 1985 and 1987, three members of the NARA's Archival Research and Evaluation Staff made separate trips to Optiram's conversion facility in London to verify the company's claims. Carrying a variety of typed, printed, and handwritten archival documents, NARA staff members directly observed Optiram staff as the materials were processed. The results were impressive. The poorer quality typed and printed materials with which the project staff had experienced so little success using commercially available systems were converted by Optiram with accuracy rates in excess of 90 percent. Handwritten materials yielded accuracy rates upward of 70 percent for penmanship with difficult legibility. For documents with 19th century penmanship of decent quality and legibility, the accuracy rate exceeded 95 percent.

The ability to convert cursive handwriting does indeed represent a remarkable breakthrough in the technology. Conventional optical character recognition technol-

ogy, as the term implies, relies upon isolating and recognizing separate characters from the raster map. Isolation of characters is facilitated for typed and printed materials because the characters consist of black pixel patterns wholly surrounded by white pixel regions. As noted before, any imperfections such as touching or intersecting characters or broken or incomplete typefaces can be expected to cause problems. With cursive handwriting, a whole new set of problems arises. There is no distinct dividing point between the end of one character and the beginning of the next. Furthermore, everyone's handwriting is different, and unless a person's penmanship is absolutely perfect, it is improbable that a person would write the same word exactly the same way twice. How then is Optiram able to do what they do? The Optiram text conversion process is performed on a minicomputer. The system uses an ordinary group 3 fax machine as an input device (raster scanner). The conversion process is performed entirely in software. The president and chief scientist of the company is reluctant to talk about his process other than to say that he uses a large complement of mathematical algorithms, developed and refined over an 8-year period. The software performs numerical analysis that works with fragments rather than whole characters. This process obviates the necessity to determine where one character stops and the next begins in cursive handwriting. It also allows for the infinite variations in human penmanship.

Much attention has been given within the last few years to artificial intelligence (AI). The AI programs are said to imitate human intelligence by mimicking inductive reasoning. A relatively new branch of AI, still largely experimental, is neural net software. It gets its name from the fact that it analyzes and processes fragmentary information in much the same way that the neurons and synapses of the human brain process sensory information. When we read a page of text, we certainly do not read it character-by-character. Within the languages and vocabularies with which we are familiar, our brain understands and remembers "pictures" of words and phrases. We can read another person's legible penmanship because, in a instant, our brain assembles the visual fragments from sensory information coming from the retina of the eye and recognizes the similarity of a pattern to one it already knows. If the pattern is not "familiar" (because the writer's penmanship is poor), we may consciously examine the words in context to see if we can interpret what is there.

We can speculate that Optiram's software is a form of highly developed neural net technology. At any rate, Optiram appears to be about 5 years ahead of the rest of the industry. It becomes necessary then to distinguish between the "state of the industry," i.e., what is commercially available for purchase or lease, and the "state of the art" that Optiram offers only as a service. To achieve an acceptable level of accuracy using commercially available OCR systems, input materials generally need to 1) be typed, printed, or zoned noncursive hand printing; 2) have 6-30 point font size; 3) be on paper with decent contrast; 4) have no broken or incomplete characters; and 5) be free of smears or extraneous intersecting marks. The Optiram process, which appears to represent the state of the art, can handle 1) typed, printed, or cursive handwriting; 2) input of marginal quality; and 3) materials with mixed mode, such as a printed form that has been filled in with handwritten entries.

In 1986, NARA's Archival Research and Evaluation Staff decided to expand its original project to evaluate the comparative feasibility and costs of retrospective conversion of different types of paper- and microfilm-based archival documents to machine-readable form by both automated and manual processes. The project staff released three Federal Government Requests for Proposals (RFPs). These RFPs requested bids for text conversion services for 1) typed and printed materials by automated process; 2) handwritten materials by automated process; and 3) typed, printed, and handwritten materials by manual key entry. The first two of NARA's RFPs had technical requirements that the vendor had to be able to achieve a minimum "uncorrected" accuracy rate of 25 percent through use of the automated process; this provision was specifically designed to disqualify a key entry vendor from bidding on the contracts for automated conversion. However, they also contained the provision that if the uncorrected conversion had an accuracy rate of less than 95 percent, the vendor had to employ correction procedures, either manual or automated, to achieve an accuracy rate greater than or equal to 95 percent. Additionally, the RFPs specified that if the paper or microfilm documents were legible to the human eye, the vendor had to convert the text, regardless of whether it could be accomplished by automated conversion or necessitated manual correction.

The RFPs listed a number of different projects (document sets) to be converted, included electrostatic copies of sample documents, and provided data preparation specifications for the desired machine-readable products. They also outlined the random sampling procedure by which NARA would determine the achieved accuracy rates of the vendor's text conversion and data preparation, including format controls, database fielding, and data consistency measures (e.g., standardizing abbreviations or calendar date formats). The RFPs divided the project work into two phases, a test phase and a verification phase. In the test phase, the vendors were required to convert a 10-percent sample from each project's document set, usually amounting to no more than 1/2 million characters. In the verification phase, the vendors would do the "production work" to convert the remainder of each project's documents. The vendors were to bid a firm fixed price for the test phase. This part of the bid was to include the vendors' expenses for performing the conversion, data preparation, any special programming required, and the statistical reporting requirements levied by the RFP. For the verification phase, the vendors were required to submit a firm, fixed price per kilobyte (1,000 characters) of output. The billable output stream would include converted text, embedded spaces, and data preparation sequences (field delimiters and format control characters). Additionally, the vendors were asked to bid separate accuracy level incentives (higher rates) for 95, 96, 97, 98, 99, and 99.9 percent. Finally, they were asked to bid separate rates for paper and microfilm documents because NARA reserved the right to provide the documents in either format. The RFP's statistical reporting requirements for each project included 1) uncorrected accuracy rate, 2) corrected accuracy rate, 3) throughput rate for the completed conversion, and 4) staff-hours to complete each project broken down by subtask.

In October 1987, contracts were competitively awarded to Optiran or both of the automated conversion projects and to Unicor, a commercial enterprise operated by the

U.S. Department of Justice, for the manual key entry projects. One year later, Optiram had completed the test phase of the contract for typed and printed materials, and had begun the verification phase. Optiram had finished the first attempts on all of the projects in the test phase of the contract for handwritten materials and was waiting for NARA to evaluate the output.

As originally envisioned, the test phase procedures and interchange between Optiram and the NARA project staff for each project under contract were to be as follows: 1) NARA would send paper or microfilm copies of documents for each project to Optiram; 2) Optiram would use its automated process to convert the test sample; 3) Optiram would use manual editing to increase the conversion's accuracy rate to an acceptable level if necessary; 4) Optiram would apply the necessary data preparation procedures or programs to the base conversion product to get it to comply with NARA requirements; 5) Optiram would place the final product on 9-track magnetic tape and ship it to NARA; 6) NARA's computer facility would dump (list) all or a portion of the machine-readable file; 7) the project staff would inspect the listing of the machine-readable file for general conformance to NARA's specifications; and 8) the project staff would perform the statistical sampling necessary to validate that the final product met the minimum acceptable accuracy rate.

In actuality, the process involved a considerable amount of trial and error. In a number of instances, Optiram's interpretation of the data formatting and preparation requirements did not coincide with NARA's interpretation. This necessitated NARA's documenting the deviations and returning the tapes and input documents to Optiram for correction of the problems. Sometimes this continued for two or three iterations. When the project staff finally determined the general format to be acceptable, they performed the statistical validation. It was found that once Optiram understood the data formatting and preparation, they never had less than the minimum accuracy rate and were usually above 98 percent.

In their narrative report, Optiram offered the following explanation for the need to rework a number of the projects: 1) there was a learning curve for the Optiram production staff on each of the projects since the data preparation specifications were different for each project; 2) there was often a misunderstanding or misinterpretation of the specifications in the RFP, which were admittedly less than clear in many instances; and 3) there were numerous anomalies in the documents themselves that were not specifically covered in the specifications.

Optiram reported that they achieved an uncorrected scanning rate of 30,000 characters per hour per input scanner, but that when time for manual correction and data preparation was factored in, the actual throughput rate was approximately one-tenth that figure or about 3,000 to 4,000 characters per hour. They also reported that the ratio of total characters shipped compared to total characters processed was on the order of 48.8 percent. In other words, because of the reworking that was necessary on some of the projects, they essentially did the equivalent of processing every input character twice.

Optiram recommended that, in the future, to be fair to both parties, contracts should be restructured into three phases. Phase I would call for development of the technical specification. In Phase I, 1) the Agency (NARA) would develop its specification; 2) the vendor would test and verify the Agency's specification on a small quantity of material, usually less than 10 kilobytes; 3) the vendor would be paid for time and charges to do the work and tune its conversion programs, estimated to be about $500 to $1,000 per project; and 4) the end product would be a "tested" specification, and the Agency and the vendor would be in basic agreement as to how it should appear.

Phase II would be a feasibility test of the conversion and data preparation called for in the specification developed in Phase I. In Phase II, 1) the vendor would test the updated specification against a larger sample of the project materials, usually about 500 kilobytes; 2) the vendor would uncover any anomalies in the data, or glitches or omissions in the specification, thus helping the Agency fine-tune the specification where necessary; 3) the vendor would be paid time and charges for any reprocessing due to changes made in the specification; 4) the end product would be a final, "set-in-concrete" specification; and 5) the vendor would then be obligated to quote a fixed price, with incentive tiers for different accuracy levels for the final phase.

Phase III would be the production phase. In Phase III, 1) large volumes (i.e., entire projects) would be processed by the vendor; 2) no changes would be permitted to the specification without renegotiation; and 3) the vendor would be permitted to invoice for individual projects upon validation and approval of the output by the Agency.

Although NARA's text conversion projects are still under way and the current series of contracts will not conclude until September 1989, a few lessons have already been learned. First, it is extremely important to know your data. Although the NARA project staff did a decent job in preparing conversion specifications to handle the general case for the documents in each project, Optiram found numerous data anomalies and special cases not covered in the specifications. In the case of archival documents or finding aids comprising formatted document sets, some of which were generated before the age of automation, the authors could not have foreseen the need for consistency in data format going from one document to the next. This situation makes it difficult, if not impossible, to prepare a simple conversion specification which can accommodate all cases. It is recommended that a skilled data analyst be asked to prepare the specification, preferably after reviewing a large sample of the input documents. A second data analyst should then be asked to play the role of the vendor, taking the draft specification and trying to apply it to another sample from the input documents. If the second analyst can interpret and apply the specification to the documents without consulting the first analyst, the specification may be fairly complete.

A second recommendation is to know what you want to do with your products. Normally this is not a problem. But in NARA's case, the text conversion project was a research project, the primary purpose of which was to gather empirical data and prepare a report comparing alternative methodologies. It just so happened that the converted machine-readable files were a potentially useful byproduct of the project. Neverthe-

less, although the materials for the projects themselves were selected by management of the custodial units responsible for the documents, those officials could not identify an ultimate disposition for the products. Therefore, the output database specifications used in the RFP were as much of a generic "vanilla" as possible. Now, in retrospect, some of the decisions that were made were not the best.

A final recommendation is to consider doing your final data editing and preparation in-house. Most of the difficulties the project team has had in dealing with Optiram were due to the complexity of the data preparation instructions. In all fairness, Optiram did about as well under the circumstances as any vendor could have been expected to do. In those cases requiring only a simple conversion of text, Optiram turned the products around rapidly and accurately. As might be expected, Optiram's bid prices for projects with complex data preparation requirements were significantly higher than those for projects will little or no special data preparation requirements. If an organization seeking text conversion services has its own data processing shop with skilled programmers, it might want to consider having the text conversion contractor do the basic conversion and then rely upon in-house resources to do the special data preparation and reformatting.

In summation, NARA's Archival Research and Evaluation Staff believes that, except for very few firms, including Optiram, the state of commercially available OCR and text conversion technology has not advanced much in the past 3 years. Hardware and software available for purchase or lease still cannot accurately handle other than high-quality, typed or printed materials. Significant competition for Optiram's level of capability appears not to have advanced beyond rudimentary, prototypical laboratory-level developments.

WILLIAM M. HOLMES, JR.

Director, Archival Research and
Evaluation Staff, National Archives
and Records Administration

As Director of the Archival Research and Evaluation Staff, Mr. Holmes has responsibility for the planning and execution of a technology assessment and research program and for assisting other executive managers at the Archives in the planning and implementation of technological applications to their programs. In addition to performing general technological assessments and acting as an in-house consultant to other parts of the Archives organization, the Archival Research and Evaluation Staff undertakes a number of research pilot activities to test various new and emerging technologies under live operational conditions. Typical of these activities are current projects involving the use of digital imaging, optical character recognition, bar coding, and media preservation techniques.

Before joining the National Archives in November 1983, Mr. Holmes served for 3 years as the Director of the Test Center Planning Staff at the Social Security Administration Headquarters in Baltimore, Maryland, the mission of which was to define, design, and procure a state-of-the-art computer software development facility to support SSA's programmer/analyst community of 1,000. Prior to his service at SSA, Mr. Holmes was the Manager of Software Validation and Testing Services for the General Services Administration's Office of Information Resources Management and was responsible for validating vendor-supplied software before its procurement by other Federal agencies. From 1973 to 1977, Mr. Holmes was a staff analyst at the Navy's Automatic Data Processing Equipment Selection Office, prior to which he was a systems programmer and a programming instructor at Naval Command Systems Support Activity at the Washington Navy Yard.

SESSION 2

GEORGE R. THOMA,
MODERATOR

Chief, Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine

George R. Thoma received the B.S. in 1965 from Swarthmore College, and the M.S. and Ph.D. in 1967 and 1971, respectively, from the University of Pennsylvania, all in electrical engineering.

As the senior electronics engineer and Chief of the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, the research and development arm of the National Library of Medicine, he has developed and evaluated systems involving image processing, document image storage on digital optical disks, high-speed image transmission, analog videodisks, satellite communications, and video teleconferencing.

Dr. Thoma's previous experience at the General Electric Company, AII Systems, and the University of Pennsylvania had been in developing concepts and systems involving the application of satellites in voice and data communications, video distribution, navigation, and surveillance. He has lectured extensively on these topics at various conferences and institutions in this country as well as in Japan, Holland, Switzerland, Canada, Mexico, and India, frequently as an invited speaker. He has also published widely in the technical literature. He serves as a reviewer for the "Information Processing and Management Journal," on the editorial review board of the "Journal of Clinical Engineering," and as a consultant for the National Science Foundation.

Dr. Thoma is a member of the Society of Photo-Optical Instrumentation Engineers and the American Society for Information Science. He is listed in "American Men and Women of Science," "Who's Who in Technology Today," and similar publications.

Digital Imaging at the Library of Congress

Audrey Fischer
Library of Congress

Abstract

After several years of planning and procurement, the digital imaging component of the Optical Disk Pilot Program at the Library of Congress became operational in the fall of 1984. As the first pilot of its kind at a time when a technology, now still in its youth, was literally in its infancy, the Library of Congress has had a hard row to hoe in paving the way for those who have followed. As with all libraries, the problems facing the Nation's library--space, service, access, and preservation--are large scale. Therefore, the questions posed by the pilot have been equally large scale. Furthermore, the solutions to such problems will have to be implemented on a large scale as well. In attempting to determine if optical disk technology can and should be implemented librarywide, the original questions have been answered while still others are being posed during the final phase of the pilot. The history and accomplishments of the digital imaging project are discussed with speculation on where it is headed as a result of what has been learned.

I am pleased to be here today to discuss the digital imaging system at the Library of Congress. I am particularly pleased about the focus of this conference, that is, application of scanning methodologies in libraries. It was not too long ago that digital imaging was a technology in search of an application and was in danger of not finding a home in the library community with our special concern for preservation.

Nearly 7 years ago, our recently retired Deputy Director, Bill Welsh, a man of vision, was a lone voice when he wondered out loud whether or not this technology could have a favorable impact upon the Library of Congress. By establishing our Optical Disk Pilot Program in 1982, he gave voice to his objective: to evaluate the use of optical disk technology for information preservation and management, to determine if such a system could be put to wider use in the future. Implicit in this objective was the hope that other libraries would follow in our footsteps and benefit from our experience. Judging from those of you in attendance, and from the many interesting applications being discussed during these 2 days, I would say that this objective has been met.

As the Nation's library, the problems facing all libraries, such as SPACE, SERVICE, ACCESS, and PRESERVATION, are large scale. This is the key to understanding why Mr. Welsh was so eager for the Library of Congress to explore a technological solution to these problems. As the first of its kind at a time when a technology, now still in its youth, was literally in its infancy, the Library of Congress Optical Disk Pilot Program has had a hard row to hoe. From the outset, we not only posed large-scale questions, but set goals which, in hindsight, appear more than a bit ambitious. On the other hand,

much of what we have accomplished may have seemed impossible to our colleagues, if not to some of us, when we let the contract to Integrated Automation (now Litton Industrial Automation) back in 1982. Yet somehow an RFP was developed which outlined our need for something that barely existed even as a prototype. With that qualification, I would like to begin the task of discussing the digital imaging component of the pilot program with an emphasis on the way we believe optical storage can address the problems of space, service, access and preservation. I will also attempt to explain where the Library is heading with the technology as we enter a postpilot transition period with plans for establishing a production system.

Let me begin by distinguishing the print pilot project utilizing digital imaging technology, that is 12-inch WORM disks (Write Once Read Many Times), for textual material as but one component of the pilot program, it is the component I will be discussing today. The nonprint pilot project utilized analog videodisk technology to store and retrieve still and moving pictures, often in color, as well as compact audio discs to capture sound recordings. The Library has also made use of CD-ROM technology, producing a Read Only disk for MARC records. While all three technologies are related, the methods for producing the disks differ markedly, as do their applications.

It was clear from the outset that text-based material, largely black and white, would be best suited to digital imaging technology, which showed great promise for image enhancement and resolution. At a resolution of 300 lines per inch, the amount of text that can be written onto one side of our disks now is 4,000-9,000 pages with an average of 14,000 on BOTH sides, depending upon the amount of data contained on each page. Estimates by the vendor of 10,000-15,000 images PER SIDE turned out to be ambitious for pages containing more than one sentence of data, not to mention those that contain pictures which we are able to capture in halftones. Halftone pictures, of course, utilize additional storage space. When the disks were upgraded from plastic to glass several years ago, and a compression problem was resolved as well, we began to experience our current level of storage--which is up nearly 100 percent over the previous level. Double-sided disks became available only recently, increasing our storage capacity per platter yet another 100 percent. Data were transferred from the single- to double-sided disks this past summer, thus reducing the number of disks in the jukebox by half, or from about 80 to 40. The original jukebox was upgraded at the time we switched over from plastic to glass disks and both the prototype and our current model hold 100 double-sided platters. The new jukebox differs from the old one in that it uses a pneumatic, or suction, method rather than a mechanical servo system to retrieve the disks and can accommodate two disk drives or players for a faster response time. It also can access double-sided disks. We are expecting delivery of a third jukebox shortly. In short, though original expectations for storage were not met, we have nearly quadrupled the storage capacity since our original equipment was delivered over 4 years ago.

We began actual scanning 4 years ago this month. Our original configuration consisted of one microfiche scanner, one multi-purpose scanner, and one quality review terminal. The microfiche scanner is capable of scanning from 24x standard 98-frame microfiche

to disk. This is truly a first-generation model, and we found it to be too labor-intensive to achieve an acceptable image. That is, much manual adjustment is needed, and many images are rejected before an optimum image is captured, though it should be noted that more proficient models are now available. Instead, we opted to concentrate on paper scanning. The multi-purpose scanner for paper scanning is capable of scanning material from 5- by 7-inch to 11- by 14-inch, though the latter would be reduced to 8.5- by 10-inch for display. It was intended to be fed either manually or automatically by using a sheet feeder. Since the sheet feeder attachment did not work optimally, we have relied upon this scanner for documents that we do not want to put through a feeder using much the same criteria that one would employ in selecting a rotary or planetary camera for microfilming. Fortunately, we added a high-speed, or semi-automatic, scanner, which captures the front and back of the page at one pass through and displays both on the two corresponding display or preview terminals. A second quality review terminal was added when it was determined that the magnetic disk buffers were filling faster than one person could perform the 100 percent quality review that we insisted upon throughout the pilot. We have also been operating only one daytime shift, from 8:30 - 4:30. It became clear early on that utilizing only two scanners during one shift per day would not be suitable for a high-volume, or production, environment. Also, our system architecture involves a video system controller (VSC) that causes a hesitation at the scanning level while the images are being compressed. This hesitation accounts for a slower input rate than we originally expected. Output (hard copies of documents written to disk) is handled by the batch, or remote off-line printer, which is a modified Xerox 5700 laser printer. In addition, two convenience printers--modified Xerox 2700 machines--are attached to two of the four user display terminals. During the pilot, these terminals were located throughout the various reading rooms in all three buildings comprising the Library of Congress. Plans are under way to supplement this equipment with Canon laser printers.

I would like to say a few words about the staffing necessary to operate during the pilot phase. A matrix managment approach was used. That is, the talents and skills of over 100 staff members from throughout the Library of Congress were tapped in order to effectively plan and operate during the pilot. This was in addition to performing their regular duties. For example, we relied heavily upon our reference staff to educate us about the needs of our users, and as such they had input into devising retrieval methodology. Similarly, staff from our Automated Systems Office programmed both our input system (known as ISOD) and our retrieval system (known as ODIS). During the past 4 years, the tasks necessary for input scanning and quality review were added to the job descriptions of staff in the Preservation Microfilming Office. Seven people are assigned to an optical disk team for a particular rotation (usually 2 months), during which time they are responsible for document preparation, scanning, and quality review. As time permits, they can return to their preservation microfilming duties. This approach was successful throughout the pilot insofar as there is a close correlation between the tasks necessary for both operations and each requires a general understanding of, and commitment to, preservation of library materials. However, toward the end of the pilot phase of the project, it became clear that there would be a marked increase in preservation microfilming activities and that the staff could no longer carry

this double burden. For this reason, we are in the process of staffing a dedicated optical disk operation unit to be located in our Automated Systems Office.

Though I have been candid about areas that have fallen short of our expectations, it is remarkable how much of what was planned, virtually in a vacuum back in 1982 in our original design document, was implemented almost to the letter. The workflow is one such area and I will now attempt to step you through the process. Scanning into the print system is made possible by an LC programmed Input Subsystem for Optical Disk (ISOD) which is maintained on a Data General minicomputer system. ISOD controls all input scanning, quality review functions, disk availability directories, image transfer to the optical disk, and transfer of document locations to the retrieval system.

In the document preparation and input process, we develop records or make use of existing machine records, produce an Input Header Sheet or target, collate the documents, complete the necessary paperwork, and finally send the document to the Optical Disk Operations Center (ODOC) for scanning. The operator then scans each document by stepping through the Input Subsystem Software which is menu driven. A decision must be made regarding the optimal scan setting, using a combination of the light, medium, dark and halftone settings in the scanner control, for a total of six options. At this point in the process, unacceptable images can be rejected, as they do not yet reside on the optical disk but rather on the magnetic disk buffer (each scanner is equipped with a separate magnetic disk buffer). In a sense, the preview terminals attached to each scanner permit a first level of quality review. However, another 100 percent quality review is later performed on images stored on the buffer. Our earliest system configuration necessitated complete rescanning of an entire document when only one error occurred. Our current configuration compares more favorably with microfilming, with which cutting and splicing must be done to correct errors detected after the film is developed. With the disk system, we have the ability to juxtapose images which are out of sequence. Extra images can be deleted and missing pages can be sent for scanning and insertion into their proper position within the document. Poor-quality images can be overwritten, that is rescanned, perhaps using a different setting, and overlayed in place. When an entire document is deemed acceptable, it is written to the optical disk contained in the disk drive in the video system controller. When an entire disk is filled, it is placed in the jukebox and is available for retrieval.

The actual transfer to disk results in the creation of a page table or locator file which is controlled by the retrieval system software known as the Optical Disk Interface System or ODIS. The file is passed from the input subsystem (ISOD) to the retrieval system software (ODIS) via a tape load. This locator file then links user requests to the optical disk address of scanned documents. The retrieval system is the interface between the digital imaging system and the Library's mainframe. It allows users to search, identify, and select documents online, using SCORPIO commands with which many of our regular patrons are already familiar. The ability to view full-text of selected items on the same terminal following a search of the database is a vast improvement upon waiting for a deck attendant to retrieve a hard copy from the stacks. On the optical disk terminals we use the tongue-in-cheek GONE TO THE STACKS message to briefly ex-

plain the retrieval process to the user. While there is a delay factor, it is certainly shorter than one which would occur if a member of the deck staff had to physically go to the stacks. The retrieval software also monitors user reaction via an on-line questionnaire.

Another area of the pilot program which was implemented with great fidelity to the original design document was the preparation of materials for scanning. While not all the collections slated for scanning were included in the pilot, specific preparation methods for each, including bibliographic control, varied little from the original plan. During Phase I of the pilot, we scanned previously machine-indexed material from a pre-existing file of public policy articles and documents developed and indexed by the Library's Congressional Research Service. This is commonly known as the Bibliographic or BIBL file. The Optical Disk BIBL file now contains the full text of nearly 14,000 items cited in this file from 1983 to the present. I will discuss this file in more detail later.

The challenge of developing new indexing systems was tackled in Phase II of the pilot, in which we created three additional optical disk files under SCORPIO. The first, the Optical Disk Congressional Record File, contains the full text of all issues of the Congressional Record for the 99th Congress. We have successfully provided retrieval for this file at the page level through a commercial index. The second, the Optical Disk Manuscripts File, is a collection of Presidential and Vice Presidential portraits and letters dated 1750-1925; the file uses custom software designed by staff in our Automated Systems Office. The third and final file in the program is the Optical Disk Serials File, which includes fully scanned issues of a selection of 79 journals from 1983-88. Retrieval is at the issue and article level through a commercial index (MAGAZINE INDEX) and some experimentation was done with in-house software. Approximately 1,200 issues were scanned. The entire print database now totals over 16,000 items, or almost 1/2 million images.

These electronic files now demonstrate that full-text documents, which are housed in the original throughout the Library's three buildings, can be searched, delivered and printed from the same user terminal. Also, as many people as there are user stations can access the same document at the same time. As you can imagine, this is of concern to the copyright proprietary community. From the beginning of the pilot, the publishers were involved in the Library's plans to test this new technology. A Copyright Advisory Committee consisting of librarians and publishers was convened, and the Library agreed to provide feedback regarding usage of the files. Copyright permission to scan, display, and print was obtained from the publishers of the 79 journals throughout the duration of the pilot. We have not charged a fee for off-line or batch printing throughout the pilot in order to encourage usage. The copyright issue will continue to receive attention during the present transition phase.

We are excited by the ways in which optical disk technology can address the critical problems of SPACE, SERVICE, PRESERVATION, and ACCESS.

Space is an acute problem for us at the Library of Congress. Our collections now number more than 80 million items stored on 863 kilometers of shelving, and they continue to grow at a rate of about 7,000 items per day. Optical storage holds great promise for resolving this problem, since we have been able to store large amounts of data at high densities. To better illustrate the degree of compaction which can be achieved on an optical digital disk, consider that one 21.8- by 25.6-cm book page occupies 5,400 square millimeters. The same page on 35-mm microfilm occupies 150 square millimeters, and on 98-frame microfiche, 70 square millimeters. By comparison, the same page occupies only 3 to 6 square millimeters on a 30.7-cm optical disk.

Of course, we are always keenly interested in improving library service and access to library materials. As you can well imagine, servicing and making available a collection of this size is a monumental job. After consulting manual and on-line bibliographic sources, patrons generally must wait for staff to deliver material to them. This can be a time-consuming activity perhaps resulting in the designation of NOS or NOT ON THE SHELF in the event the item is either missing or being used by another patron. But if we have written the item to disk, this will be indicated in the on-line bibliographic citation. If the patron is seated at an optical disk terminal, the item can be retrieved, displayed, and printed. Although the actual collections are currently stored and made available in 20 reading rooms throughout the three buildings, this technology holds the promise of conducting multicollection research from one remote location. The preservation challenge is perhaps the greatest one facing the Library of Congress today. Many of our most brittle collections are decaying at an alarming rate. Reducing the handling of such collections by electronic document delivery may have major implications for preservation. Many optical disks are currently projected to last 10 to 20 years, and some vendors are even claiming 100 years. In most circles this would not be considered "archival," but the quality of the data can be electronically monitored and transferred to new disks as necessary. The cost of disk duplication will undoubtedly decrease as the technology finds a permanent place in the information industry. Such periodic transfer of data will be analogous to records retention and disposition schedules under a sound records management program. The Library of Congress plans to work closely with the National Institute of Standards and Technology and other standards groups so that we can share the results of tests aimed at approximating disk longevity.

While we believe that this technology compares favorably with microfilm as a preservation tool, not unlike microfilming, the process of transferring documents to a secondary storage medium is enormously labor intensive. Tackling a large quantity of material will certainly require a system that would be configured on a larger scale than the preproduction system with which we have been operating. All things considered, we have been most encouraged by the technology's advantages with respect to service and access of current and/or high-use materials. The Congressional Research Service has been especially pleased with what was accomplished in Phase I of the pilot. In that phase, images of their Bibliographic File were scanned and made available in the Library's three Capitol Hill buildings. As a result, CRS has decided to further explore the possibility of applying optical disk technology to its SDI (Selective Dissemination of Information Document Delivery System). The SDI is a service provided to Mem-

bers of Congress, their staffs, and CRS researchers. Subscribers specify particular topics of interest in the areas of public policy. They receive a weekly set of cards listing the most recent material published on the topics they have preselected. They may then request items in hard copy, which CRS provides from its complete collection of microfiche. Each month, approximately 30,000 pages of material are put onto microfiche, and about 150,000 pages of hard copy are made at the request of Congressional offices. Blowbacks, or paper copies from microfiche, are produced on three Xerox 970s. As these machines were acquired over 10 years ago, they have become a problem to maintain. They will not meet the high volume needed by CRS much longer.

Exactly 1 year ago this month, a CRS test plan was implemented which involved scanning current material in tandem with the ongoing microfilming effort. It was soon discovered that we could easily meet the daily input scanning requirements. The daily output was more problematic and plans are underway to upgrade our off-line printing equipment, with the financial support and approval of the Congress.

We are all excited by the prospect of perfecting this particular application in the near future. Some of my colleagues envision a time in the not-too-distant future when these documents will be beamed directly to Congressional offices.

Though much needs to be done, we have come a long way in our quest to find a home for optical disk technology in the Library of Congress.

AUDREY FISCHER

Optical Disk Operations Ccordinator,
Library of Congress

Audrey Fischer is the Operations Manager of the Optical Disk Digital Imaging Project at the Library of Congress, with responsibility for daily oversight of input-scanning activities. She is the former Head of the Preservation Section (an in-house microfilming operation) of the U.S. Copyright Office. Ms. Fischer earned a B.A. in History and English at the State University of New York at Binghamton, and an M.A. in American Studies and Archival Management from Case Western Reserve University where she was awarded the Fenn Archival Fellowship. Her outside activities include writing and consulting in the areas of micrographics and optical disk systems design.

Issues In Document Conversion

Frank L. Walker
Lister Hill National Center for Biomedical Communications,
National Library of Medicine

Abstract

Although capturing the contents of printed material in machine-readable form is necessary for the creation of products such as searchable full-text databases, the first step in the process is the capture of the material in bit-mapped form. This is not always easily done with commercially available OCR systems employing specific input devices, usually handling only loose-leaf pages and only document pages that are strong enough to be handled by a mechanical transport or feed mechanism. Problems arise in the handling of bound volumes and brittle paper. As part of a research program into the electronic conversion of biomedical documents for preservation, the Lister Hill National Center for Biomedical Communications, the R&D arm of the National Library of Medicine, is performing electronic imaging research spanning the gamut from scanning, digitizing, image processing, storage and archiving to retrieval, transmission, display, and image manipulation, with special attention to fragile and bound documents. This paper addresses two key issues in scanning such documents. The first issue is converting the paper-based collection to electronic form as efficiently as possible. Conversion efficiency is governed by such factors as image compression, method of conversion, and the system architecture that performs the conversion. The second issue is that of image enhancement. Electronic imaging offers a distinct advantage in that images of deteriorating documents exhibiting poor contrast or stains and other disfigurements inhibiting legibility may be enhanced when they are scanned. Findings related to these issues will be discussed.

Background

Research into alternate means of document storage and retrieval at the Lister Hill Center is motivated by the National Library of Medicine's congressional mandate to preserve and provide access to the biomedical literature. First, the Library must preserve the information in paper materials before it is irretrievably lost to decay. As of 1986, approximately 9 percent of the collection was found to be brittle, thus needing immediate preservation in an archival form [1]. More recently this figure has been found to be closer to 12 percent. This conversion is presently proceeding through microfilming of about 35 to 40 million pages. At the same time, easy and rapid access to the archived documents is considered a desirable goal. Electronic storage of images of these documents on optical media, with fast computerized methods of retrieval, has the potential to simultaneously solve the archival storage and access problems. The Lister Hill Center is performing research into all aspects of electronic document preservation, access, and imaging. Its specific interest is in the electronic preservation of bound volumes and fragile documents which characterize the part of the collection that

is deteriorating. As part of the Lister Hill Center's research, it has developed bookscanners capable of scanning bound volumes and fragile documents, and prototype systems for evaluating problems encountered in document conversion and access.

This paper discusses two key issues of current interest: conversion efficiency and image enhancement. Document conversion efficiency is critical because conversion can be an expensive and time-consuming process. Factors which can make document conversion an efficient task are discussed. Image enhancement is critical in document conversion because by improving image quality before optical character recognition (OCR), the time to perform accurate OCR can be minimized, thereby increasing throughput. Techniques for image enhancement will be outlined.

Conversion Efficiency

The efficiency with which paper documents can be converted to electronic form can be measured by cost and throughput. The cost of archiving the electronic documents can be minimized by storing as many document images as possible on the archival medium. This is accomplished through image compression. Throughput also determines conversion efficiency, since the higher the throughput of document conversion, the lower the cost. The reason for this is that document conversion is an operator-intensive task. By minimizing the amount of time spent by operators, the cost of human labor is minimized. This time can be lessened by the careful choice of system architecture for document conversion. Conversion methods and system architectures for performing the conversion will be discussed.

Image Compression

Any method for converting paper documents to electronic form will be beneficial if the resulting electronic representation of the document is compressed. Compression is the systematic removal of redundancy from an electronic signal for efficient storage and transmission. If an 8.5- by 11-inch page is scanned at 200 points per inch resolution, the result is a bit-mapped image of almost 4 million bits, or 1/2 megabyte. Clearly, if these data can be reduced or compressed, it will be possible to store more pages on an electronic archival medium such as a CD-ROM (compact disk read only memory) or WORM (write once read many) disk. In addition, since there will be less data to retrieve per page, the retrieval speed will be faster than that for an uncompressed page. Table 1 lists several methods of compressing electronic documents.

Table 1: Compression techniques

Technique	Typical Compression Ratio
1. Text compression using Optical Character Recognition (OCR)	100.0
2. Bit-mapped image compression	
-CCITT Group 3 one dimensional compression	8.9*
-CCITT Group 4 two dimensional compression	14.4*
-Progressive encoding for gray scale or color image compression	[not available]

*Actual data for biomedical documents in the NLM Collection

By far the best currently available technique for compressing text documents is optical character recognition (OCR), also sometimes called omnifont character recognition. Several techniques are commercially available for performing OCR on the bit-mapped image representation of a document page. The output of the OCR process is the text portion of the document in the form of ASCII codes. A typical compression ratio achievable through OCR is around 100. For instance, for a page containing 5,000 characters, the resulting size of the compressed page is 5,000 bytes, or about 1/100th the size of the corresponding uncompressed bit-mapped image. However, OCR cannot be used to compress graphics appearing on the same page as the text, a common occurrence in the scientific journal literature. For those pages which contain graphics such as photographs or drawings, several techniques for bit-mapped image compression are available. Experiments have shown that the CCITT Group 3 one-dimensional compression algorithm produces an average compression ratio for biomedical documents of about 8.9 [2]. The CCITT Group 4 two-dimensional compression algorithm is somewhat better, giving an average compression ratio of about 14.4 for a typical biomedical document. A CCITT subcommittee is considering a compression method, progressive encoding, that will be able to compress color images or black and white images with a gray scale [3].

Conversion Techniques

There are four basic methods of converting paper documents to electronic form. Each of the four methods uses one or a combination of the techniques of image compression listed in table 1. Table 2 lists the 4 methods of conversion.

Table 2: Methods of document conversion

Method
CASE 1: Convert document to bit-mapped images
CASE 2: Convert document text to ASCII codes
CASE 3: Convert document text to ASCII and convert compound pages to bit-mapped images
CASE 4: Segment compound document pages into text and graphics regions; convert text regions to ASCII and store each graphics region as a bit-mapped image

The choice of which method to use depends partly on how the documents are to be retrieved once archived and in what form they should be displayed. If the user wants the retrieved electronic document to resemble closely the original paper document, cases 1 and 4 are best suited for performing this conversion. If the user wants to be able to see only the text from the paper document, case 2 is ideal. If the user wants to be able to perform a full-text search, cases 2, 3, or 4 would be suitable. The four cases will be considered individually, the advantages and disadvantages being given for each case.

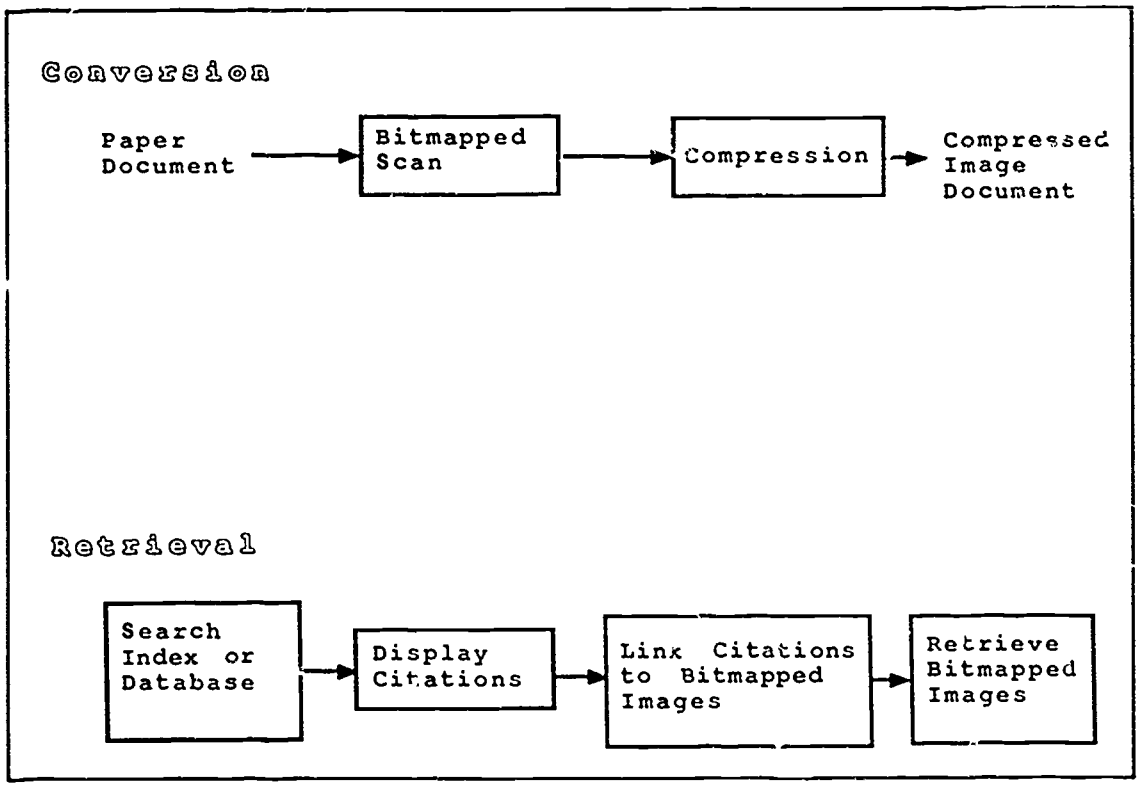


Figure 1. Conversion/Retrieval of Images: Case 1.

The first case of document conversion, illustrated in figure 1, shows that the document pages are converted to bit-mapped images through a bit-mapped scanning process. Such a process is normally done by a charge-coupled device (CCD) camera, which produces an image of thousands of bits, each representing (or mapped to) a small part of the original document. Before archived bit-mapped images may be retrieved, a method of accessing them must be devised. One such method is to create a database of document citations, with each citation giving such information as title, author, publisher, etc. Then, as shown in figure 1, the user may search the citation database, perhaps using keywords, and display each pertinent citation. For each citation in which the user is interested, the citation may be linked to the corresponding document images, and those images may be retrieved. The advantage of this method of document conversion is that the retrieved images closely resemble the original document. One disadvantage is that it may be time consuming and costly to create the citation database unless, as in biomedicine, it already exists as Medline® or one of a family of related

databases. A second disadvantage is that the size of the image database is the largest of all four methods of conversion, since solely bit-mapped compression is used, rather than OCR.

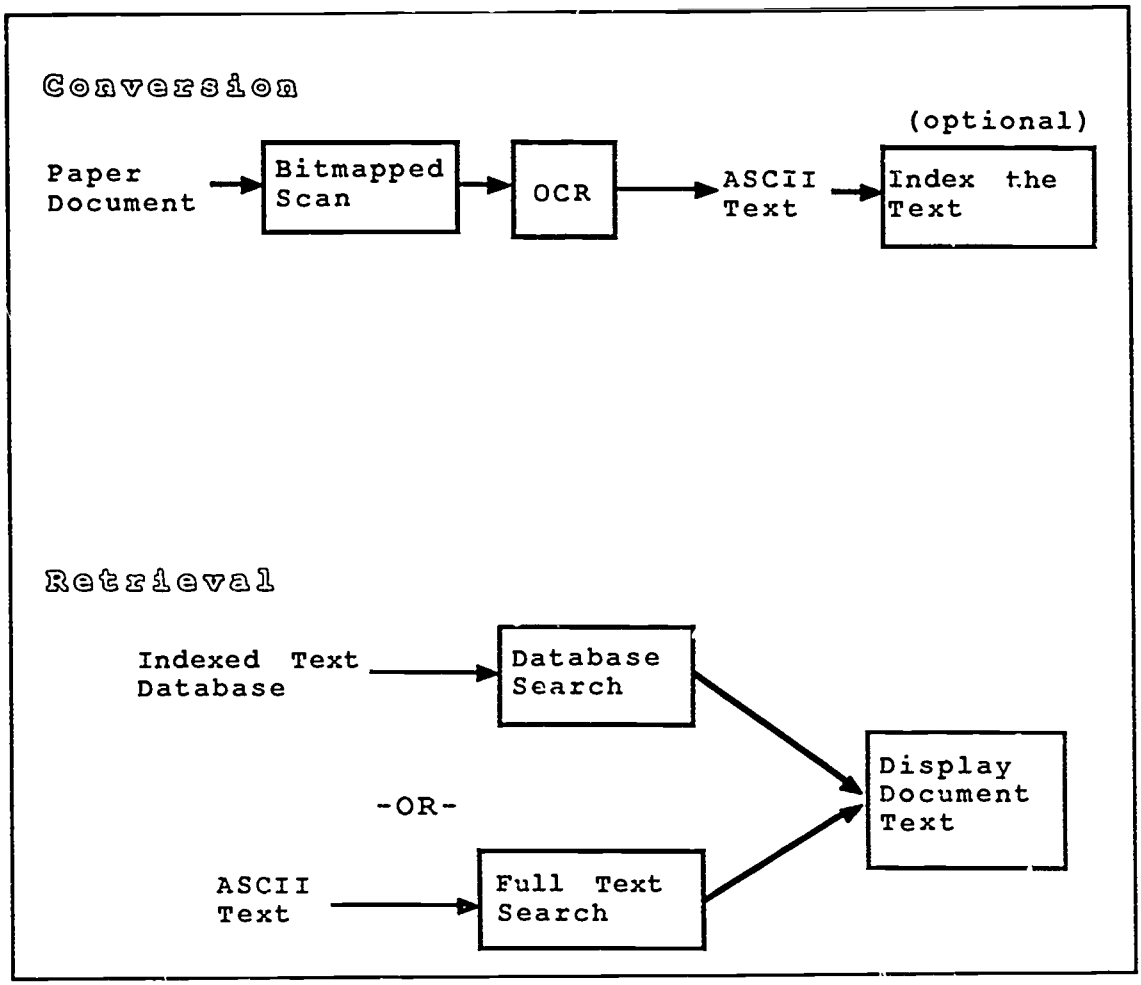


Figure 2. Conversion/Retrieval of Images: Case 2.

The second method of document conversion is shown in figure 2. Here the document text is converted to ASCII codes using an OCR process. The document is initially scanned, producing a bit-mapped image which is then sent through the OCR process, the output of which is ASCII text. As an optional step, the text may be indexed automatically, producing an indexed text database. To retrieve the documents, two methods are possible. First, a search may be performed on the indexed text database, and the documents subsequently retrieved. Second, a full-text search may be done on the ASCII text itself, yielding the documents of interest. It is quite possible to combine both methods of retrieval to form a powerful retrieval engine. An advantage of this method of conversion is very high compression resulting from the OCR process, allowing a large number of documents to be stored. A second advantage is that the indexed text database search is fast. On the other hand, a disadvantage is that the full-text database search can be slow. Another disadvantage is that all graphics from the original paper document are lost and are not displayable using this technique. Finally, the

original font and type size information is usually not retained during most commercially available OCR processes, so that in addition to eliminating the display of graphics, the displayed text itself will not be similar to that in the original paper document.

Briefly considering OCR, current technology permits thousands of fonts and many type sizes to be automatically recognized with fairly good accuracy. However, recognition accuracy is critical with OCR since the less accurate the process, the more time an operator must spend correcting the errors found during OCR. The effect of recognition accuracy may be appreciated by considering a typical 5,000-character printed page. If the OCR process is 99 percent accurate, the product will contain about 50 errors which must be corrected after the OCR is completed. If the accuracy drops to 97 percent, this translates to 150 errors per page which must be manually corrected. Once the accuracy drops much below this, it might be more efficient to type the document if full text is desirable. Several factors affect OCR accuracy. These include the OCR algorithm itself; some commercial algorithms are better than others. Second, the font size is critical since OCR has a more difficult time handling very small type. If the type is too small, resulting in a large number of errors, the scan resolution must be increased. Third, image quality plays an important role in OCR accuracy. Optical character recognition systems tend to prefer documents which have clean, crisp, black text on white background. Any deviation from this results in lower OCR accuracy. Documents which are stained, yellowed, or which have poor contrast are not suitable for OCR unless the image quality can be improved. Image enhancement will be addressed later in this paper.

Figure 3 illustrates the third case of document conversion. This method combines the first two methods and provides the best features of both. As in the first two cases, the paper document is scanned, initially producing a bit-mapped image. If any resulting image contains any kind of graphics, each is stored as a bit-mapped image. All images containing text are sent through an OCR process, producing ASCII codes for the text. As in case 2, an additional step might be to index the text. Retrieving the electronic documents is similar to case 2, in which the indexed text database may be searched or the full text may be searched. Either way, the document's text may be retrieved and displayed. For any image stored in bit-mapped form, that image may be linked to the corresponding page of the ASCII text and also be retrieved and displayed. The advantages of this method of conversion are that all graphics are retained from the original document and that the indexed text database search technique is very efficient. Furthermore, this technique could result in better image compression than case 1, resulting in lower cost of electronic document storage. This assumes that the proportion of compound pages (mixed text and graphics) is sufficiently low. The disadvantage is, as in case 2, that the full text search retrieval technique can be quite time consuming.

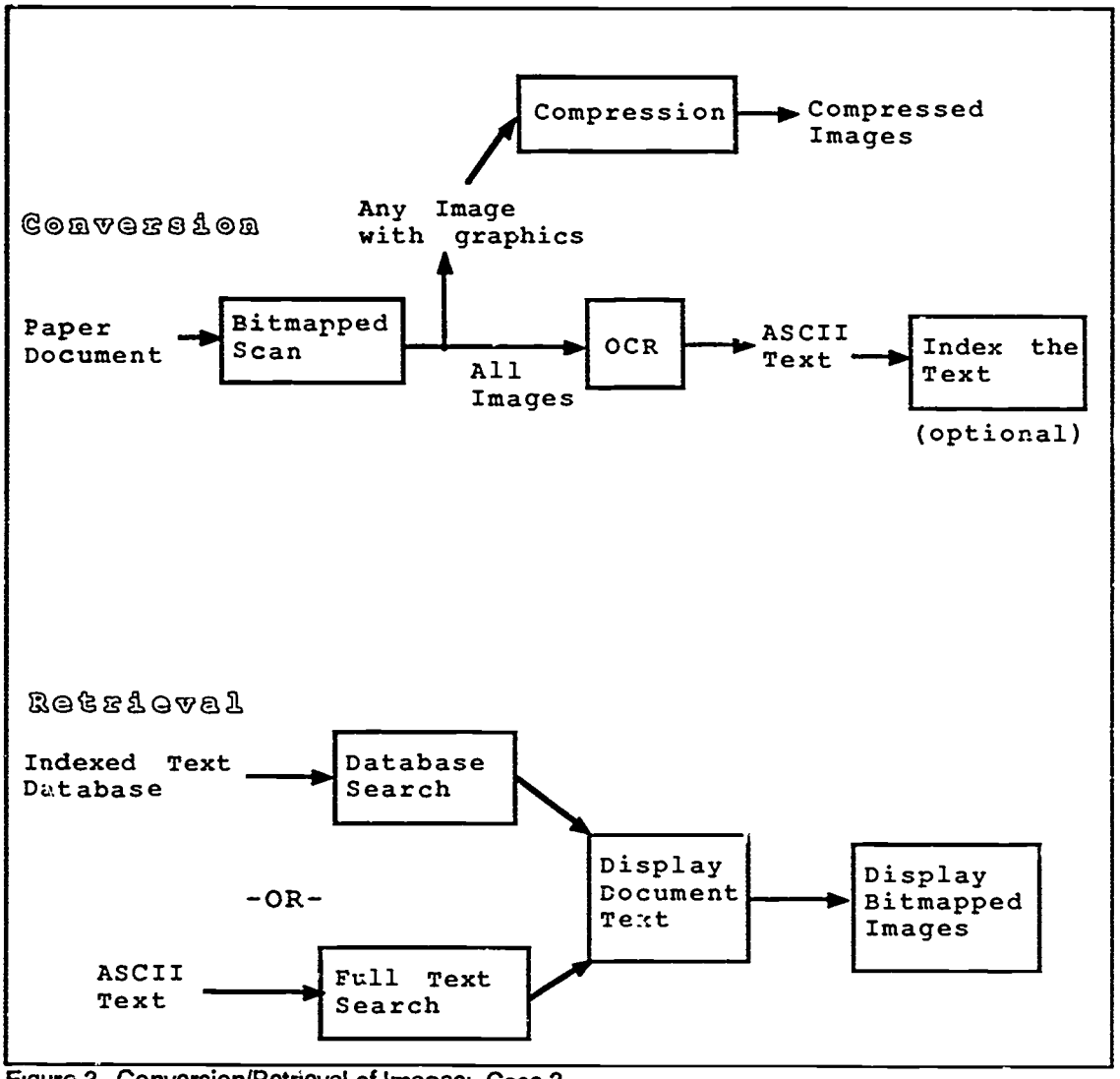


Figure 3. Conversion/Retrieval of Images: Case 3.

The fourth method of converting documents to electronic form is illustrated in figure 4. As in the first three cases, the document is scanned, producing a bit-mapped image. Then the image is segmented into its component text and graphics subregions. Each text region of the image is sent through an OCR process, producing the ASCII text, and optionally is indexed. Each graphics region is stored separately as a bit-mapped image. Document retrieval is the same as for the previous two cases: either search an indexed text database or search the full ASCII text to retrieve the document. Each image displayed is a composite image composed of its member regions.

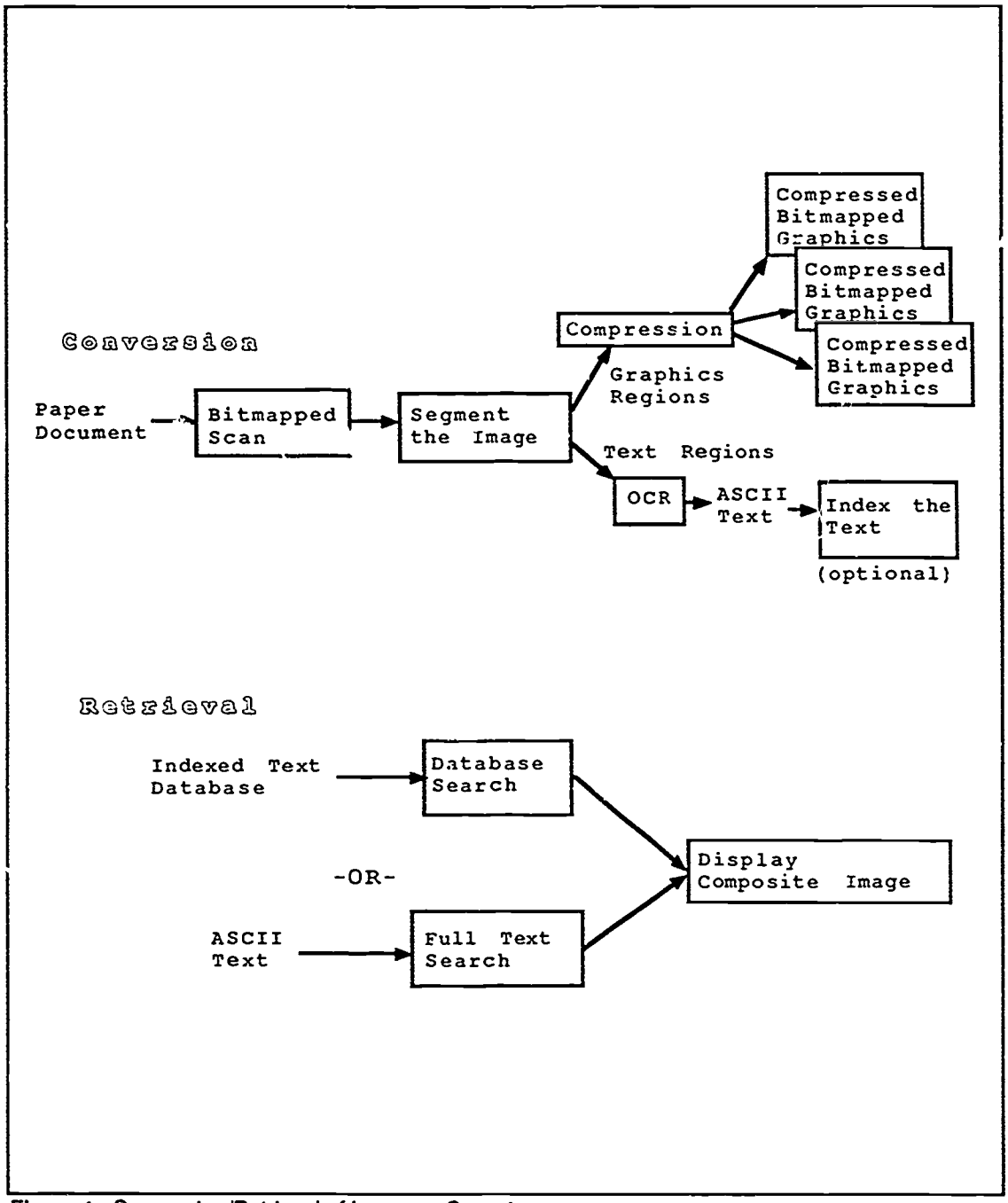


Figure 4. Conversion/Retrieval of Images: Case 4.

Figure 5 is an example showing how the composite image may be formed for a text region and two graphics regions. To form the composite image, the location of each region must be retained from the segmentation process. The recombination process involves techniques similar to those used by electronic publishing packages commonly available for personal computers. One advantage of this method of conversion is that it results in the best possible compression ratio (ASCII for the text and bit-mapped compression for the graphics regions). In addition, if it is possible for the font and type size information to be retained from the OCR process, the composite image can be

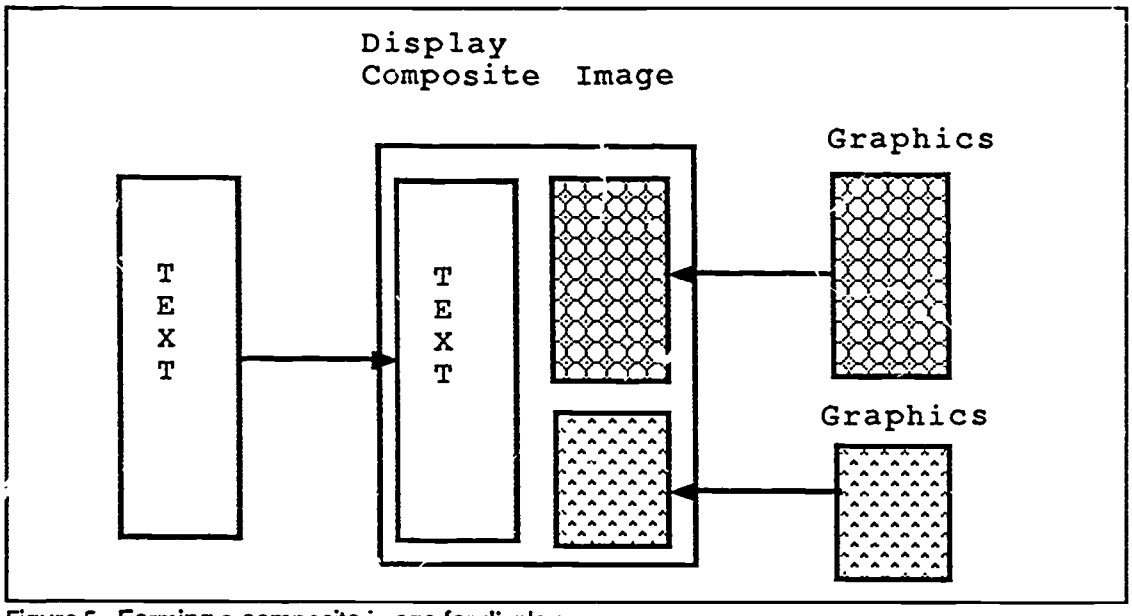


Figure 5. Forming a composite image for display.

made to resemble the printed page. As in the second and third cases, the indexed text database provides a very quick method of document access. The only disadvantage of this approach is that it is not easy to segment the image. Technology is available to permit this to be accomplished manually by using mouse-controlled cursors on the image screen, but not automatically; automatic segmentation or decomposition of electronic images remains a classical image processing problem. A consortium of Japanese companies is working on this problem, but does not expect to have it solved until the early 1990's [4].

System Architectures for Conversion

Once the method for converting paper-based documents to electronic form has been selected, the system architecture must be considered for large-scale document conversion. For a relatively small number of document pages, perhaps a few thousand, any single-user small computer system should suffice for document conversion. However, if hundreds of thousands or perhaps millions of pages must be converted, a larger computer system with many operators may be necessary for adequate throughput. Having several operators perform parts of the document conversion task simultaneously increases the conversion efficiency by reducing the overall time to perform the conversion. The question is: what type of computer architecture is best for this situation? Two example architectures will be considered and compared.

The Lister Hill Center has developed two experimental prototype systems for document conversion, one centralized and one distributed. Both systems are designed to perform the method of document conversion as described above in case 1, in which bound and fragile biomedical documents are converted to bit-mapped images which are subsequently compressed and archived on optical disk. The first system, possessing a centralized architecture based on a Digital Equipment Corporation PDP 11/44 minicomputer, is illustrated in figure 6 [5].

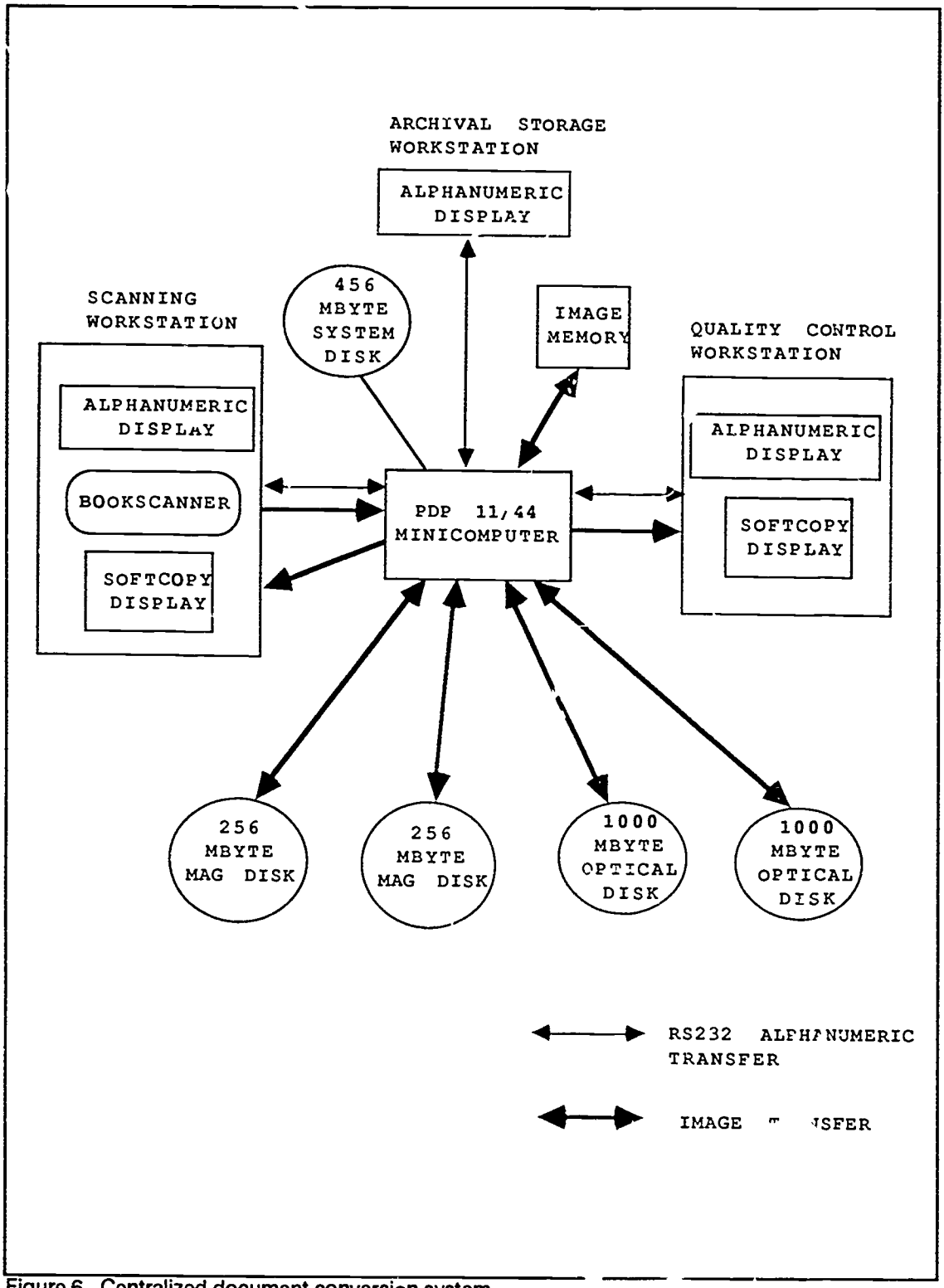


Figure 6. Centralized document conversion system.

It has three basic types of workstations: scanning, quality control, and archiving. The scanning workstation has three components: an alphanumeric display terminal for user interface; a bookscanner, which is a bound-volume scanner; and a softcopy display

device for displaying images of scanned pages. While a document is scanned at 200-points-per-inch resolution, its images are collected on one of two 256-megabyte magnetic disk drives. After scanning, an image document is available for quality control. The quality control workstation operator uses an alphanumeric display terminal for operator/machine interface and a softcopy display terminal for displaying the document images at 200-points-per-inch resolution. If a document contains scanning errors, the scanning operator is notified and the problem pages are rescanned. Once a document has passed quality control, it is available for archiving. The operator of the archiving workstation oversees the transfer of all magnetic disk-based image documents to optical disk. The optical disk used by this system is the Optimem 1000, which uses 1-gigabyte capacity media.

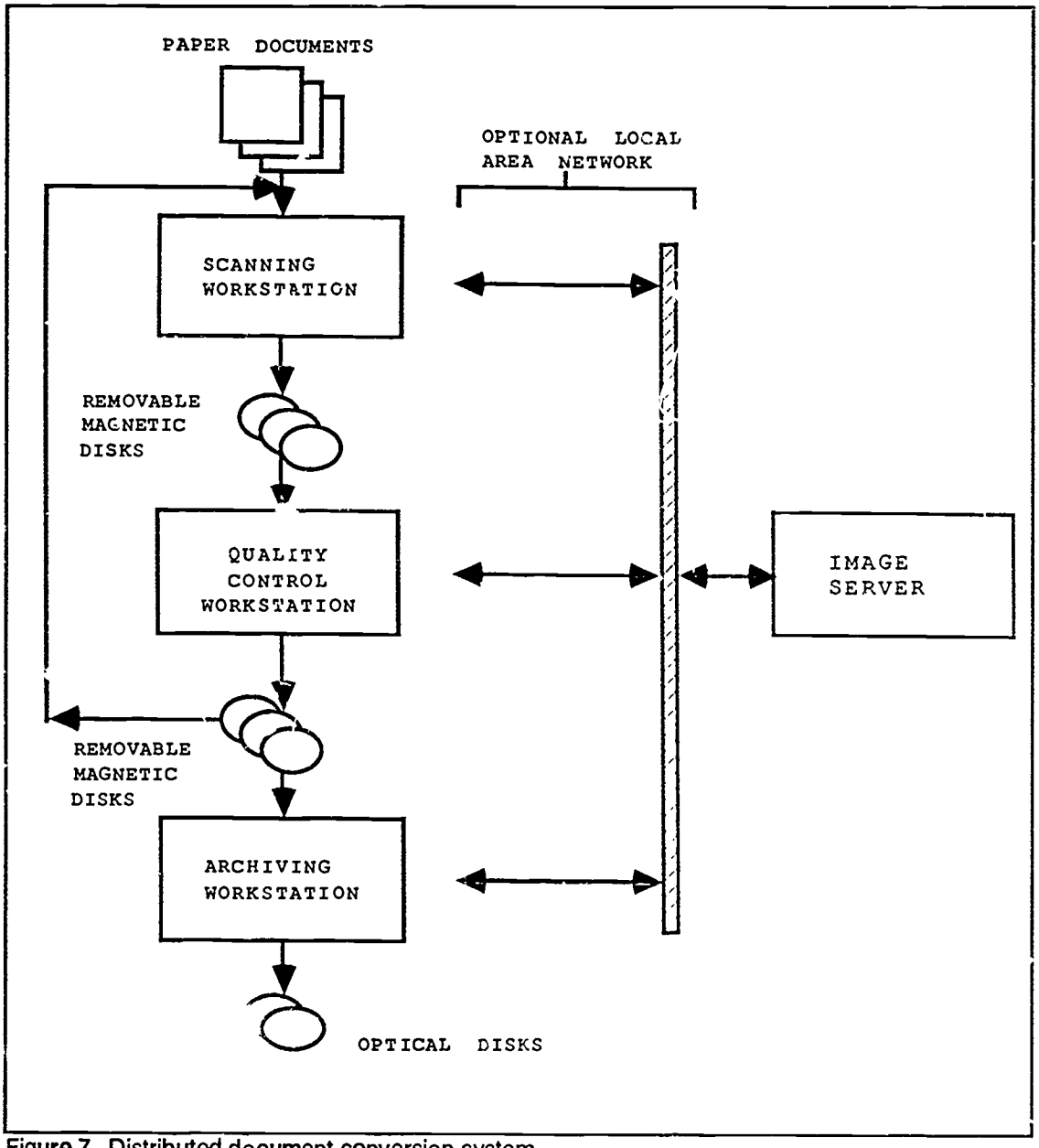


Figure 7. Distributed document conversion system.

An alternative to the centralized, single CPU architecture is a distributed architecture. Figure 7 illustrates a prototype distributed architecture recently developed by the Lister Hill Center [6]. In this architecture the processing functions are spread among several computers, each based on an IBM PC-AT-class computer. As in the centralized architecture, it has three basic types of workstations; each is a self-contained unit running under its own CPU. The inputs to the scanning workstation are paper documents and the outputs from the archiving workstation are optical disks. Two methods have been created for transferring document images from one workstation to another. The first is a loosely coupled design which utilizes removable magnetic media in the form of 20-megabyte Bernoulli cartridges. These cartridges, which plug into disk drives, are easily removable for physically transferring image documents from one workstation to another. The other method for image document transfer is a more tightly coupled design based on a local area network. This is a 10-megabit-per-second token ring design employing an image server having a large magnetic disk drive for storing and retrieving the image documents by the 3 workstations.

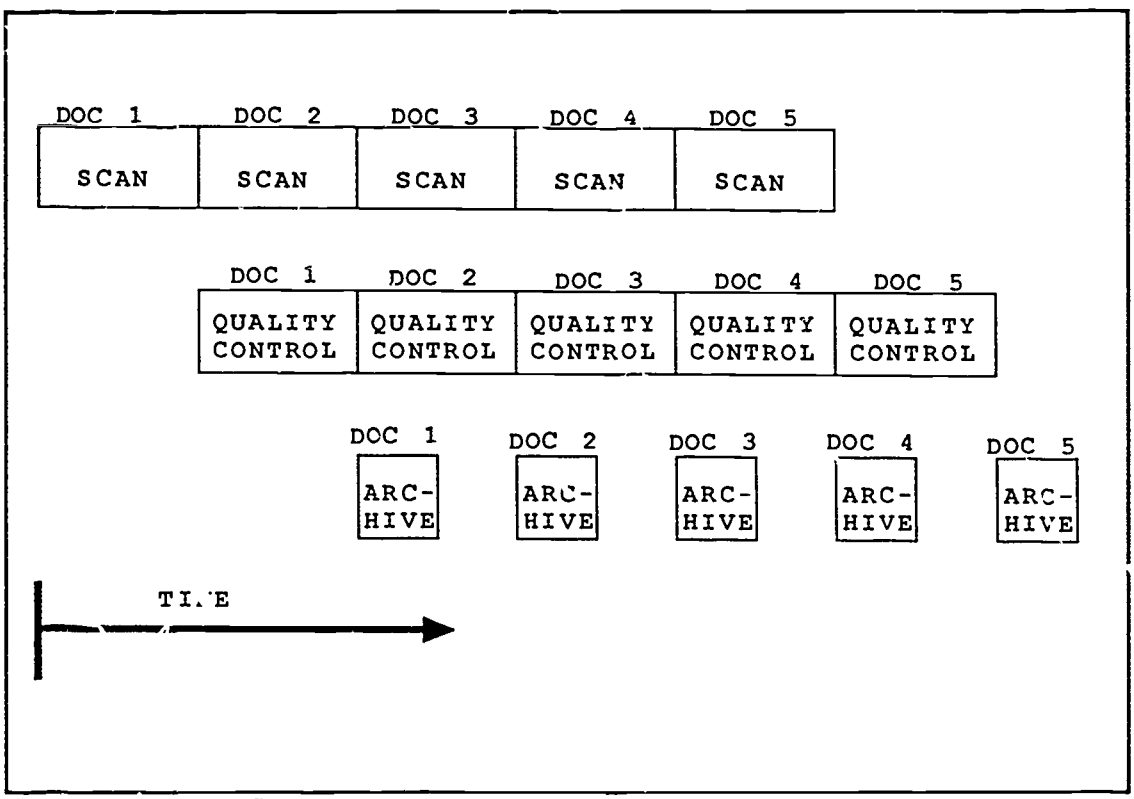


Figure 8. Procedural flow for document conversion.

Both the centralized and distributed architectures permit operations to be overlapped, as depicted in figure 8. Overlapping the operations by using several workstations, each to perform a piece of the document conversion process, permits higher overall throughput than that obtained using a single computer. While the two architectures are suitable for converting a large collection of documents to bit-mapped images, they are also applicable to the other techniques of document conversion. Scanning documents to produce bit-mapped images is the first step of all four methods of document conver-

sion. If OCR is introduced into the process, it would probably be best suited for insertion between the scanning process and the quality control process. This way the quality control workstation operator can correct the errors resulting from the OCR process. As to the question of which architecture is better, several factors indicate that the distributed architecture is preferable to the centralized architecture. First, tests indicate that a distributed architecture can deliver higher throughput than that obtained with a centralized architecture. The reason for this is that as a centralized architecture gets more and more users, those users all contend for the same set of shared image resources, namely the single CPU, the one I/O bus, the same main memory, and the same disk drives. And, since the task of document conversion is an I/O-intensive process because of the relatively large amount of data transferred for every image, as the number of users increases, each user's task will take longer. This is not true in a distributed architecture, particularly the loosely coupled design. Here, since each workstation is a dedicated computer performing a specific task, that task is barely affected by the number of other operators working on the document conversion process. A second advantage of the distributed architecture over the centralized one is that the entire system will be dysfunctional less often. In the centralized architecture, all operations cease if the CPU malfunctions. In the distributed architecture, if a single CPU goes down, only one workstation is affected, not the entire operation. The third advantage of the distributed architecture is that maintenance is easier. Since each workstation is based on a personal computer, only a few hardware components may fail in each workstation. This makes troubleshooting a relatively simple task. Maintenance is more complicated in a large centralized design since there are a large number of components; troubleshooting is more complex.

Image Enhancement

The last issue to be addressed is image enhancement. Each of the four methods of document conversion can be improved if the document-scanning process permits the produced "raw" image to be enhanced. The first, third, and fourth cases of document conversion as listed in table 2, all of which produce bit-mapped images, can benefit from image enhancement if the resulting images are made as legible as possible. The second, third, and fourth methods of document conversion may also benefit from image enhancement because they employ OCR as part of their processes. Optical character recognition can be a very time-consuming process, especially if error rates are high. As mentioned previously, OCR error rates can be reduced by enhancing the image during the scanning process. Image enhancement may include such tasks as removing borders, improving contrast, and removing stains.

One problem often encountered with the images of a scanned bound document is that a border, or "garbage" region, sometimes appears around the edges of the image. This border may be created by shadows cast by the pages behind the page currently being scanned, or by the background behind the document. The border is undesirable for several reasons. First, it is extraneous to the OCR process; it will slow down character recognition and perhaps result in lower accuracy. Second, for the case of bit-mapped images, the border region does not appear esthetically pleasing. Finally, for bitmapped images, the border reduces the overall compression ratio. By taking a representative

sample of the NLM collection, it has been shown that by removing the border from the image of a typical biomedical library document, the average compression ratio can be increased by 80 to 90 percent. Border removal is accomplished through a manual process at the scanning workstation. By using a mouse device with cursors, a scanning workstation operator can segment the image, entering the upper left and lower right coordinates of the image appearing within the borders. The computer can use this information not only to remove the borders, but also to center the image within the display.

Poor print-to-background contrast in the original document can affect the OCR recognition rate to a significant degree. Contrast problems could arise from the browning or yellowing of the paper due to age or might be a result of darkly colored paper stock. The images of poor-contrast documents may be improved during the scanning process either manually or automatically. Many commercial scanners permit a fixed scanning threshold to be set manually; this often can compensate for poor-contrast material. However, if the page exhibits variable contrast due to stains or mold, for example, a fixed threshold level might not suffice in all regions of the page. Few commercial scanners employ dynamic thresholding, an automatic technique. Dynamic thresholding acts like an automatic gain control that can change the scanning threshold based on the region of the page being scanned. This technique can also compensate for variations in lighting across the page [7]. The basic algorithm for dynamic thresholding compares the gray level of a picture element (pixel) with the average of gray-level values in the neighborhood about the pixel of interest. If the pixel is considerably darker than the average value of the neighborhood, it becomes a black dot. Otherwise, it becomes a white dot. It is possible to incorporate the algorithm into hardware to achieve fast real-time processing.

Poor contrast due to stains not only increases OCR error rates, but also degrades the quality of displayed bit-mapped images. In addition to dynamic thresholding, as mentioned above, another technique called peak detection may be used to remove stains. This is an electronic process which, by analyzing the frequency and amplitude of bit-to-bit changes in an image, brings out print obscured by stains [8].

Summary

Two of the key issues in converting paper documents to electronic form are conversion efficiency and image enhancement. Documents may be converted efficiently if they employ some method of image compression; several techniques for compression are available. Four methods of document conversion, each implementing one or more of the methods of compression are outlined, along with the appropriate retrieval techniques for each. For large-scale document conversion, the system architecture is critical, with distributed systems exhibiting several advantages over centralized systems. Image quality can affect conversion efficiency; several techniques are available for image enhancement.

References

[1] Report of the Task Force on Preservation Methods/Techniques, National Library of Medicine, 1985.

[2] F.L. Walker, et al.: "Issues in Archiving the Medical Literature Using Electronic Imaging Techniques," Proceedings of Electronic Imaging '88 International Electronic Imaging Exposition and Conference, Boston, Massachusetts, October 3-6, 1988, pp. 590-595.

[3] K.L. Anderson, et al.: "Image Compression Algorithms," Proceedings of Electronic Imaging '88 International Electronic Imaging Exposition and Conference, Boston, Massachusetts, October 3-6, 1988, pp. 398-401.

[4] Akio Tojo: "The Interoperable Database System Project and OSI promotion in Japan," DOCMIX State-of-the-Art and Market Requirements in Europe Electronic Image Banks Final Report, March 1988, pp. 226-234.

[5] G.R. Thoma, et al.: "A Prototype System for the Electronic Storage and Retrieval of Document Images." ACM Transactions on Office Information Systems, Vol. 3, No. 3, July 1985, pp. 279-91.

[6] F.L. Walker, et al.: "A Distributed Approach to Optical Disk-Based Document Storage and Retrieval," Proceedings of the 26th Annual Technical Symposium of the Washington, D.C., Chapter of the ACM, Gaithersburg, Maryland, June 1987, pp. 44-52.

[7] J.M. White, et al.: "Image Thresholding for Optical Character Recognition and Other Applications Requiring Character Image Extraction," IBM Journal of Research and Development, Volume 27, No. 4, July 1983, pp. 400-411.

[8] C. Hornbaker: "Recent Advances in Image Scanning," Proceedings of Electronic Imaging '88 International Electronic Imaging Exposition and Conference, Boston, Massachusetts, October 3-6, 1988, pp. 33-37.

FRANK L. WALKER

Senior Electronics Engineer,
National Library of Medicine

Frank Walker received both the B.S. and M.S. in electrical engineering from the University of Maryland. As a senior electronics engineer on the staff of the Lister Hill National Center for Biomedical Communications, the research and development arm of the National Library of Medicine, he has designed, developed, performed research, and published a number of papers on complex computer systems utilizing electronic imaging, primarily for the purpose of electronic document storage and retrieval.

Scanning and Digitizing Technology Employed in the National Agricultural Text Digitizing Project

Pamela Q. J. Andre, Nancy L. Eaton, and Judith A. Zidar
National Agricultural Library

Abstract

A brief summary of the National Agricultural Text Digitizing Project, its history, objectives, participants, and anticipated outcomes will be presented. The remainder of the paper will concentrate on the actual technology employed in the project. The system integrator chosen for the project, Science Applications International Corporation (SAIC), is utilizing a combination of off-the-shelf hardware and is writing software to integrate the hardware for a full text/graphics storage and retrieval application. Material is scanned and stored as bit-mapped images, then converted and stored as ASCII text. SAIC software controls the scanning and conversion and creates indexes and hypertext links between indexes, ASCII text, and bit-mapped images. Four different retrieval software packages are being tested for effectiveness of full text retrieval. Experience with scanning error rates and editorial requirements will be shared.

I. Introduction

The National Agricultural Library (NAL) and the land-grant libraries have entered into a cooperative project to assess imaging technology as a means of capturing and distributing textual information. The project will evaluate a turnkey scanning system to determine whether it is now possible to provide in-depth access to the literature of agriculture while at the same time preserving it from rapid deterioration.

The project was really started on September 12, 1986, when a panel of land-grant library directors was assembled to see a prototype system demonstrated at the National Agricultural Library. In addition, two consultants, Nancy L. Eaton, Director of Libraries, University of Vermont, and Edwin Brownrigg, the then Director of Library Automation, University of California System, were invited to evaluate the system's potential for use with agricultural information. At the conclusion of the meeting, participants unanimously approved a cooperative project between NAL and land-grant universities to pursue investigation of this technology.

This 2-year pilot project is jointly managed by NAL and the University of Vermont in the person of Nancy Eaton. The land-grant library community was invited to participate in the project and 44 libraries agreed to participate and donated funding to the effort. In addition, each participant agreed to supply a microcomputer to test the resulting disks while the project would supply CD-ROM players and high-resolution workstations.

A Technical Advisory Panel of librarians was established to guide the overall effort. They include: J. Beecher, North Dakota State; J. Harrar, University of Maryland;

N. Aldridge, Montana State; P. Gherman, Virginia Polytechnic Institute; and T. Shaughnessy, University of Missouri. Clifford Lynch, current Director of Library Automation at the University of California System, was asked to serve as technical consultant.

II. Project Description

The purpose of the project is to test the feasibility, costs, and effectiveness of newly emerging technologies for capturing page-images, providing access to their content, and disseminating them. This technology converts full text, including graphics and illustrations, to a bit-mapped digitized page-image and then processes that page-image through a "recognition engine" which converts the text to digitized ASCII code. The ASCII code is then processed by computer software and indexed. The resulting digitized text, bit-mapped page-images, and indexes can be stored on a variety of electronic storage devices for dissemination and retrieval at microcomputer workstations. For this project, CD-ROM will be the distribution medium.

The study will determine the costs of this method of capturing and converting textual data and the limitations of the processes used. Retrieval effectiveness using only the ASCII text will be compared against using both ASCII text and bit-mapped page images, both for analysis of retrieval satisfaction and accuracy and for production costs. Possibilities for electronic document delivery of the digitized collections via telecommunications networks also will be examined.

Although the scanning system developed and marketed by Science Applications International Corporation (SAIC) is very promising, it is still a developing technology. Reliability of the equipment, production costs, and effectiveness of the indexing and retrieval software currently available have not been field-tested with agricultural materials. This project will utilize the technology in a demonstration environment to gather experience with the technology and to gather concrete information on production costs and software retrieval design capabilities.

The National Agricultural Text Digitizing Project (NATDP) will be broken into three distinct phases:

- **Phase 1,** The Pilot Project, which will test the scanning system and a variety of indexing/search software systems. An in-depth evaluation will be conducted by the University of Vermont.

- **Phase 2,** System Refinement, which will include various enhancements to hardware and software in an in-depth project on acid rain.

- **Phase 3,** Telecommunications, which will include a study of alternatives and a pilot project.

III. Central Scanning Facility

The SAIC scanning system was installed at NAL in January of 1988. Using this system, a CD-ROM on the subject of aquaculture, the first of four pilot study disks we plan to publish, is almost completed. Each disk will be on a different topic and use a different retrieval package. The 44 land-grant libraries will evaluate these disks and, based on their responses, a retrieval package will be chosen for the multidisk set on acid rain to be published in phase 2 of the project at the end of 1989.

Herein is a brief description of the scanning system and an overview of the operational procedures. A discussion of the rather amazing abilities of this technology, as well as some of its inherent problems and limitations, will follow. Preliminary findings from our not-yet-completed in-depth analysis of conversion accuracy will also be presented.

System Description

In describing our system, it must be emphasized that this is a microcomputer-based scanning system. It is designed for handling thousands, but not millions, of pages of text and graphics. All the hardware components are off the shelf. The integration of these components, however, required very sophisticated interfacing software, which was custom designed for our project by SAIC.

The heart of the system is a PC AT 286 microcomputer with 640 K of RAM, a 230-MB hard disk, and a 1-MB video card for image display. The software employed is the PC DOS 3.1 operating system and the interfacing and operations software developed by SAIC for this project.

The Ricoh high-speed scanner has both a flatbed scanner, for scanning books and other bound publications, and an automatic document feeder for high-speed scanning of loose pages. The scanner looks and works like an office photocopier, except that it is controlled from the computer keyboard using various function keys. Pages are scanned at a resolution of 300 dpi. When a page is scanned, a digital bit-mapped image of that page is created and stored on the computer hard disk using CCITT Group 4 compression.

The Palantir RS 9000 recognition engine performs the text conversion from the page-images stored in the computer. It takes 15-20 seconds to create a page of ASCII text from a page-image.

The LaserView high-resolution monitor is used to display the page-images at a resolution of 150 dpi. The images are of an extremely high quality, although the image format required by this monitor is not a standard one (such as TIFF). The monitor contains a DOS window which allows display of ASCII text and the use of standard DOS commands.

Both the text and the page-images are archived on a Maxtor 5 1/4-inch WORM disk. These disks can hold approximately 400 MB of data on each side. A 9-track tape drive is also interfaced with our system and is used for premastering the CD-ROMs. The

printer used for this project is the Ricoh laser printer. It can print both text and page-images at a resolution of 300 dpi.

Operations

Although each project of this nature contains unique applications, some common aspects can be seen throughout. In order to help those of you who are still in the planning stage of your projects, the actual operational procedures of the NATDP will now be discussed. These operational phases can be defined as data pre-preparation, data input, image scanning, conversion to ASCII, and text editing.

Data Pre-preparation. For the NATDP, publications, primarily reference materials, are being scanned. These include books, journal articles, pamphlets, dissertations, bibliographies, etc. After all the publications have been selected, a number is assigned to each, beginning with 1 and going up to 62. Next, a list of the publications is prepared with their corresponding publication numbers. Then, each publication is examined to determine how it should be divided into documents. In the database, each document represents a record. Defining a document is a critical issue since the system can treat each page, each chapter, each book, etc., as a separate document. For the purposes of this project, each chapter of a book constitutes a document in the database; in the case of small publications such as pamphlets, the entire pamphlet constitutes a document. A worksheet is then prepared for each document, giving title, page numbers, and other pertinent information. This pre-preparation is completed before scanning.

Data Input Phase. Once data pre-preparation is complete, the data input phase is initiated. The operational software is menu driven, and when the option for scanning a new document is selected, a unique number is automatically assigned to that document and the input screens are called up. The document number and date are entered in a handwritten log, along with a brief description. Later, the date that the document completes each processing step is entered into the log. By maintaining this log, the status of each document can be seen at any point in time. The operational software also tracks the documents through the various processing steps, but it tracks by task, whereas the log tracks by document number.

After logging the document number, bibliographic data about the document are entered using the input screens. Every scanning system requires the keying of a descriptive record for each document for system control. This descriptive record is called a relational header, the nature of which is dependent on the kind of data being scanned. Our headers include the title of the publication, the title of the document (in the case of a book chapter, for example), authors, publication date, pagination, and so on. These bibliographic data will later be attached to the text of

the document and become a part of the database. Once all the required data have been entered, the next phase is implemented.

Image Scanning. As previously mentioned, this is somewhat like making a photocopy, except that the functions are controlled from the computer keyboard. It is also more complicated than photocopying because the operator must be sure that each page is straight and must tell the computer whether the image is text or graphic, single- or multi-column, and landscape or portrait in orientation. As a page is scanned, its bit-mapped image appears on the high-resolution monitor where the operator can view it and decide whether it is acceptable. If it is not acceptable, it is rescanned; if it is acceptable, the operator goes on to the next one. It is from this bit-mapped image that the ASCII conversion is performed.

Conversion To ASCII. This is usually done as an overnight batch job. All the documents scanned in a day are selected as a group and then run through the Palantir recognition engine at night. The Palantir outputs an ASCII text file for each document scanned. In the morning, each ASCII file is checked to determine if all the text pages were converted and if the conversion was done properly. If, for example, it is discovered that a page is missing, the document can be retrieved, the missing page inserted, and it can be run through the Palantir again. The Palantir will insert the page of text in its proper place in the ASCII file.

Text Editing. This is a step we did not expect to do originally, the need for which will be discussed under system limitations. WordPerfect is the word processing software package used to edit the ASCII text files. When the editing of a document is complete, the system attaches the bibliographic data to the text file. The document is then archived to a WORM disk for permanent storage.

For the Aquaculture disk, full MARC cataloging records are also being added, one for each publication. These are stored in ASCII in a standard OCLC display format. They are searchable as text and, like all other elements in the database, can be downloaded for local use.

After all documents for a database have been archived to WORM, the index is created by running the ASCII files through indexing software. This is usually done as an overnight batch job.

Once the indexing is complete, premastering is initiated. The archived documents on the WORM are used to simulate the database so it can be tested and reviewed. Links are created between certain documents--for example, all the chapters from a book are linked and the MARC cataloging records are each linked to their associated documents. The bit-mapped page images are automatically linked to their ASCII text files

by the software. After review of the database is complete, the entire database is written to 9-track tape which will be sent to a mastering facility for reproduction on CD-ROM.

The first NATDP CD-ROM, on the topic of aquaculture, is expected in February 1989. The database will be made up of the following elements:

- Page-images (one for each page)
- Full-text ASCII files (one for each document)
- Bibliographic records (one for each document)
- MARC cataloging records (one for each publication)

Technology Assessment

We will continue to assess the technology as we go through the pilot study and on to phase 2 for the multidisk acid rain material. Some of the system capabilities and limitations have already been identified and are discussed below.

System Capabilities. Page scanning is fast and easy to perform, and it produces high-quality digital page-images. For pages that are to be converted to ASCII, the operator must take extra time and care with settings for the brightness, page orientation (landscape vs. portrait), and columnation (single vs. multiple). This does require a high degree of concentration, but the settings are made with function keys or can be selected from a menu.

Text recognition is also rapid and easy, and can be done in a batch mode on as many pages as are ready to be converted. Several page variations can be accommodated, including portrait or landscape text (though not both on the same page); single- or multi-column; and page sizes from 3- by 5-inches to 11- by 17-inches. There is a "dithering" setting for capturing halftones in pictures, and this works well for all except very dark photographs. The recognition server appears to have omnifont recognition--to date, we have not encountered a font that it could not convert to ASCII, including italics and dot matrix, and can work with character sizes ranging from 4 to 28 points. These features appear to be common to most midprice-range text recognition engines.

Limitations. There are certain limitations inherent to the technology, most of them having to do with the text recognition step of the process.

1. Pictures vs. text: When a page contains both text and graphics, it may be necessary to choose between the setting that is best for the graphic and that which is best for the text, or the page may need to be scanned twice. The dithering option that allows the capture of halftones in pictures prevents the recognition server from identifying the text on the same page. As mentioned above, very dark photographs do not

turn out well, even with the dithering option. These are worsened when the brightness setting must be darkened for the capture of text.

2. Complex page formats: More than any other limitation, the complexity and diversity of page formats in published literature present the most problems. Although the recognition server utilizes artificial intelligence to do a remarkable job of discerning page layouts, it can make mistakes in text sequence during conversion of multicolumn pages. This is especially true when different page layouts are used within the same chapter or article. The result is ASCII files that take appreciable human effort to correct during the editing phase. Multicolumn pages that include tables present further problems since these must be converted as single column if the table is to be included in the ASCII. Then the editor must manually rearrange and concatenate the columns of text in order to preserve retrieval integrity.

3. Graffiti: When the recognition server works on images that contain both text and pictures, it uses artificial intelligence to identify (and then ignore) the pictures. During this process, however, it will often output nonsense characters until it concludes that a picture is present. We refer to these nonsense characters as "graffiti." Graffiti may also be output if the image contains a wide, dark border caused by the shadow of the book binding (the same kind of shadow border that appears in a photocopy). This graffiti must be edited out of the ASCII text, or it will play havoc with the indexing software used to create the full-text index. The editors also delete multiple lines of blank spaces which the recognition server has dutifully output for pages of text where the text ends near the top of the page but is numbered at the bottom, with blank space in between. This type of deletion is done so that users do not have to page through blank screens as they read the ASCII text.

4. Problem characters: Certain characters present special OCR problems to the recognition server. In some plain fonts, such as Helvetica, the lower case 'el' (l), the upper case 'eye' (I), and the 'one' (1), all look very much alike to the server. The 'zero' (0) and 'oh' (O, o) may also be indistinguishable. The degree sign cannot be distinguished from a 0 or an O, since the size of the character is irrelevant to the server; thus, '60°' looks like '600.' Non-Roman characters, such as those that commonly appear in chemical formulas and mathematical equations, are not recognized at all. All of these may appear in the ASCII text as errors, presenting special editing problems because they are difficult and time consuming to find and correct.

There are two other limitations which should be pointed out, and which may or may not present problems, depending on how a system is configured.

1. Boldface, underscore, etc.: The text output by the NATDP recognition server is in generic ASCII code. It contains no high-level control codes for boldface, underscore, italics, etc. However, the Calera server can be configured to output SGML codes for these characteristics, and most other mid-level servers can probably do the same. Then, special word processing or retrieval software can display the ASCII text with the appropriate boldface, underscores, etc.

2. Text wider than 80 characters: Many publications contain text that is wider than 80 characters, and the text output by the recognition server will be just as wide. If the text includes formatted data, such as tables, there are two options: (a) the data must be manually reformatted during the text editing process; or (b) the retrieval software must allow the display of wide data, either by allowing the user to window over or by shrinking the text displayed on the monitor. Many full-text retrieval packages will automatically "wrap" data that are wider than 76 or 78 characters; this is acceptable for text that is unformatted, but is not acceptable for tabular and other formatted data. Option (b) will greatly reduce data preparation time and should be a critical factor when selecting a retrieval package.

To summarize, any page within the size limitation of 3- by 5-inches to 11- by 17-inches may be scanned, and the resulting page-image will closely represent the page contents. But for text recognition, the best results are gained with pages of text that closely approximate typewritten pages, i.e., fonts such as Times Roman (where the serif helps demarcate each character), single column, and not right justified.

Error Analysis

NAL is currently analyzing the first 2,000 pages of converted text to determine the recognition error rate. In this analysis, we:

1. Count word errors, not letter by letter. A word with three characters wrong is no more difficult to correct than a word with one error; we do not have to decide how to count an error where one character is converted to two, or two are converted to one.

2. Take the percentage of errors within a document (e.g., a book chapter), not page by page. This seems to be a more realistic approach since this is how the text editing is done.

3. Factor in the "format" errors and the graffiti. These can take longer to edit than the OCR errors, and must be taken into account.

Preliminary results show a wide variation in error rates (from 1 to 15 percent). Important factors affecting these rates appear to be:

- Font size
- Font style
- Quality of original (letter clarity, contrast)
- Page format

The data we have gathered will be statistically analyzed, and the results will be published at a later date.

IV. Phase I: The Pilot Project

The first test disk of the pilot project will contain 4,000 pages of aquaculture material, using Textware retrieval software. There will be three additional test disks, as follows:

- Agent orange materials / Personal Librarian Software (PLS)
- Food irradiation materials / Quantum Leap software
- International agricultural research materials / KA2 software

The agent orange and food irradiation collections will be scanned by NAL, which will also produce the test disks. The PLS and Quantum Access software are being provided for demonstration purposes under special agreements between those software vendors and the University of Vermont. SAIC is adapting its image software to link the images into the PLS and Quantum Access software so that the images can be displayed on the full-text/graphics workstations.

The fourth disk, containing the international agricultural research materials, is being produced by the Consultative Group on International Agricultural Research (CGIAR), which is supported by the United Nations and the World Bank; CGIAR's mission is to improve the quantity and quality of food production in developing countries. CGIAR is working with Knowledge Access Corporation in the conversion of the data and the production of one test disk for field testing in Third World countries having research experiment stations. This disk will be made available to NAL for testing with the U.S. land-grant libraries as well as the international agricultural community.

There will be an in-depth three-part evaluation process for these four pilot disks and associated workstations. The evaluation is the responsibility of the University of Vermont. The three areas of evaluation are the retrieval software, the central scanning facility, and the field tests of the disks and workstations.

> **Selection of retrieval software.** Considerations that have gone into the selection of the retrieval software include the requirement that the software be able to link the ASCII text and the bit-mapped images; indexing and retrieval features; ease of use (including quality of the documentation) by the user in the file; type of hardware required; and the ability to truncate.

Central scanning facility at NAL. NAL staff will be responsible for obtaining data on error rates, amount of editing required, types of materials which do not lend themselves to conversion from the bit-mapped images to ASCII text, production controls, problem in and refinement of the editing processes, and any problems in the actual production of the CD-ROM disks. They will recommend refinements or upgrades to the central hardware and software. The expectation, once we complete the pilot project, is that NAL will make this facility available to land-grant and other appropriate libraries or agencies as a centralized scanning facility for our own materials and for others' materials on a contract or fee basis.

Fieldtests. All 44 of the participating land-grant libraries plus CGIAR libraries for the fourth disk will be asked to conduct formal evaluations of the disks and workstation configurations. There will be two workstation configurations initially, though a third configuration is of interest, based upon our experience to date with the scanning of materials.

The first configuration, an ASCII-only microcomputer workstation equipped with CD-ROM disk drives and laser or dot matrix printers, will be tested by all but five of the land-grant libraries. These stations can search and print only from the ASCII text. Images which could not be digitized are omitted; notes are added to the text, indicating that tables or graphics are missing from the ASCII text. We are interested in the acceptance of ASCII-only without graphics and in evaluating acceptance under different levels of error rates.

The second configuration, a full-text/graphics workstation using the LaserData compression/decompression system and laser printers and having the capability to search ASCII and also display and print bit-mapped images, will be tested by five land-grant libraries. These are Texas A&M University, Clemson University, the University of California at Davis, Virginia Polytechnic Institute, and the University of Hawaii. The University of Vermont is an alpha test site along with NAL and, as such, will be involved in premanufacturing quality control testing as well as testing of the full-text/graphics workstation with actual users.

The third configuration we are interested in evaluating, though it was not in the original project concept, is an ASCII workstation which cannot view the bit-mapped image but which could print out the bit-mapped image. The experience which has led us to want to explore this third configuration will be discussed below under phase 2.

All 44 land-grant libraries will test the pilot disks with 10 individuals: the project manager, 3 librarians, 3 research faculty members, and 3 students. This will provide a test group of 440 individuals for each disk. In addition, CGIAR will test the fourth disk with approximately 50 libraries in Third World countries. The University of Vermont's biometry unit has aided in the design of the survey instruments and will key and con-

duct the statistical analysis for us. The fieldtests will have users of the systems complete scripted searches and a questionnaire as they go through the searches.

V. Phase 2: System Enhancements and the Acid Rain Collection

Once the 4 test disks have been produced and field-tested, phase 2 will begin. It includes enhancements to the scanning system/editing station, the retrieval workstations, and the software as well as production of the first major collection for distribution to the land-grant libraries. This collection is a full-text database of Canadian documents on acid rain, the information being collected and produced by the University of Vermont under a Higher Education Act, Title II-C, grant.

Scanning system/editing station. Some enhancements to the central equipment and editing station have already taken place during the first year of the project. These include an upgrade to the Calera (previously Palantir) intelligent recognition engine which converts from the bit-mapped image to ASCII; this was done to increase the speed of digitizing by two to three times. A second enhancement was the addition of a new editing software package which allows dual columns, to compare unedited and edited text. A local area network was also added to the configuration to allow the scanning station and editing station to be used simultaneously. We are now considering an upgrade to the editing workstation controller to increase speed. Also under consideration are upgrades from the single 5.25-inch WORM drive to the larger capacity 12-inch drive or to a dual 5.25-inch drive unit; the dual drive unit would make copying of data from WORM disk to WORM disk easier. SAIC has been asked to help evaluate the desirability of adding a microform scanner to handle conversion of microform collections; a new Minolta film/fiche scanner is being considered. Finally, a Meridian Data CD-Publisher system has been ordered to aid NAL in premanufacturing testing and quality control of the databases; the CD-Publisher simulates the database/disk layouts, file structures, and resulting retrieval response times. Other improvements to the digitizing and editing systems undoubtedly will become available before this project is complete.

Retrieval workstations. SAIC will have to develop the software linkages between its image software and the retrieval software packages being tested in order for the full-text/graphics workstations to be able to view the bit-mapped images on the graphics monitors. For PLS software to be used with the second disk, ASCII workstations must be upgraded and standard monitors added to the graphics workstations because the LaserData units cannot display WINDOWS. KA2 software may require specific graphics boards for compatibility and also requires a standard monitor for the full-text/graphics workstation. A third configuration which is of much interest, is an ASCII station upgraded with a compression/decompression board and the addition by SAIC of a

software link between the ASCII text and bit-mapped image in order to allow the ASCII station to print the bit-mapped image.

Software. The evaluation process can be expected to produce user feedback on retrieval software capabilities, which would be desirable in the retrieval package finally selected for production of the acid rain disks. Likewise, editing staff has already identified desirable changes in the text editing software capabilities; linkages between the SAIC image management software and retrieval software have to be provided.

After the four test disks have been evaluated, one of the tested retrieval software packages will be selected for use with subsequent agricultural collections to be produced on CD-ROM. The first such collection will be the full-text collection of Canadian documents on acid rain. The Title II-C grant provides for scanning of 35,000 pages of text; these will be distributed to land-grant libraries and cooperating Canadian libraries and agencies. The ASCII text and bit-mapped images will be stored onto CD-ROM disks. At the estimated 4,000 - 6,000 pages per disk, this will require approximately 6 to 9 disks. In addition to making this material available in each State, since these documents are virtually unavailable in the United States at present, it will also test the ability to store and retrieve collections of disks in a library setting.

VI. Phase 3: Telecommunications and Document Delivery

A third objective of the NATDP is to experiment with electronic document delivery of the data being produced on CD-ROM disks. The first stage of this effort was to contract with Clifford Lynch, Director of Library Automation, the University of California System. Dr. Lynch has just completed a survey of state-of-the-art telecommunications technology available to NAL for electronic transmission of full text and graphics and has suggested a series of actions and projects to begin to test these options. A draft of this report was shared with the NATDP Advisory Panel at its June 1988 meeting in New Orleans, and NAL has already begun to act on some of its recommendations.

NAL is in the process of completing a cooperative agreement with North Carolina State University. Under terms of the agreement, Dr. Henry Shaeffer, Vice Provost for Information Systems, and Susan Nutter, Director of Libraries, will work with NAL to connect two sites at the NCSU campus to NAL via the Internet to transmit documents from NAL to NCSU. Although there is considerable interest in the ability to transmit documents internationally, the NATDP Advisory Panel has indicated that such projects should be viewed as separate from the NATDP because of the complexity of such an undertaking. Any investigation of international telecommunications requirements for international document delivery may go on at NAL in parallel to the NADTP, but will not be reconsidered as a part of this project. In the interim, CD-ROM products which can be distributed internationally without a telecommunications network are of interest to this project, as evidenced by the participation of CGIAR in the NATDP. It is felt that a large part of the need for international electronic dissemination of information can be met via use of CD-ROM technology, particularly in Third World countries where telecommunications networks are unreliable or nonexistent.

Much has been learned during the first year of the NATDP. It is an exciting pilot project and a very successful collaboration among the land-grant libraries, the National Agricultural Library, and CGIAR.

PAMELA Q. J. ANDRE

Chief, Information Systems Division,
National Agricultural Library

As chief of the Information Systems Division, Pamela Andre is responsible for planning and implementing automated systems for the National Agricultural Library. Specific projects implemented include the procurement and installation of an integrated library system; the development of two digital videodisks with mixed media content; the development of an analog videodisk with historic photos from the USDA Forest Service; the development of a CD-ROM containing the 2.5 million record AGRICOLA database; and the implementation of the National Agricultural Text Digitizing Project in which textual information is captured for distribution via CD-ROM. From 1968 to 1984 she held various positions at the Library of Congress, including Assistant Chief of the MARC Editorial Division from 1981-84, in which she was responsible for planning and directing the conversion of bibliographic and authority records to machine-readable form and during which she served as Operations Manager for the planning of the Optical Digital Disk Project. She was also Senior Staff Analyst (1978-81) and Computer Systems Analyst (1968-78) in the Automated Systems Office.

She is active in the American Library Association (ALA) and the American Society of Information Science (ASIS). She has been a member of the ALA Library and Information Technology Association (LITA) Emerging Technology Committee (1985-87) and a chair of the LITA Optical Systems Discussion Group (1986-87) and is presently a member of the Library of Congress Network Advisory Committee. She has presented lectures and papers on various library automation topics.

NANCY L. EATON

Director of Libraries & Media Service,
University of Vermont

Nancy L. Eaton received her B.A. degree fror Stanford University in 1965 and her M.L S. from the University of Texas at Austin in 1968. She has held various professional positions specializing in library automation and management since 1968, including positions at the University of Texas at Austin (1968-74), the State University of New York at Stony Brook (1974-76), and the Atlanta Public Library (1976-82). She became Director of Libraries at the University of Vermont in 1982, and was promoted to Director of Libraries and Media Services at the University of Vermont in 1987. In addition, Ms. Eaton has been President of the Library and Information Technology Association, a division of the American Library Association, and currently is a member of the OCLC Board of Trustees, the New England Library Information Network Board of Trustees, and the Center for Research Libraries Board of Directors. In July of 1988, she was elected to chair the newly formed National Agricultural Information Network. Under a cooperative agreement between the National Agricultural Library and the University of Vermont, she is National Project Manager for the National Agricultural Text Digitizing Project, a 3-year project to convert full text to optical disk for storage, retrieval, and distribution.

JUDITH A. ZIDAR

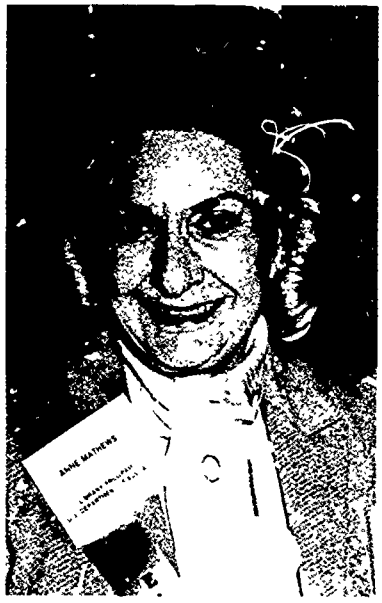
Operations Coordinator, NATDP,
National Agricultural Library

Judi Zidar began working with electronic databases in 1983. Since then, she has served as the database manager for NAL's full-text laser videodisk pilot project, which evaluated optical storage and retrieval of text plus graphics. She has also served as database manager for NAL's second full-text videodisk project (Laser II), which evaluated conversion of both machine-readable and optically scanned data into a specific retrieval format for storage on videodisk. She is currently the operations coordinator for the National Agricultural Text Digitizing Project.

SESSION 3

ANNE J. MATHEWS,
MODERATOR

Director, Library Programs,
U.S. Department of Education



Dr. Anne J. Mathews is Director of the Office of Library Programs for the U.S. Department of Education. Before accepting her position in Washington, D.C. in 1985, she was a professor at the University of Denver's Graduate School of Librarianship and Information Management, where she taught courses in administration, reference, marketing, public relations, and public services.

Dr. Mathews has held a number of positions since receiving her M.A. in Library Science and Ph.D. in Speech Communication, both from the University of Denver. During her career, she has served as Reference Librarian at Oregon State University, Program Director of the Central Colorado Library System, Consultant to the Colorado State Library and to ABC-CLIO, and has had more than 15 years of experience as an interviewer and trainer for the National Opinion Research Center, Louis Harris Company, and Colorado Market Research Services, Inc.

In addition to administering the Office of Library Programs, Dr. Mathews is a consultant to the U.S. Information Agency and the U.S. Agency for International Development. She currently serves on the National Advisory Committee of the Library of Congress, the Visionary Leaders Committee of the American Library Association, and the Book and Library Committee of USIA. She is a former member of the Council of ALA, has served on the Board of Directors of Friends of Libraries, U.S.A., and is a past President of the Colorado Library Association.

Developing an Optical Disk System for Adult Education Manuscripts: The Kellogg Project at Syracuse University

Terrance Keenan and Elizabeth Carley Oddy
Kellogg Project, Syracuse University

Abstract

Beginning in September 1986, funding from the Kellogg Foundation gave the Kellogg Project at Syracuse University the opportunity to place a large collection of adult and continuing education manuscript materials into an image system using write-once optical disks. In our work, we became very much aware of the not-always-obvious differences between archives and libraries. We found ourselves focusing on vocabulary and information organization issues: how should we describe and define "things" consistently in an archival environment? Could we use OCR to index content words automatically? Should we try to apply subject terms? How might we incorporate provenance information? The work we have done suggests that there is promise in optical disk technology for preservation and compact storage. There is also potential for alternate means of access to historical documents through a combination of browsable images, direct searches of content or descriptive words, and access to provenance information. The kind of system we are designing with our vendor, Plexus Computers, Inc., could change the way historians approach research. A prototype system was implemented in December 1987, and the large-scale version will be installed in December 1988. Larger issues remain. Copyright law does not yet adequately address the problems presented by electronic and optical systems. Institutions are beginning to join together to design cooperative documentation strategies for collection development. Disciplinary centers are evolving. New technology changes how and what kind of information can be used; it also complicates the practical and ethical issues of choosing just what to keep from the past.

Introduction

We are going to describe an innovative project involving original manuscript materials and optical disk technology. Our presentation is a little unorthodox and has an informal air about it. We try to describe the system's evolution, from concept--so new as to be like a dream--to realization and implementation. Lessons were often hard, and the dream did not always fit nicely into the reality we encountered. The purpose of the effort, however, has been to break new ground. We think that narrating the history of what we have done, with all its foibles and contradictions, is a good way to convey how our project happened.

We are enthusiastic about the work we have put into this enterprise. At the same time, we are reminded of a little story. A wise man was walking with some friends one day, when an adept ran up to him shouting, "Master, I've done it! I've finally done it!" "What is it you have done, my friend?" the wise man said. "I have at last learned to

walk on water," he replied. "And how long did it take you to be able to do this marvelous thing?" the wise man inquired. "Ten long years of hard labor, concentration, and sacrifice," was the reply. "That is a great pity," said the wise man. "For a few coins you could have taken the ferry." Innovation is not good in itself; it is merely something new. On the other hand, searching for new answers to old problems allows one to discover the possible.

First steps

One man's dream lies behind all our efforts. "The adult education library-for-the-world is within our grasp," said Roger Heimstra, Chair of the Program in Adult Education at Syracuse University, when he proposed an eccentric and revolutionary approach to research in manuscript materials in 1986. The Kellogg Foundation was sufficiently impressed with Dr. Hiemstra's dream to award $3.7 million to the School of Education in September of that year, to create the Syracuse University Kellogg Project. Just what was this dream? And why Syracuse? And why so much money from such a large foundation? And, further, what was there about adult education that warranted such rich attention?

The answers can be found by stepping back a bit into the recent past. Although work in adult education existed earlier, it was in the late1920's that "adult education" became a self-conscious, defined concept. Academic study of that concept followed, along with creation of organizations that focused on adult education activities throughout society. "Adult and continuing education" includes institutionally based efforts such as community and evening colleges, as well as extension programs, self-directed learning, literacy projects industrial and military education, the settlement and Chatauqua movements, coteries, and the countless community organizations that bring learning to adults. It touches everything from the Peace Corps and civil rights to military training programs and the women's and labor movements.

At the close of World War II, there was great change in the makeup of the work force. Employers hired returning veterans, displacing the women who had worked during wartime. Men who had learned new skills in military service returned to old jobs. The GI Bill permitted people to go to college who had been unable to go earlier. Industry had to realign production, meeting consumer needs rather than those of the military. A new world emerged, with a mature, aware population that had real, if poorly defined, educative needs.

Dr. Alexander N. Charters, a young and aggressive member of several continuing education organizations and member of the Syracuse University School of Education faculty, began in the 1940's what became a career-long effort to acquire records of significant organizations in the adult education field. In time, he was also able to bring in the personal papers of key individuals, people who had been colleagues and pioneers. Papers of the Adult Education Association of the United States, the Fund for Adult Education, the Commission of Professors of Adult Education, Alexander Liveright, Malcolm Knowles--some 60 collections in all--found their way over 40 years into a manuscript collection at Syracuse University.

When Roger Hiemstra joined the faculty in 1980, he found a large research collection, but one that had grown haphazardly. At least one-third was unprocessed and inaccessible to scholars, while the available materials had been rarely used and poorly understood. Dr. Hiemstra is a chronic optimist and a proponent of new technology, especially when it concerns "distance education," that aspect of adult education in which students can learn without having physically to attend an academic institution. New technologies hold promise for great flexibility and innovation in research and practice. One of the most exciting promises comes from optical disk technology. Not only can enormous amounts of original, unique material be captured for compact storage and preservation, but it is theoretically also possible for researchers to examine images of documents without having to come to the library to do their work. Remote access to original documents of historical significance is the core of the dream behind the Kellogg Project.

If such a system were then hooked to systems or collections at other institutions, what then? Today's bibliographic networks, such as OCLC and RLIN, allow people to exchange information about collections, but they don't include actual contents of those collections. The concept--storing images on optical disk for later retrieval over networks--presents a means of bringing real things to real users, the scholars and practitioners, the true end-users of our system.

At Syracuse, adult and continuing education is represented by records of all types, from handwritten manuscripts to interoffice memos, to conference proceedings, to thousands of pieces of correspondence in all conceivable formats. The materials constitute the single largest collection of English language historical documents in the field Further, included among the documents are thousands of photographs, audio and video tapes (both cassette and reel-to-reel), and color slides. Such mixed media do not usually lend themselves to a common means of access or a common database of artifacts. However, through digitization of both image and sound, the improbable will become actual.

At least that was the sales pitch Dr. Hiemstra and his small team used to sell the idea to the Kellogg Foundation and, ultimately, to the university. This was no easy task. The scheme was complex and required the cooperation of several schools and departments. Anyone connected with a large university knows that these social and political environments are as interwoven and subtle as those of any government.

The university and the Kellogg Foundation engaged in lengthy negotiations. Several things had to be clarified. From which departments would staff be paid? To whom would the general office of the project report? These questions brought in the School of Education and its Adult Education Program. The Syracuse University libraries also became part of the discussion because they maintain and house the collections and would inherit the optical disk system at the close of the project. The School of Computer and Information Science and the School of Information Studies joined the group of interested parties, because considerable expertise about computers and information

retrieval was needed. Finally, consultation with Academic Computing Services was required before development of an expensive computer installation could take place.

Although the Kellogg Project was conceived as a vehicle for expanding access to our adult education manuscripts, it grew to include several other mechanisms for communicating among adult educators and about adult education. The International Information Sharing Network, linking adult educators from over 50 countries; the Visiting Scholar Program, which brings visitors to our collections to do research; AEDNET, an electronic adult education network; New Horizons in Adult Education, an electronic journal produced by graduate students and distributed over AEDNET; and our experimental efforts to teach graduate courses at a distance using computers--all became important components of the Kellogg Project. However, the centerpiece remains the optical disk system and improved access to our research collections.

Special nature of the collections

Some basic characteristics of our collections should be understood before we describe our system. Many people equate libraries with archival or manuscript collections. After all, both collect quantities of information. How do they differ, and why does it matter?

In a general sense, libraries maintain books and periodicals. These are published items, duplicated in many places. By contrast, archives and manuscript collections generally hold unpublished, often unique, written or printed material. Usually, a collection is available in a single repository.

Archives contain records generated by an institution or an organization and maintained by that operation. When an organization becomes defunct or when its records are transferred to a repository for preservation, they become manuscript materials. Manuscript collections are most often thought of as the personal papers of authors, but they include organizational records as well, if such records are no longer being collected or managed by the generating organization.

People come to libraries looking for works by particular authors, for specific titles, or for information about subjects. Card catalogs, whether manual or automated, are organized to respond to such needs. Archival and manuscript materials, on the other hand, tend to be organized by type of document: correspondence, business records, minutes, membership lists, etc. Grouped corporate correspondence, for example, is unlikely to have a single author, or an obvious subject. Archivists help researchers to find appropriate information by matching researchers' questions against archivists' knowledge of collections and the entities that produced those collections. Which committees would have dealt with the subject? At what time was the subject under active discussion? Were particular personalities associated with the subject? Relationships among people and groups and ideas provide the subject access links. Giving this kind of assistance requires extensive understanding, not just of the material in collections, but also of the environment in which activities documented by a collection took place.

The concept "Provenance" describes an important distinction between libraries and archival or manuscript collections. There are three working definitions of provenance: 1) The entity that created or received the records in the conduct of its business; or the source of personal papers; 2) A record of successive transfers of ownership and custody; 3) the principle that one set of archives must not be intermingled with another, thus maintaining the sanctity of the original order. Information about the provenance of documents is customarily kept along with information about the documents themselves. In other words, the context in which the documents were created and the continuum of functions they represent are important. The documents are evidence of relationships within an ongoing saga.

Intellectual issues

The system we rather naively agreed to develop, now named KLARS (Kellogg Library and Archives Retrieval System), was unlike anything currently available. Existing approaches to intellectual access for original and often unique documents had to be examined carefully. To what extent would they be applicable in the new electronic environment? We wanted our work to fit within the general pattern of current archival practice, and at the same time to exploit the considerable power added by computer and imaging technologies. Several issues had to be explored before decisions could be made.

Providing intellectual access to our material proved difficult. The process of finding relevant material in a collection of personal papers or organizational records is very different from the parallel process in a library. Papers are usually organized in groups rather than as individual items (Correspondence 1950-52; Committee reports; etc.). Subject headings like those attached to books in libraries are rarely applied except in a very general way to large groups of documents. Collections of personal papers and organizational records contain a great variety of material, most of which was not created with use by scholars and historians in mind. Documents are seldom "authored" toward an unseen audience; they are the artifacts of carrying out the business of life.

Making sense of this mass of functional information is a challenge. How do we make information within thousands of unique documents and artifacts available to researchers who are often uncertain of the nature of the documents? How do we help them define and refine their questions so the system can help them find answers? What aspects of the role traditionally played by archivists could fruitfully be taken on by a machine; what aspects would have to remain within the purview of archivists?

We had to find a way to describe documents so that researchers could locate specific information without losing a sense of the context surrounding that information. We decided that describing individual letters and small documents was neither feasible nor especially desirable in our situation. Instead, small groups of documents would be gathered into units for description--small enough for effective word-level querying. Whole collections would also form descriptive units. In addition, we decided to include information about the larger field of adult education by incorporating records about

people, organizations, and major topics. Searches of these additional data would show relationships that could be used to gain access to documents stored in the system.

Another issue was the relationship between our optical system and the work that archivists have done on standardizing a machine-readable (MARC) data format for information exchange. We concluded that each of our collections would be described in the standard way for inclusion in the RLIN database. The RLIN identification number associated with each collection will appear in descriptive records for all groupings associated with that collection in our optical system. The RLIN and KLARS systems are compatible but designed to serve different purposes. Standardized records like those in RLIN show researchers where collections are located, and show the general subject matter and amount of material in them. KLARS, on the other hand, gives researchers access to facsimile copies of the documents themselves and encourages research online. The important thing is to keep a link between the two systems. We will use the RLIN identification number for that purpose.

Vocabulary control also had to be addressed. To what extent should we attempt to control descriptive vocabulary? Rudimentary facilities for checking entries against previously chosen lists of words have been provided in the software. We must still, of course, choose which words to put into the lists. We will not use a full-scale thesaurus to apply controlled subject headings to all records. We are developing preliminary authority information for personal names, form, function, names of places, organizations, major events, and projects.

Technical issues

The Kellogg Project began with a mandate to create a system, based on optical digital WORM (Write Once Read Many) disk storage of images, that would make our collections accessible to a much wider audience. WORM disks are similar to CD's, except that they are not published in multiple copies from masters as CD's are. We can write to these disks ourselves, only once, but then can read the data as often as we wish. WORMs can hold huge amounts of data. We expect to store up to 20,000 page images per side.

Images will be stored on the optical disks as facsimile bit-maps--high resolution grids of dots or cells, each dot representing one "bit" of memory in the system. One bit is an on or off impulse, noted as either a filled or an empty cell. The image is only a map, usually at a resolution of 300 dots per linear inch. The words we humans can read in bit-mapped images of texts appear to the computer only as pictures, as sets of random dots, rather than as machine-readable character codes that can be joined into words.

KLARS incorporates free-text searching, which means that researchers will be able to look for words contained anywhere in the database, as long as they are in machine-readable form. ASCII code, an international standard for digital representation of textual information, is used to represent characters in text. Machine-readable data, in effect, provide an index to themselves automatically.

Descriptive data that we create to describe groups of records will, by definition, be in machine-readable form. However, in order to make full use of available search power, we are going to try translating bit-mapped images of documents into machine-readable text codes using an optical character reader (OCR) device. OCR devices look for patterns in bit-maps, compare the patterns to letter shapes, and then substitute ASCII codes for the characters "recognized." OCR effectiveness depends on the quality of the images presented to it. Our documents are a jumble of clear copies, printed pages, fuzzy carbon copies, telegrams, handwriten notes, poor thermofaxes, etc. Much of our material is going to challenge a top-of-the-line OCR device, and at least 30 percent will be impossible to cope with (handwritten documents, for instance). It remains to be seen how much can be translated successfully, and even how we should define success.

One of the most demanding problems we have faced is deciding how to structure description of the content of our collections. We had to balance giving intellectual access to individual documents against making huge quantities of material available to users. We settled eventually on six basic types of data to capture: collection level data, data within a collection substructure, citations for published material, organizational histories, personal histories, and subject precis. After extended discussion, we decided to combine all six types into one database structure represented in a template of fields (see illustration, p. 86).

A person wishing to look only at one type of data (or a staff member inputting only one type at a time) can choose one view, one subset of the full data structure, to study. Views, however, affect only what is displayed. Regardless of the chosen view, the information from all of the fields will always be available.

Full-text searching with some form of relevance feedback (a tool for redefining a query) was something we looked for from vendors. The system we are installing includes these capabilities. In addition, search results will be presented in order of presumed relevance. Embedded in our system is Personal Librarian, marketed by Personal Library Software, Inc. It is a commercial descendent of the experimental SIRE system and is a product that takes advantage of information retrieval research to enhance the pertinence of query results.

Much thought was put into user interface design. Our system is primarily built for serious researchers. Workstations with large screens, controlled by "mice" as well as by keyboards, work in a windowing environment. Windows can display images, ASCII text and control information, and can be opened, closed, resized, and moved around on the screen. Bit-mapped images can be viewed at normal size or magnified two, three, or sometimes four times (depending on the resolution at which they were originally scanned). We will use varying levels of training to prepare people for use of the system's power. Although some learning will be required for almost all who come to the system, we have attempted to make its use intuitive and enjoyable. In time, a simplified interface may be developed to enable people to look at some of the data over telecommunication networks.

Syracuse University Kellogg Project **Master Template** DRAFT ecoddy 10/22/88	Citation			Org/vita	Person/vita	Subject
	Coll	item	pub			
Record type (Code?)	*	*	*	*	*	*
Plexus ID (auto. assigned)	*	*	*	*	*	*
RLIN ID	*					
* Main Entry ("*"= Authority?)	*		author	name	name	keyword
Title	*		*			
Summary description	*	*	* abstract	* mission function...		* incl. vocab issues
Provenance	*					
Content (OCR'd text)	* inv.	*	*			
* Related subjects	*	*	*	* ?	* ?	*
Location	*	*	*	* ?	* ?	
Folder name (from Inventory)		*				
Inclusive dates	*	*		* begin demise	* birth death	
* Form	*	*	*			
* Function (genre?)		*				
Restrictions	*	* ?				
Citation			*			
Degrees held					*	
* Places assoc with				* ?	*	* ?
* Organiz. assoc with				*	*	* include subsections
* People assoc with	* ?	* ?		*	*	*
* Events assoc with	* ?	* ?		*	*	*
* Projects assoc with	* ?	* ?		*	*	*
Publications				* ?	*	* ?
Organiz. struc/descrip	* ?	* ?		*		
Search strategy hints	*			*	*	*

While development of KLARS has been going on, there has been parallel development of a knowledge base of information about adult education, stored on a VAX mainframe computer. These data can be queried using logic-based software; the data are in machine-readable (ASCII) form. A set of rules for making deductions about relationships between data elements is being devised. We expect to use this resource to gather information about organizations, people, and subjects for the extended parts of our data structure. For the moment, writing logical queries on the VAX is a complex and difficult task, requiring special training. Initial exploration of the knowledge base will be done by members of the project staff. Eventually, the logic programming team (from the School of Computer and Information Science) hopes to design a simpler menu-driven interface that will allow scholars to use the power of logical search directly on the data in the KLARS system.

System development

To create the mandated system, we needed equipment. A request for proposals (RFP) was distributed to potential vendors in January 1987. The group that defined the RFP included representatives from the Syracuse University Library, from the School of Computer and Information Science, and from Academic Computing Services, as well as from Kellogg Project staff. Two months later, we received full proposals from nine vendors, five of whom eventually came to Syracuse to discuss their offerings. A committee of about eight people, half from the project and half from around the University, evaluated the proposals and interviewed the vendors. Each of the final three candidates visited at least twice. In April, after intense discussion, we chose Plexus Computers, Inc. to build our system.

A deciding factor in the choice was Plexus' understanding of the problem we faced. Most vendors were accustomed to providing systems for organizations like insurance companies--data-intensive systems that use prescribed forms and require responses to predictable questions. These vendors were prepared to deliver modified database software, well adapted to such tasks. Plexus, on the other hand, perceived our need for an image system with information retrieval software. The material we proposed to store on optical disk came in all conceivable forms--typed business correspondence, handwritten notes on scrap paper, yellow-lined legal pad, bound account books, poor quality thermofaxes, yellowed and torn newspaper clippings, audiotapes, and color slides. We could not predict the questions that might be addressed to our proposed system. Scholars can come up with an unlimited variety of questions to be answered. Neither could we predict what would constitute a "correct" answer to many queries. In addition to their understanding of the problems facing us, Plexus showed willingness to work with us on design, and then to take the programming responsibility for implementing the design we jointly drafted, all within our fixed equipment budget.

When we began the design process, we had some urgent questions: For whom were we designing this system? Adult educators? Historians? International scholars? Practitioners? Archivists? Whose needs should influence the design? What would be put onto optical disk? How should we choose? Could we include audiotapes, videotapes, slides, photographs? How and where would the system be used? By how

many users? How sophisticated could the system be? How much training time could we expect of users? Should we provide for network access? What sort of user interface would be required? Because there were so many questions, and because answers to these questions were unclear, the design process began in an unsettled atmosphere. It took considerable flexibility and imagination on everyone's part--and a high tolerance for ambiguity--to maintain balance in that uncertain environment.

Under our current contract with Plexus, the optical system is being developed in three phases. For each phase, Kellogg staff and the Plexus design team develop an external specification, showing how the system will appear to its users. Phase I, installed in December 1987, includes most of the hardware we are purchasing, along with rudimentary software. During phase I, we are using the machines, learning effective scanning techniques for our particular documents, and building a body of scanned material.

Phase I hardware includes a UNIX-based Plexus P-95 computer with added storage on 12-inch optical disks, 2 digital scanners, 2 laser printers, and 4 graphics workstations (2 designed for staff use, and 2 for the use of visitors to the collection). All of this equipment is in one large room in the university library, connected by a high-speed communication network called an ethernet. The ethernet is linked to the university-wide network as well.

In Phase II (due to be installed in late fall 1988), the OCR device and a jukebox that can hold and manipulate up to 51 optical disks will be added to the hardware configuration. The more fundamental changes, however, will be found in the software. We will have full use of the Personal Librarian search facility, optical character recognition, the full record structure with all six views of data, and the intuitive interface designed by Plexus.

Hardware configurations for the Syracuse University Kellogg Project's optical disk information system, acquired from Plexus Computers, Inc.

Phase one (installed December 1987)

1 Plexus P-95 DataServer

- 8 megabytes of main memory
- operating system: Unix 5.2 (with BSD 4.2 enhancements)
- uses DataManager (Plexus' version of Turbo Informix)
- supported by 572 megabytes of magnetic disk storage
- and by 1 Optimem optical disk drive
- uses 12-inch Optimem WORM disks, 1 gigabyte per side

4 Plexus XDP WorkStations

- Wyse 286's with 19-inch Sigma Designs high-resolution monitors
- running MS-DOS 3.3 and Microsoft Windows 2.1
- telecommunications link to university mainframe computers
- one floppy disk drive in each machine

2 Fujitsu Digitizing Scanners, model 3094A
- include automatic document feeders (up to 50 pages)
- scan legal-, letter-, A4-, A5-, B4-, B5-sized paper
- 200-, 240-, 300-, or 400-dots-per-inch (dpi) resolution

2 Fujitsu Laser Printers, model M3727
- print 12 pages per minute at 300-dpi resolution

Phase two additions (Installation expected December 1988)

1 Cygnetics Jukebox, model 3070A
- includes two 12-inch Optimem optical disk drives
- stores up to 51 optical platters

1 Calera 9000 OCR Server
- omnifont character reader
- reads typewritten, typeset, laser-printed, dot-matrix styles
- copes with proportional or uniform character spacing

Phase three probable additions (Installation late 1989?)

FAX capability
Machines for digitization and display of color slides
Machines for digitization and playback of audiotapes

nplementation

s design issues were settled, challenges arose around implementation of the new system. A computer system takes up physical space. This may seem obvious, but it is a "little thing," and can be forgotten in the enthusiasm of an enterprise. In our situation, it made sense for the library to house the system. Indeed, to library staff, one of the project's selling points was that in time they would own the system and be able to use it for their own purposes, such as preservation. The library building, though relatively new, has been overcrowded for some years. It was partly because the new Director of Libraries envisioned a major reconfiguration that he was able to find space for the Kellogg system at all. Inserting the system into the ongoing Library environment was almost like creating a new department. Both physical and bureaucratic spaces were needed, implying shifts of people, materials, and facilities. Who would manage the room once it was completed? Who would make decisions about that room beforehand? What would the lines of authority be? The priorities of the library were different from those of the project; keeping communication lines open was not always easy. A new room was eventually created, clearing the way for system implementation.

Before the hardware could be delivered, a representative from the Kellogg Project had to go to California to take part in a test of the system, making sure that it did in fact meet the specifications we had so carefully worked out with Plexus. Bugs identified at that time were fixed before the system was transported to Syracuse. After the hardware was installed, representatives from the design team at Plexus came to Syracuse to do a similar acceptance test. Payment for each phase is completed only after the final acceptance sign-off.

Putting people on the job was another issue altogether. The Project is staffed largely by professionals and graduate assistants. The library is staffed by professionals and by unionized support personnel with carefully graded job positions. One cannot simply seek out an employee because a job has to be done; rather, one has to define the work parameters very specifically--the do's and the don'ts of each task. Union rules about self-nomination, bumping, and seniority must be considered along with applicant qualifications and compatibility.

Even though we were given space intended originally for computer use, preparing it to fit our needs was more complicated than we had expected. We had to ensure clean, painted space with facilities for wiring and cables, adequate lighting, equipment security, furniture for staff and equipment, and proper temperature control. (Computers generate heat. Our basement room temperature soared 20 degrees in as many minutes when the old air-conditioner crashed.)

Operating a system is not just a matter of turning a switch. Our vendor made it very clear (as did subsequent bugs in the software) that we needed a systems administrator to keep the shop running and the sensitive machines happy. We found ourselves training the person we hired to manage the space to manage the system as well.

Two positions for assistants were approved. For these, we had to combine processing and scanning work, two very different activities. Processors are rather high on the union scale, scanners much lower. Negotiating job levels and dealing with budget and pay constraints complicated matters further. In the end, we redefined each position somewhat, and allowed enough supervision by management to meet union requirements. All of this had to be accomplished before work could even begin. It is important to recognize that the set-up time for a major project can take up to one-third of the time the project is allotted.

Complex software such as ours invariably contains a few bugs. The question is not so much "Are there bugs?" as "How well are problems solved by the vendor?". We installed an electronic mail hookup between Syracuse and the California office of Plexus, so that the design team could look directly at system screens to see what was going on on our machines. The disadvantage of maintaining an installation so far away from its developers has been lessened by the use of this communication link.

Future developments

Design of Phase III will begin as soon as Phase II is installed. In Phase III, we will experiment with including digitized audio recordings and color images in our optical system. Additional hardware for digitization and playback of audio signals and for digitization and reproduction of color images will be installed. We will also have to modify our processes and data structures to deal with the new data types.

High-speed, wide-band telecommunication networks capable of transmitting images as well as ASCII text are growing rapidly. A variety of graphics workstations has come into the marketplace. Technology will soon make it possible for complex imaging sys-

tems such as ours to be accessed fully at great distance. To make such access a reality, 1) we would have to prepare software to allow different commercial brands of workstations to link to our system, 2) we would want a simplified interface for distant access, 3) training and help materials that do not assume the presence of library staff would have to be developed, and 4) various copyright and privacy issues would have to be decided. We don't expect current funding to cover provision of network access to images, but it is something on our wish list for future proposals.

In the less remote future, we may be able to make the ASCII portion of our system available to researchers at a distance. Descriptive information exists in ASCII form, as does document content for those items amenable to optical character recognition. Distant users could search the index, then order FAX or xerox copies of chosen pages, as long as no restriction on printing is in effect for those documents. Again, copyright issues affect what can be put into the network and how it might be used. This kind of network access would require a much simpler, less graphic user interface. Our current contract with Plexus, however, does not include development of one.

Larger issues

As it became clear that we could not scan everything in our collections, it also became clear that, although our collections are large, we have neither the mandate nor the facilities to gather everything worth saving in adult and continuing education. Our own priorities must focus on rounding out incomplete collections, processing our backlog, and developing policies that reflect these priorities. For policies to be realistic and effective, and for us to comprehend the scope of the adult education discipline, we felt that a broad cooperative effort with other repositories would be helpful. Cooperative decisions might involve coordinating institutional priorities, identifying collection overlaps, resource sharing at some level, and a dozen ragbag physical, fiscal, and metaphysical issues.

The decision to pursue a documentation strategy for adult education grew out of the idea that the Kellogg Project could realistically mature from a 5-year funded project into a long-term disciplinary history center. Devising a collection strategy is a fundamental building block toward that end. A strategy is really a plan that coordinates the efforts of three types of people: creators, users, and keepers of documents. It is usually multi-institutional and can result in broad sharing of resources, open lines of communication, and a clear picture of who has what, where. In such a cooperative environment, the creative use of an image network--Hiemstra's dream--might substantially increase possibilities for resource sharing.

Last September, we brought together creators of adult education documents, users of such documents (largely historians in the field), and document keepers (usually archivists), to explore the feasibility of developing a documentation strategy for adult education. A major problem was the amorphous nature of the field to be defined. Nonetheless, we agreed that such a strategy was a viable project and we chose which steps to take first.

However, all this talk of cooperation, sharing, and remote electronic access to unique information has raised difficult questions. Copyright is the largest single issue. The law has not kept up with technology, and the whole question of fair-use is not fully understood. To comply with the law as it now stands, we have written to all donors to explain our situation and to request permission to scan the material and make it available through our system. This is only a temporary solution to the problem. And it does not answer questions of invasion of donor privacy or how our need to maintain records on research might affect the researchers. Will the changed environment affect how donors give materials in the future? Will we or they impose more restrictions? Could the integrity of the data on the network be compromised in some way? We do not have answers. These are questions that the entire information community must address. However, we feel that the value of our work is directly proportional to the ethical standards we maintain.

System benefits

In the long run, systems such as the one we are developing have the potential to change the way people approach historical research. Searchers can combine browsing among related documents, looking for relationships across an area of inquiry a field, and conducting searches for specific words in text. Any of these three approaches would be inadequate alone. Interacting, the three substantially increase access power.

Researchers will be able to move among these search modes without leaving the screen. In the past, it was necessary to look for information in guides that simply made reference to large paper collections. A second step, physically locating the documents, was needed before anything could actually be read. In our system (after we have had time to put an interesting amount of data into it), a searcher can look up an organization, person, or subject; can find out which people are associated with it, which organizations, at what times; and can identify which terms have been used to describe activities. He or she can then take those relationships and terms into the search for documents. When an item of interest is located, it can be called immediately to the screen. The researcher might then choose to browse through surrounding documents, giving serendipity a chance to play a role. All of this can be done while sitting in one place.

Researchers will be able to take notes, kept on floppy discs that they bring to and take away from online sessions. Our workstations are based on MS-DOS machines, and have word processing software built in. Inexpensive printers are connected to the workstations for use by researchers, so they will be able to make prints of their notes as they work. (These printers will not be used to make copies of the digitized documents--for this, the laser printers hooked into the main optical system network will be required. Controls over the laser printers will be more stringent than those over the draft printers intended for scholars' notes.)

Notes and a caveat

1. When planning an optical disk project, be sure to allow adequate funds, staff, and time for software development. Purchasing equipment and off-the-shelf software is only a preliminary step--making the pieces work together to accomplish a goal remains

a challenge. The cost of software development can equal or even exceed these initial costs.

2. In an environment that changes as rapidly as the optical disk/computer field, it is important to plan systems that are sufficiently flexible to take advantage of unforeseen changes. One way to do this is to evaluate proposed systems on the extent to which they employ open architecture. How easily could other kinds of hardware be attached? To what extent does the vendor work within guidelines and standards agreed upon in the computer/information industry?

3. Because systems for wide access to original material raise new and thorny issues of privacy and copyright, it is important to design into such systems adequate, but flexible, controls on access and use. Preserving the integrity of data and maintaining some control over its use must be carefully addressed.

4. One thing we have learned down to our pores--allow enough time. Everything takes more time and effort than you think it will. Design is complex, and involves many people. Software development can be depended upon to take longer than you hope it will. Preparing space for and staffing new machines is slow. Setting up the process for analysis and input of data requires much thought, followed by extended staff time. Developing effective training material is another difficult and time-consuming challenge. And when the system finally starts to take shape, be sure to build in sufficient staff time for patron service and show-and-tell! We weren't prepared for the amount of time required for formal and informal presentations, and find that the stream of interested visitors, although important and enjoyable, is occasionally overwhelming.

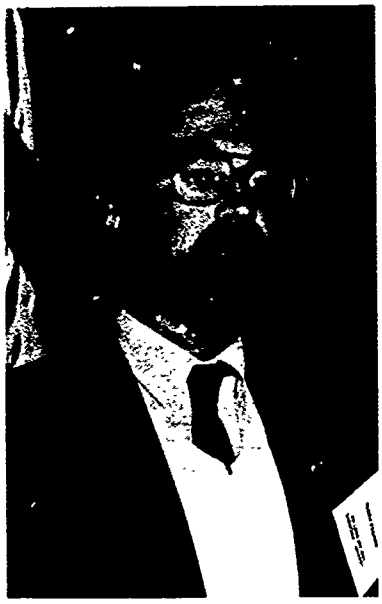
5. The primary objective of the Kellogg Project has been to facilitate communication in a wide variety of ways. Our use of optical disk technology should be viewed in that light. Preservation of documents is not the main focus. For that reason, we are less concerned than many with the longevity of optical disks as media. We plan to build in a replication schedule, and assume that transferring digital information from one medium to another will be feasible at a reasonable cost. In any case, such replication will probably be necessary to conform with evolving hardware, as the machines that read the disks change. Increased availability of adult education documentation while our system functions, however long that may be, remains valuable.

Conclusions

How close are we to Dr. Hiemstra's dream? What are we actually able to do? Well, we have equipment and staff. Our long-range software is being delivered. Processing of the collection itself is underway. Our knowledge of collection contents has grown substantially. Word about the Kellogg Project at Syracuse University is getting around. Our confidence in adult education and its value for society is being borne out. Research in the collections has increased dramatically, leading to at least one prize-winning paper. On the other hand, long distance communication of images over networks seems rather far off. The foundation for Roger's dream is in place, although we have a long way to go before it is complete.

At the risk of taking an image too far, one can say that a ferry is a good, solid, traditional way to cross water. Everyone knows how to do it; everyone knows the schedule. We recognize that our work may show that, in fact, the traditional way is best. But until our experiment and others like it have a chance to prove themselves, we will not know if there are other ways across, ways not dependent on weather or schedules. We have been given an opportunity to see what is possible. We recognize that our work is experimental; a research effort in which errors can be as productive as successes. The ferry may turn out, after all, to be cheaper; but we will soon be able to tell you whether it is better. This presentation has been an attempt to share our experiences, so we can all move into the future with eyes and minds open.

TERRANCE KEENAN



Adult Education Manuscripts
Librarian,
Kellogg Project, Syracuse University

Terrance Keenan is the Adult Education Manuscripts Librarian for the Kellogg Project in the George Arents Research Library at Syracuse University. He received his M.L.S. in Rare Books and Special Collections from Syracuse in 1986, and his B.A. from Hamilton College in 1970. He is a former bookseller and adult education practitioner. He has published extensively in small presses, including three volumes of poetry.

ELIZABETH CARLEY ODDY

Information Transfer Specialist,
Kellogg Project, Syracuse University

In addition to coordinating the development of the optical system, Elizabeth Carley Oddy serves the Kellogg Project as information counselor to an international network of adult educators. She majored in English at Carleton College, completing a B.A. in 1966. At Syracuse University she earned an M.L.S. (1976), was an academic administrator and adjunct instructor for 6 years, and is now working on a dissertation for a Ph.D. in Information Transfer.

Access to Little Magazines: An Index of Optically Scanned Contents Pages

Stephen M. Roberts and Robert J. Bertholf
University Libraries, SUNY, Buffalo, NY

Abstract

The Poetry Collection of the University at Buffalo includes a comprehensive archive of little magazines. These magazines are edited literary publications, printed in limited runs, produced by nonprofessional staffs, and distributed outside the commercial distribution system of major publishers. They are neither comprehensively indexed nor widely collected. However, little magazines contain information which is vital to an understanding of 20th-century literature. The intent of this proposal is to provide detailed and convenient access to the contents of a considerable number of little magazines. By scanning the contents pages of the collection's archive, we intend to generate a cost-effective index to this material. The index will form the basis for an information search and dissemination service that will significantly augment research.

Unlike most units of the University at Buffalo Libraries, the Poetry Collection strives to be comprehensive. The collection's curator is not bound by the same judgmental dictates facing the selectors of the other libraries in the system. The collection development policies of those selectors are of necessity more practical; they are molded by the educational priorities of the institution and by the fluctuations of the State-controlled budget. The Curator of the Poetry Collection, however, carries on a tradition of paying particular attention to poetry, and of enforcing a single policy of acquisitions laid down in very explicit terms by Charles D. Abbott during the 1930s: that is, to collect poetry published in English in the 20th century, along with the working papers, manuscripts, and associated correspondence of the authors collected.

The Poetry Collection began in 1937 as "the Poetry Project," and since then has held firmly to its single-minded collection policy. The initial, narrow restriction of collecting 20th-century poetry in English has produced a research library of great depth. Critical and reference books that support the primary texts have also been added to the collection. At the beginning, Charles D. Abbott wrote to poets, literally asking for the contents of their wastebaskets, and poets responded by mailing in drafts and early versions of poems. Wallace Stevens, for example, sent the holograph manuscript of his poem *The Man With The Blue Guitar.* In the 1950s and 1960s, large manuscript archives came to the Poetry Collection. The first purchase of James Joyce materials was followed by the purchase of the Sylvia Beach collection of Joyce notebooks, books, and correspondence. A huge archive of William Carlos Williams manuscripts took their place next to the major collection of manuscripts of Robert Graves and the notebooks of Dylan Thomas. In recent years, the archive of the Jargon Society (a small but very important publisher), the library and papers of the British poet Basil Bunting, the manuscripts and correspondence of the American poet Robert Duncan, and the papers

of the contemporary Irish poet John Montague have been added to the manuscript holdings.

From the beginning, the Poetry Collection has purchased runs of little magazines. Runs of *Hound & Horn, The Little Review, Poetry: A Magazine of Verse,* and *The Dial,* to name just a few, now provide the foundation for the collection of little magazines at the University at Buffalo. Over the years, about 3,500 runs of little magazines have been acquired. The Poetry Collection presently subscribes to about 1,100 little magazines and spends a good deal of time and effort searching for more new titles. Frederick Hoffman, in his study of little magazines, points out that the golden era of the genre occurred from about 1910 to 1930; throughout this century, however, little magazines have continued to play a central role in defining literary tastes and styles and in leading "the battle for a mature literature" [1].

Little magazines are not commercial publications, and, for the most part, they do not have strong fiscal backing. As Felix Pollak notes in an interview, these publications begin with more enthusiasm than literary sagacity, more raw determination than marketing experience, and more verve than strategic planning [2]. Little magazines usually depend on the financial and literary effort of a few people, especially the editor. They usually last for only a short time--getting five issues published represents an actual success, but as Michael Anania says, getting the third issue out is a frustrating barrier. These publications are not usually sponsored by colleges, universities, or corporations. However, starting in the 1960s and continuing into the 1980s, State arts councils gave grant money to small presses and little magazines, while the Coordinating Council of Little Magazines and the National Endowment for the Arts provided even more money to aid these publications. That so many new titles come out each year is a testimony to the continuing strength and energy of the Nation's writing community; it is also a demonstration of the strong tradition of the little magazine as an important, alternative mode of publishing.

Little magazines offer more than just an alternative place for writers to publish. Traditionally, they have also been places to publish experimental writing that cuts across the accepted patterns of literary taste. It was no mistake, therefore, that James Joyce's *Ulysses,* the major novel of the 20th century, was serialized in the United States in Margaret Anderson's *The Little Review.* Nor was it a mistake that Wyndham Lewis' and Ezra Pound's *Blast,* which appeared in two issues in 1914 and 1915, proposed "Vorticism," a new literary theory, the idea for which was swallowed up in the vortex of World War I. Little magazines are also a place for trial publications, a place for writers to test their new forms and ideas. Although it is usually the case that an editor or group of writers constitute a group in support of the magazine, as was the case with Robert Creeley's editing of *The Black Mountain Review*, it is not unusual for the testing out to lead the contributors to a national reputation. This is certainly the case with Cid Corman's editing of the three series of *Origin,* and the poets it supported: Robert Creeley, Robert Duncan, Charles Olson, Denise Levertov, Paul Blackburn, and William Bronk. Before the days of the endless flow of photocopies, publishing in little magazines was a direct way to communicate with other writers, to reach and cul-

tivate an audience for a particular kind of writing. The example of Walter Arensberg and Alfred Kreymborg's magazine *Others* comes in here as a telling example of a magazine which provided a forum for communication and debate within a group effort to reform the state of American writing. Whatever the regiment and whatever the cause, the little magazine is, unequivocally, the central and most important vehicle for publication in modern literature. That T.S. Eliot's *The Love Song of J. Alfred Prufrock,* and Wallace Stevens' *Sunday Morning*--two of the most important poems of this century--appeared in March and November in *Poetry: A Magazine of Verse* in the same year, 1915, demonstrates that the little magazine is an integral part of, and a generative source for, modern literary history. To ignore little magazines is to negate huge amounts of data about the history of our time.

Other commentators have recognized the importance of the little magazine in literary history, and some have made attempts to provide access to the contents with indexes. Marion Sader's multivolume edition of the *Comprehensive Index to English-Language Little Magazines 1890-1970* was a very ambitious project which supplied access to some magazines [3]. However, Sader chose to index magazines, like the *Kenyon Review,* the *Sewanee Review,* and the *Hudson Review,* which had institutional support and which, therefore, also had longer lives and editorial policies which cultivated acceptance within the traditional literary audience. Scarecrow Press' *Index to American Periodical Verse,* which appeared in annual volumes from 1971 through 1984, tried to list by title and author the publications in many little magazines [4]. Its cessation has left a gap in current research tools. By far the best guide to little magazines is Len Fulton's and Ellen Ferber's *The International Directory of Little Magazines and Small Presses*--now in the 22nd edition with the 23rd in preparation; however, the directory is just that, a directory, a publication which does not give access to the contents of the little magazines. Its companion is the the *MLA Directory of Periodicals [5].*

The project we are proposing, then, developed from an awareness of the lack of access to a massive body of literature and literary history. In another sense, it developed out of the tradition of the Poetry Collection, mainly to collect research materials long before researchers ask for them. We are supplying an answer to a series of complex research questions before the questions have been formulated.

Optical scanning equipment and microcomputer-based indexing software will be used to generate an index to the title pages of one of the largest archives of small press literary magazines in the country. This index will open vast new prospects for research using the Poetry Collection. It will facilitate work with materials that are not routinely collected at other institutions, materials whose worth has yet to be fully realized, let alone tapped. In so doing, the index will not only increase access to materials in the collection, but also will help to justify the continuation of comprehensively purchasing these specialized materials in an era when the materials in other, more practical disciplines are facing budget cuts. We are confident that even the advertisement and demonstration of the index will encourage researchers to delve more deeply into the available resources at the University at Buffalo's Poetry Collection.

We make no claim to be wizards of technology. We marvel at the revolution that is taking place in computing and communications and appreciate what vision and genius have made available. We focus our attention on using that vision and genius to meet the research needs of our own libraries. This project will use only currently available, off-the-shelf hardware and software. The workstations on which the index will be produced fit easily on a desk and can be purchased for a reasonable amount of money. All hardware is available locally and serviced locally. The word processing software that will be employed to edit files is the same package used throughout the Libraries and the University. The indexing package is readily available and was selected on the basis of positive reviews in popular computing magazines. The methodology is straightforward and uncomplicated. Virtually all work will be done by student assistants. C.P. Snow's proposal that advancing technologies produced "two cultures" is radically out of place in the late 20th century, when literary and scientific minds work best in a world of fictions, of "made-up" pragmatic solutions that are then tested in empirical situations. The project clearly demonstrates that computers are not the enemies of literature; these machines have been unjustly accused of dividing culture into camps of incomprehension.

Keep in mind that although we are using technology available to everyone, the end result of this project, the index to the contents of little magazines, will definitely be unique and the information it includes will satisfy a wide range of research questions. It will provide access to information which until now has been virtually unindexed. Further, the technology enabling the production of the index with anything approaching cost efficiency has only recently been realized and never before employed at the University at Buffalo. Ten years ago, the cost of doing this work would have been prohibitive, and it would have taken too long. Five years ago, the cost of the scanning equipment would have been a stumbling block. The Kurzweil 4000, when first available, cost about four times what the Kurzweil Discover costs today. And, although the 4000 is trainable, it is a slow student. We used the Kurzweil 4000 to scan individual volumes of poems, and then prepared the text for *The Collected Poems of George Butterick* by coding irregular spacing, internal margins, italics, bold face, etc., which then could be read by the typesetting program of Printing Prep, a Buffalo typesetting house. The book was produced without rekeying the text, run through a program that we developed with Printing Prep; this program had already been used to produce the type for two complicated bibliographies, *A Descriptive Bibliography of the Private Library of Thomas B. Lockwood* and *Robert Duncan: A Descriptive Bibliography* [6].

The Kurzweil Discover scanner is somewhat less flexible than its older sibling, but it is quicker and incredibly more independent. The new scanner is well suited to this project, mainly because little magazines often produce contents pages with several different type fonts and point sizes of type. The Kurzweil 4000 can be trained to read these smoothly, but given the circumstances that magazines frequently change fonts from issue to issue and that they often cease publication before their third issue, it would have been inefficient to use the 4000 to prepare a training set for each magazine. Still, it was during our experimentation with the Kurzweil 4000 that the idea for this index was hatched. We spent hours feeding samples of poetry and text into the old machine,

wondering at the scanner's ability to generate computer files that we could store and manipulate. We outlined a series of grant-fundable projects that would meld this technology with our desire to expand access to the collection. The release of the Kurzweil Discover prompted us to launch our effort. In our view, that release signaled a state of technological development that made it feasible for practitioners such as ourselves to embrace scanning as a viable tool. It freed us from focusing on the technical process of scanning and allowed us to concentrate on making up procedures through which we could provide access to our unique collection of little magazines.

The basic workstation on which the *Buffalo Index to Little Magazines* will be produced consists of a Zenith 286 PC with hard disk and monochrome monitor, the same machine that is being installed throughout the Libraries as the NOTIS integrated system workstation. Each of our machines is loaded with Enable, an integrated software package that our entire staff has been trained to use. The word processing package can read ASCII text files such as those generated by the Kurzweil Discover, and its powerful windowing capabilities are useful during the editing process. We will use one indexing workstation and two scanning workstations for this project. The indexing workstation will employ a dual, 20-megabyte, removable-cartridge Bernoulli box on which the index will be stored and maintained. The project as outlined will not require extensive storage capacity. Each scanned and edited contents page will require approximately 1.5K; the associated index of each file, about 0.5K. The dual, 20-megabyte Bernoulli box storage device will have more than enough capacity and speed to meet the needs of this project. The two scanning stations will be attached to Kurzweil Discover Model 30 optical scanners.

The Kurzweil Discover Model 30 Scanning System is an automatic intelligent document processor designed for use with IBM XT or AT compatibles. The scanner can operate in either of two modes, a text mode or a graphics mode. In text mode, the Discover employs intelligent character recognition to convert almost any document printed in a nonscript, nonornamental type style between 8 and 24 points into an ASCII text file that it then transfers to a PC where it can be manipulated and, in the case of this project, indexed. The Discover Model 30 has a document feeder and flatbed configuration that facilitates the scanning process and increases efficiency. The Libraries have been familiar with Kurzweil equipment for a number of years and have been consistently impressed with the technology and innovation demonstrated by this small company. We have several Kurzweil readers in our Libraries that are used by visually disabled faculty and students. The Poetry Collection was given a Kurzweil 4000 scanner several years ago when we first started thinking about the possibilities of this project. The Libraries purchased their first Kurzweil Discover machines to participate in a Title II-D (U.S. Department of Education) demonstration grant investigating the viability of using telefacsimile and scanning technologies to foster effective interlibrary loan and cooperative collection development efforts. We have found the Discover machines to be easy to use, dependable, and remarkably effective. Students familiar with PCs can be trained to use them in 1 or 2 hours. Local service is available near campus. Therefore, it seemed logical to build on our positive experience with Kurzweil and use the Discover for this project. The Kurzweil Discover costs approximately $8,000.

ZyINDEX Plus by ZyLab is a program that indexes all major words in a selected body of files and provides a query language to retrieve specific information from those files [7]. The capacity of the Plus version is limited only by the storage limits of the PC on which it is mounted. Its speed is impressive even on standard PCs, and speed is significantly enhanced by the faster processor used in our Zenith 286 machines. In developing the test database for this project, we scanned about 40 contents pages; ZyINDEX Plus was able to read a 1,500-word file and index every significant word in that file in less than 15 seconds. The program was donated to us by ZyLAB when we explained the premise of our experiment. ZyINDEX is a powerful, easy-to-use, program which, once told what to do through a clear set of on-screen prompts, performs flawlessly and accomplishes a tremendous amount of complex work without operator intervention. Again, students can be taught to use the program in one sitting.

ZyINDEX has two main components, an index generator and a search engine; the program is accompanied by a series of useful utilities for updating and maintaining indexes. The index generator creates a list of the unique words in the database being processed. The program maps these words to the database files in which they are found. Finally, it sets pointers to the exact locations within the files where the words appear. The index program includes a user-modifiable "noise word" list containing articles, prepositions, pronouns, and common abbreviations. Words can be added to or deleted from the list using any word processor. The search engine is powerful and convenient. Queries consisting of content words, Boolean connectors that relate to or help to limit the content words, and wild card symbols which take the place of parts of content words and which allow retrieval of variations on a given word are constructed and run against the index list to isolate relevant citations for examination. Retrieved files can be viewed on screen, marked, moved, saved to disk, or printed.

The methodology for this project was developed during the process of scanning and then indexing a trial group of little magazine contents pages. The test database consists of the contents pages of all issues of 10 little magazines. During the actual project we will scan and index the contents pages of 3,500 titles over the course of 2 years. Other than selecting the titles to be included and monitoring the process, two jobs that will remain our responsibility, virtually all of the work will be accomplished by part-time students. The following is a step-by-step description of what we plan to do.

The curator will select titles for inclusion in the index based on four criteria: first, whether the editor or the writers for a particular magazine became known to a national audience; second, whether the writers continued to write for little magazines and established a provocative position within the world of the little magazine; third, whether the magazine was a part of a literary movement which in some way influenced the scope and progress of 20th-century literary history; and fourth, whether the Poetry Collection owns a complete run of the magazines which fit one or more of the first three criteria. Those titles that comply will be tagged, and the contents pages of all issues will be photocopied.

While we were experimenting with ZyINDEX, we decided that the value of the final indexed product would be significantly enhanced if we invested some time in editing the raw scanned files.

When a search is performed using ZyINDEX, the program shows a screen listing the names of the files in which hits occur and displaying the first 50 characters of each of those files. We decided that this set of 50 characters should, to the extent possible, show a standard bibliographic citation to the title and issue represented in the file. The student assigned to the project, therefore, will mark the photocopy of each issue with the citation which will be inserted into the first line of the file after it has been scanned. That citation will appear in the following format:

TITLE, VOLUME NUMBER.ISSUE NUMBER (DATE) (TYPE)

We elected to designate magazine TYPE to provide a delimiting context for narrowing searches:

(I) American Titles 1900-1945

(II) American Titles 1945-Present

(III) British Titles; this category includes all publications from the British Isles 1900-1945

(IV) British Titles; this category includes all publications from the British Isles 1945-Present

(V) Other; this category includes publications from Canada, New Zealand, Australia, South Africa, and elsewhere

An example of a formatted first line would be:

Temblor, 2.1 (March 1986) (II)

We also decided that searches would be greatly enhanced if we could indicate the role of all individuals cited on the contents pages. Working from conventions we established, students will also mark all names on the photocopied contents pages according to their roles as poet; author of article, letter, photograph, play, review, or story; editor; illustrator; composer; interviewer or interviewee; translator; recipient of letter; or subject of article, photograph, or review. A one- or two-letter code in parentheses will be written onto the photocopy after each name.

Specific name c des:

(P) -- Poet
(E) -- Editor
(AA) -- Article/Author
(DA) -- Drama/Author
(RA) -- Review/Author
(SA) -- Story/Author
(PA) -- Photograph/Author
(LF) -- Letter/From
(LT) -- Letter/To
(PS) -- Photograph/Subject
(AS) -- Article/Subject
(RS) -- Review/Subject
(I) -- Illustrator
(IQ) -- Interviewer
(IA) -- Interviewee
(T) -- Translator
(C) -- Composer

The little magazines with marked photocopies of contents pages will then proceed to the two scanning workstations where the contents pages will be scanned and the output from each page will be saved as a separate ASCII file. An alphanumeric file-naming procedure has been established to facilitate identification and database maintenance. An eight-character file name will be assigned to each file as the scanned data are saved. The name will have the following format.

> TTTVVVII where TTT is three significant letters from the title of the magazine, where VVV is the volume number of the scanned magazine, and where II is the issue number of the scanned magazine.

Each file will be edited using the Enable word processor. When necessary, the actual titles of the poems, articles, etc., appearing in the issues will be inserted into the file. (The convention followed on the title pages of many magazines is to give only generic indications of an author's contribution, to print, for example, *Two Poems* on the contents page rather than the actual titles of the two poems.) The edited files will then be saved to disk as text files.

The edited text files from the two scanning stations will regularly be added to the index database on the third computer.

Using ZyINDEX Plus, newly added files will be indexed, thereby updating the index database. A backup copy of the updated database will be created on one or more Bernoulli cartridges.

We will use the Enable database to update a comprehensive file of the titles and runs within those titles that have been scanned and indexed. The database will reflect the content of the RLIN serial records that have been created for each of the little magazines appearing in the index. The Name Authority File, developed through the cataloging of books and manuscripts in the Poetry Collection for the past 6 years, as well as the in-house Name Authority File, will be used to verify names in the search procedures.

Once it has been created, the existence of the index will be well publicized. At least one article regarding the project will appear in print before the index is ready to use. The Poetry Collection will produce a brochure and mail copies to the main research libraries in the United States and in Canada. The brochure will list the titles and issues of the little magazines that have been included. Advertisements will be placed in the leading library and research journals covering 20th-century literature, and every effort will be made to inform scholars of 20th-century literature that the index is available for their use.

At least initially, contact between researchers and the Poetry Collection will take place by phone or mail. Eventually, we will set up a BITNET address to which queries can be sent via electronic mail. Incoming requests for information from the index will be analyzed and a search strategy will be developed. After consulting the Poetry Collection's name authority files, the search will be run against the index. Searching with ZyINDEX is very similar to using most bibliographic search services. Single- or multiple-word queries can be entered. Names, magazine or poem titles, individual words or phrases from those titles, and dates can be searched. Boolean operators such as "and," "or," and "not" can be used to link topics of interest. Proximity indicators, truncation symbols, and wild card characters can be used. The editing codes that we inserted next to each name in the contents pages can be entered to limit a search, as can the codes we placed in the titles. The only feature not present that would be useful is the ability to use "equal to," "greater than," "less than," etc., in conjunction with dates; it is not possible, for example, to search for all occurrences of a name since 1950 with a single query. ZyINDEX is fast and logical to use. As stated earlier, when the query is entered, the program responds with a list of hits. By moving the cursor to the desired line, the actual file can be called up on screen and then used or rejected.

Depending on the needs of the researcher, citations or copies of appropriate materials will be forwarded once located using the index. Photocopy and telefacsimile service will be offered.

We anticipate that some researchers may find it useful to have a copy of the text files from which the index is generated. We expect that the collected files will be of a size that is quite manageable as well as transportable; many scholars may want only a specific subset of the full database and such subsets would be easy to create. We intend to offer subscriptions to the scanned text files.

In the volume, *Return to "Pagany": The History, Correspondence, and Selections from a Little Magazine,* the poet William Carlos Williams wrote to Richard Johns, who proposed a new journal Pagany:

"Thus I take you seriously but so much do I pant for what you offer that I am doubly dubious of anyone's ability to make good. Yet what you envision IS the future -or rather the present if anyone could have the ability to put it across. I'd back it and it would be the center of every literary interest after a patient murderous ascent extending over several years of effort. I'd expect to give it my life--in short"[8].

Not only are little magazines the starting place and the place of demonstration for literature in this century, but they are, in all their idiosyncratic ways and unregulated impulses toward the new in expression, the principal repository of modern literary history. They nurture new writing and new writers, like William Carlos Williams, nurture them by continuing to accept poems and offer financial support. It does little toward the understanding of literature history to announce that most little magazines are not worth the effort of a reading. The record of literary accomplishment and patterns of literary tastes must take account of all levels and varieties of writing to found a comprehensive view of literary history. Just as the study of Denham and Waller reveals the maturation of the couplet and intricate metrical systems in a more obvious way than the refinements of the same forms in the work of John Dryden, and just as a study of the work of Al Held reveals the vocabulary and gestures of abstract expressionism more obviously than the perfections of Willem DeKooning, the examination of the writing in little magazines illuminates the prevailing modes of expression and charts alternative and equally valid writing and theories of writing. Little magazines complete the literary record.

The procedures of scanning, editing, and indexing we have outlined here activate a massive body of orphan publications that are now dormant in specialized libraries. Leafing through runs of little magazines looking for references will soon be a procedure of the past. Further, the methodology of the *Index to Little Magazines* discussed here can be transferred easily to similar bodies of dormant information. Collections of materials such as these are indeed highly specialized, but our procedures for providing access to them are not. Without divisions, and without a rent in the social fabric, technology and the world of poetry can be combined.

References

[1] Frederick J. Hoffman, Charles Allen, and Carolyn F. Ulrich, *The Little Magazine: A History and Bibliography.* Princeton, NJ: Princeton University Press, 1947, 1-17; esp. 1.

[2] "Felix Pollak: An Interview on Little Magazines," with Mark Olson, John Hudson, and Richard Boudreau, *Triquarterly,* No. 43 (Fall 1978): 34-49.

[3] Marion Sader, ed. *Comprehensive Index to English-Language Little Magazines, 1890-1970.* Millwood, NY: Kraus-Thomson Organization, Limited, 1976.

[4] *Index of American Periodical Verse: 1984,* Sander W. Zulauf, et al. 1971-80; Rafael Catala et al. 1981-84. Metuchen, NJ: The Scarecrow Press, Inc., 1973-86. See also Stephen H. Goode, *Index to Little Magazines.* Four volumes. New York: Johnson Reprint, 1967; [*Index to Little Magazines,* then published annually.] Jefferson D. Caskey, *Index to Poetry in Popular Periodicals 1955-1959.* Westport, CT: Greenwood Press, 1984.

[5] Len Fulton and Ellen Ferber, eds. *The International Directory of Little Magazines and Small Presses: 22nd Edition 1986-87.* Paradise, CA: Dustbooks, 1986; Eileen M. Mackesy et al., *MLA Directory of Periodicals: A Guide to Journals and Series in Language and Literatures: 1984-85 Edition.* New York: The Modern Language Association of America, 1984.

[6] Robert J. Bertholf, *A Descriptive Catalog of the Private Library of Thomas Lockwood.* Buffalo, NY: University Libraries, SUNY at Buffalo, 1983; Robert J. Bertholf, *Robert Duncan: A Descriptive Bibliography.* Santa Rosa, CA: Black Sparrow Press, 1986. The University of Chicago Press has published a revised edition of its *Chicago Guide to Preparing Electronic Manuscripts* (Chicago: University of Chicago Press, 1987), but the guide is too complicated, contains too many commands to be efficient. This guide was rejected in favor of the program developed by Printing Prep and the Poetry Collection.

[7] Ernest Perez, "ZyIndex: Text Retrieval Program Offers Blazing Speed with Ease of Use," *Info World,* 9.45 (9 November 1987): 46-48.

[8] Stephen Halpert and Richard Johns, eds. *A Return to Pagany: The History, Correspondence, and Selections from A Little Magazine, 1929-1932.* Boston: Beacon Press, 1969: p. 4.

STEPHEN M. ROBERTS

Associate Director, University Libraries,
State University of New York, Buffalo

Stephen M. Roberts holds a B.A. from the Johns Hopkins University, an M.A.T. from Brown University, and an M.L.S from the State University of New York at Buffalo (SUNY/UB). Roberts joined the SUNY/UB Libraries director's office staff in 1977 to help coordinate the move to the University's new campus in Amherst, New York. He was promoted to Assistant Director for Systems in 1982 and supervised the purchase, installation, and operation of a GEAC circulation system. Since 1986, as Associate Director, Roberts has inititated the implementation of an integrated library system using NOTIS software and has served as Acting Director of the Lockwood Memorial Library, the graduate social sciences and humanities library at SUNY/UB.

ROBERT J. BERTHOLF

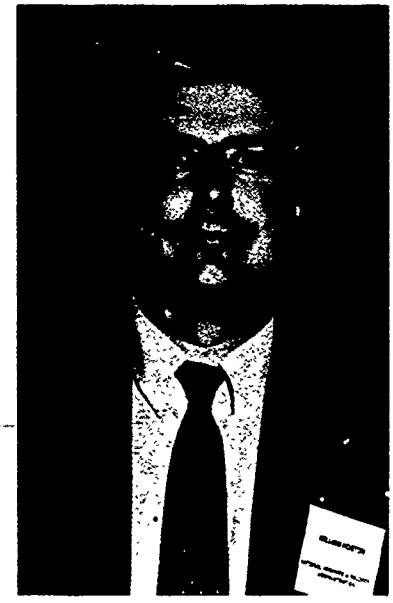
Curator, Poetry/Rare Books
Collection,
State University of New York, Buffalo

Robert J. Bertholf holds an undergraduate degree from Bowdoin College and graduate degrees from the University of Oregon. In 1979, he left a career in teaching to become the Curator of the Poetry/Rare Books Collection, State University of New York at Buffalo. He is the author of articles and reviews on modern poetry, as well as the author of "A Descriptive Catalog of the Private Library of Thomas B. Lockwood" (1983) and "Robert Duncan: A Descriptive Bibliography" (1986). He is now editing a multivolume edition of the collected works of Robert Duncan, and working on digital and computer applications to the management of special collections, and to the process of editing modern texts.

SESSION 4

WILLIAM L. HOOTON,
MODERATOR

Director of the Optical Digital Image Storage System,
National Archives and Records Administration



William L. Hooton's responsibilities include the chairmanship of the Digital Image Applications Group and membership on the Steering Committee of the Federal Council on Computer Storage Standards and Technology as well as memberships in the American National Standards Committee on Optical Digital Data Disk-X3B11, the Association for Information and Image Management (AIIM) Publications Committee, the AIIM Standards Committee on Electronic Imaging, and the Recognition Technologies User Association. He is responsible for research and system development involving digital image, OCR, and optical disk technologies at the National Archives.

Mr. Hooton has been with the Federal Government since 1970 and has been involved in digital image and optical disk technologies since 1978. He authored the concept and original design for the FAISR optical disk system at the IRS before joining the National Archives in 1983. Over the past 10 years, Bill has frequently spoken at national and international conferences and seminars on the subject of digital image and optical disk technologies. He is frequently interviewed for national publications and has published several papers and articles on the subject.

Experience with an Optical Disk System at FDA

Kenneth Krell
Center For Devices and Radiological Health, Food and Drug Administration, Department of Health and Human Services

Abstract

A FileNet optical storage and retrieval system was leased by FDA in October 1986 to evaluate the applicability of this new technology to the Agency's paperwork management problems. A 1-year evaluation period was undertaken by a staff of 3 members of the Center for Devices and Radiological Health (CDRH), one of the 7 major components of the FDA. The evaluation plan called for demonstration of the system's capabilities to staff members of the various FDA components, development of operational and administrative pilot applications, and publication of a final evaluation report. The system became operational in January 1987 and, during that year, many demonstrations were given to FDA staff as well as to persons in the public and private sectors. Several pilot applications were developed during that period. The main conclusions stated in the evaluation report were that optical storage technology could solve many of the paperwork management problems of the FDA, that there was no need for a central or standard system throughout the Agency, and that each FDA component should evaluate the technology as a potential solution for its problems.

Today I am going to describe how FDA got involved in optical disk technology, how we evaluated its potential effectiveness in paperwork management, what is presently being done, and what is planned for the future.

Initial Agency Evaluation

Background: As early as 1984, Stuart Carlow and Neil Goldstein of the Center for Devices and Radiological Health (CDRH) and others within FDA were keeping track of a newly developing information storage technology, high density information storage on optical disks. As with other new technological advances, the perspective of the CDRH group was:

TECHNOLOGY LOOKING FOR PROBLEMS TO SOLVE

vs.

PROBLEMS LOOKING FOR A TECHNOLOGICAL SOLUTION.

Decision to evaluate - By 1986, we at CDRH decided that optical disk storage and retrieval technology had reached a stage at which evaluation was practical. The question was whether to avoid the time and effort spent trying to work with incompletely developed systems by waiting until the systems were better defined, or to jump in im-

mediately, hoping to significantly influence the course of development. We chose to do the latter.

FDA decided that CDRH should lead the evaluation effort. Procurement was initiated with a Request for Proposal (RFP) for an off-the-shelf system. There was only one respondent, FileNet.

Description of the evaluation plan - The evaluation plan was to acquire hardware and software on lease. After acquisition and familiarization, we would, for 1 year, demonstrate the system to the various agency components and, together with them, develop pilot applications. At the end of the year, we would submit a final report based on our experience.

Progress of the evaluation

Hardware and Software - In the same year (1986), FDA awarded a 4-year, $520,000 contract to FileNet Corp., of Costa Mesa, California, for an optical disk system. The contract was a lease-to-purchase agreement which left FDA with the option to cancel if the evaluation results were negative.

Under the contract, FileNet provided a 64-disk "jukebox," a 400-dots-per-inch laser printer, an image workstation, and a combination image/document entry workstation. An additional image/document entry workstation equipped with a different scanner was added later. FileNet supplied the basic software for document entry (scanning), indexing, and retrieval, and the following additional software: a document editor, forms generator, VT100 terminal emulator, and WorkFlo, a high-level application development language.

Applications - At FDA the problems weren't hard to find. As a regulatory agency, FDA receives a constant stream of documents, some small and others as large as 80,000 pages. These documents relate to the marketing of foods, drugs, and medical and electronic-radiation-emitting devices. The control, tracking, and review of these documents was always a formidable task, and may soon become unmanageable.

SIREN - The Center for Food Safety and Applied Nutrition (CFSAN) has assembled over the past 11 years a reference collection of documents, scientific papers, policy statements, precedent correspondence files, food additive petitions, etc. The collection numbers over 10 million pages, each of which has been indexed in a large database system (Model 204) running on an IBM 3090 mainframe computer. The reference system is known as SIREN (Scientific Information Retrieval and Exchange Network).

The original documents in both hard copy and microfiche form are maintained in a central repository. A user must, after searching the database, copy the reference number and then request the hardcopy or microfiche. On many occasions the document is already signed out, may be lost, or is incomplete. If the user is physically distant from the repository, the document/fiche must be delivered.

We developed a pilot application for SIREN as follows:
A number of documents were scanned onto optical disk in the FileNet system and were assigned a unique index value identical to the key index value of that document in the Model 204 database record (sirenid). As the documents were entered, a queue was built containing pairs of FileNet document identification numbers (docid) and sirenids. After input of a batch of documents, communication with the Model 204 database was established and as each Model 204 record was retrieved using the sirenid, it was updated with the docid.

Using the WorkFlo application development language, we created a search and display-or-print procedure. The screen at an image workstation displays a window emulating a VT100 terminal and a window for display of document images. The user establishes communication with the Model 204 database through the VT100 window and searches for desired documents. In order to display a document when a desired reference (record) is retrieved, the user presses the "d" key. The docid from the record is transmitted from the Model 204 database to the FileNet system and the first page of the document is displayed in the image window. The user can then page back and forth through the document or display a specific page. If the user is at a nonimage terminal (e.g., a VT100), a hard copy printout can be requested. The request is electronically transmitted to the FileNet system; the document is printed on a laser printer together with a routing slip, and is delivered by messenger (usually within 1 day).

FOI - The FDA receives Freedom of Information Act (FOI) requests numbering in the tens of thousands each year. The CDRH alone receives over 20,000 requests per year. Any confidential information must first be removed from the copy of any document requested under FOI. At CDRH, the most commonly requested document type is stored as microfiche. The FOI clerk prints out any pages requiring redaction. The confidential information is obliterated and a photocopy of the page is made. On a copy of the microfiche, the images corresponding to the redacted pages are excised with a sharp blade. The altered microfiche copy and the photocopied pages are then sent to the requester. A copy is retained to respond to any subsequent requests.

The FOI pilot application involves scanning the requested documents onto optical disk. Although microfiche scanners exist, hard copy printouts from microfiche were scanned onto optical disk during the pilot test. The image editor subsystem was then used to electronically redact the document. The redacted copy was then stored on optical disk and a copy was sent to the FOI requester. It should be noted that the original document images on optical disk are not altered.

Demonstrations - The FileNet system, its software subsystems, and the pilot applications were demonstrated to numerous groups and individuals throughout 1987. As a result, much was learned about the capabilities of the system and about numerous paper handling problems which exist in a variety of situations.

Evaluation Report - The evaluation period covered the 1987 calendar year and at the end of that period an evaluation report was prepared and submitted to the Office of

the Commissioner of FDA. The report concluded that optical disk storage and retrieval systems could significantly alleviate many of the paperflow problems of the Agency. In addition, there seems to be no need for a single central system, nor for one single type of system since there is relatively little document interchange among the seven FDA centers.

The SIREN pilot application was very successful. The remote host computer and database were easily accessed and document images and/or printouts were readily produced. The report recommended that CFSAN plan for acquisition of a suitable system.

The FOI pilot application quickly proved that FOI requests could be efficiently and rapidly handled with the combination of optical disk storage and retrieval and electronic document editing.

The report recommended that the Center for Drug Evaluation and Research begin to evaluate the possibility of reviewing new drug applications and managing the adverse drug reaction collection.

Finally, it recommended that CDRH procure a system in order to better manage many of its paperflow problems.

CDRH Implementation

The implementation plan takes into account the relocation of the Office of Device Evaluation (ODE) and Office of Compliance (OC), about 400 employees, from Silver Spring, Maryland to a new building in Rockville, Maryland and assumes that the move will be completed by July, 1989.

Step 1: Installation of a data and image-carrying ethernet network throughout the Rockville building to be completed by the time ODE and OC have occupied the building. Also, during this period, an image-carrying link will be established between the Rockville building, where the central optical disk facility (jukebox and image management system) will be located, and the other CDRH buildings.

Step 2: Concurrently with Step 1, PC-type image workstations will be procured and existing PCs will be modified to function as image workstations by the addition of high resolution monitors, ethernet and compression/decompression boards, and additional memory. Depending on the type of central optical disk facility utilized, either scanners and printers or scanning and printing workstations also will be procured during this step.

Step 3: (carried out concurrently with Steps 1 and 2) Installation in the Rockville building of a central optical disk facility. This location will permit the majority of image traffic to be carried on the local ethernet rather than over the remote link. A sufficient number of image workstations, scanners, and printers will be provided at those locations where high-priority documents arrive, are generated, or are processed.

Step 4: Input of high-priority documents depends on the type of central optical disk facility utilized. If the pilot FileNet system is utilized, document input can begin at once. However if a different system is procured, document input most likely will not begin until the equipment is up and running. This is because there is an absence at this time of a standard format for image data on 12-inch optical disks--disks are not interchangeable between systems of different manufacture.

Step 5: (ongoing) Definition and development of various applications utilizing optical disk image storage and retrieval. Applications already under development or completed include:

1) scanning and indexing documents with a number of checks on the index information and integrity of the documents,

2) development of the Summary Database (a secondary database containing summary information derived from various primary databases) with access to document images,

3) the ability to obtain paper copies of referenced documents when accessing the secondary database from a nonimage workstation,

4) electronic editing of documents for response to FOI requests, and

5) an electronic "paperclip" feature with which a reviewer can easily build a dynamic personalized index for a document.

Step 6: Bringing the system and applications into production mode and training users.

The Future

For CDRH, the future consists of planning for growth of the system until all those with need can access the system from their own desktop image workstation. For FDA, the future probably will hold a satellite-based system that reaches district offices throughout the country. For both, the application of character recognition will perhaps allow automated indexing and/or full-text search.

KENNETH KRELL

Associate Director, Office of
Information Systems,
CDRH, FDA, Department of Health
and Human Services

Kenneth Krell received his B.A. in History from the University of Pittsburgh in 1954 and performed his graduate studies at Columbia University, College of Physicians and Surgeons, where he received his M.S. in Human Nutrition in 1967 and Ph.D. in Biochemistry in 1970. From 1970 to 1974 he conducted a postdoctoral study at NIH on the topic of "Molecular genetics and biochemical studies of phage lambda and its host." In 1974, he joined the Food and Drug Administration as a Commissioned Officer in the U.S. Public Health Service. He initially carried out research on radiation effects in mammalian cells and currently has over 40 research papers published. He then moved to research management and then to information management and has been working with optical storage technology for the past 2 1/2 years.

Desktop-Digitization: Prospects and Problems

Bradford S. Miller
Thousand Oaks Library

Abstract

"Desktop digitization" is a term used to denote small-scale digitization ventures using off-the-shelf microcomputer components in on-dedicated configurations. A "minimalist" experimental system assembled at the UCLA Graduate School of Library and Information Science during 1987-88 is described and discussed. The discussion highlights the technical feasibility of desktop digitization and indicates important operational efficiencies and inefficiencies in the UCLA/GSLIS system, particularly regarding the use of VHS videotapes as a mass storage system.

Desktop Digitization: Prospects and Problems

The digital approach to preservation and access issues is no longer a novelty in the library world. By the late 1980's, digital imaging strategies and system designs have already received considerable professional and scholarly attention. We are moving rapidly toward system implementation in institutions where the advantages of the digital format for preserving and managing large quantities of information have been recognized and appreciated. Operationally, the approaches taken to implementation have clustered around resource-intensive strategies involving mass conversion of materials via bit-map production and optical character recognition that compete with large-sca deacidification and micrographics projects. Institutionally, these strategies are characteristic of large and relatively well-endowed research institutions and archival collections with regional or national service responsibilities. The smaller library or archive with few discretionary resources and a circumscribed service area is not yet seen as a player in the digital arena; moreover, there is, perhaps, an unspoken assumption that the only realistic role for such institutions is that of customer for packaged digital materials. Alternatively, a small library could be a minor member of a consortium, feeding candidate materials into the production/conversion system.

This assumption, if it is indeed operative in the profession, does not take full cognizance of three interlocking tendencies in the information field: 1) the spread of increasingly powerful and cost-effective microcomputer-based, digital image-processing technologies designed to support desktop publishing and other graphics-oriented applications; 2) the spread of "digital awareness" through familiarity with new microcomputer interfaces, sophisticated word processing and spreadsheet applications, and desktop publishing; and 3) the will to autonomy--particularly with reference to locally significant materials--that is a part of the makeup of many librarians and archivists. Together, these tendencies could lead to different kinds of applications for scanning methodologies in libraries: the locally produced, processed, and maintained digital archive supported by a microcomputer that is not necessarily dedicated to digital scanning but that may be used to perform other administrative tasks, or an OCR text con-

version service directed toward library users undaunted by 5-percent error rates. Hitching a ride on the popularity of the term "desktop publishing," I have dubbed this kind of application as "desktop digitization."

Technologically, some of the same economies of scale that have made dedicated digitization and data conversion systems attractive to research libraries and large archival institutions also tend to make it feasible for the smaller library to produce and access an archive of digitized materials using commercially available and moderately priced microcomputer hardware and software that were not designed primarily for that purpose. A hand-held, 200-lpi scanner designed to support desktop publishing applications can now be purchased at the local shopping mall for $200, while graphics-capable laser printers are beginning to crack the $1,000 barrier. Optical character recognition capabilities--admittedly of limited scope and reliability--are available as options and upgrades for personal computers and scanners. In the late 1980's, the microcomputer is a multi-purpose computing device of considerable power and efficiency. When coupled to a desktop scanner and laser printer, there is nothing to prevent a microcomputer, otherwise used to analyze circulation statistics or produce the institutional newsletter, from being used to preserve an archive of newspaper clippings or to generate access copies for a set of delicate photographs or to offer the undergraduate the opportunity to convert his typescript into ASCII code.

Ten years ago, no microcomputer had been marketed that could support digitization applications, and scanners and laser printers were customized devices married to mainframe computers or dedicated hardware such as newspaper page-production systems. Now, a microcomputer system that matches most of the capabilities of a 1970's mainframe-based digitization system can be purchased off the shelf for less than the cost of the feasibility study that would have been necessary in 1978 before installing such a system. Acquisition of microcomputer technology in libraries has not quite reached the "petty cash" stage of things, but it has reached the point at which public libraries purchase Macintoshes or PC-AT's in lots, like typewriters. Indeed, a corporate librarian might find one placed on his or her desk without having asked for it. In such an environment, the financial and operational barriers to desktop digitization may be considerably lower than those raised by in-house microform production. The current vogue in graphic interfaces and graphics-oriented application software to run on microcomputer platforms can only contribute to the acclimatization of the user to the pixel, the vector, and the bit map.

The point of this speculation is that bright prospects for desktop digitization may mean that we will have to broaden our view of the institutional scope of scanning methodologies in the library. We may find that we are dealing with a medium that cannot be contained within the organizational mold of the research library consortium or the technical mold of the dedicated turnkey system. If the marginal costs of adding scanning, output, and mass storage capabilities to a microcomputer platform continue to drop as they have in the past decade, the future of the digital image in the library could be a tiered structure of high-volume, capital-intensive optical disk production ventures at the top; large-scale cooperative archiving and data conversion projects in the mid-

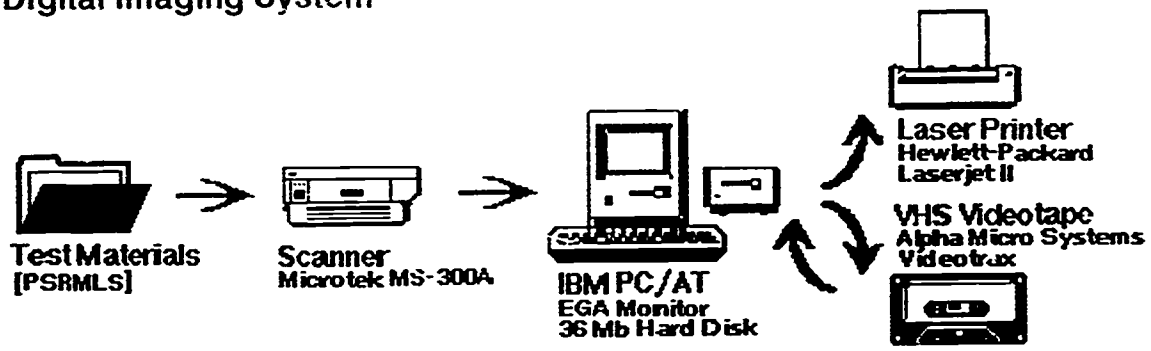
dle; and an anarchic mass of localized, small-scale archives and service offerings at the bottom. On the other hand, it may be the case that the data processing and mass storage requirements necessary to effectively support the production and maintenance of even a modest archive of digital images or a minimally useful conversion service may be beyond the capabilities of microcomputer systems not specifically configured to meet them. It may also be the case that skills obtained from experience with other microcomputer applications may not readily transfer to a facility with desktop digitization and that, consequently, training and familiarization costs will be too high to make the small-scale operations an attractive prospect. If so, the institutional base for digital applications may be limited to organizations capable of making a substantial commitment to the approach and its technology. Smaller institutions would then be largely passive participants in a system completely dominated by commercial vendors and large libraries and archives. Such a future would be more orderly, but perhaps less rich in possibilities than the anarchic alternative.

The promise of scanning methodologies as a solution to the preservation crisis and as an attractive access medium has been the motivation for large-scale experiments and pilot programs initiated by several of the institutions represented at this conference. The prospects and problems of desktop digitization in the library have not received the same level of attention. In fact, the microcomputer-based system assembled at the UCLA Graduate School of Library and Information Science (UCLA/GSLIS) in 1987-88 under the direction of Dean Robert M. Hayes to support a study for the Council on Library Resources may be one of the first experiments in the development of a small-scale digital imaging system from commercially available components not especially configured for the application. The purpose of the study was to identify the operational characteristics of such a system and to evaluate its feasibility as an approach to preservation and access enhancement. Specifically, this entailed intensive examination of digital scanning techniques, on-line and off-line storage of image files, file access and retrieval procedures, and hardcopy output processes. Examination of optical character recognition and hypertext capabilities was specifically excluded from the study. Both are information-processing techniques of great potential value to the library profession, but they are applications and manifestations of images already digitized. It is sometimes difficult to remember that OCR and hypertext are both possible "fates" for a digitized image, but not necessary outcomes of the scanning process.

Experimental System

The UCLA/GSLIS image digitization system was assembled entirely from commercially available hardware and software components and integrated without customized connections or enhancements. Some items had already been acquired before the study began; other components were purchased with the experiment in mind. Although there was no "minimalist" predisposition in the selection of components, the configuration that emerged was clearly at the lower end of the technological and financial spectrum; that is, for any particular component used in the UCLA/GSLIS system, a more sophisticated, powerful, and expensive alternative that could improve system performance is on the market. The approach, then, was minimalist in effect. Total system cost at current prices is approximately $10,000. During the course of the study, the system was

UCLA/GSLIS Microcomputer-based Digital Imaging System



Software: MS-DOS 3.1 [Microsoft], Eyestar [Microtek], Videotrax [Alpha Micro Systems], dBase III [Ashton-Tate].

Aggregate System Price: Approximately $10,000

Imaging Capabilities: Maximum image size: 8.5" x 11"
Maxin um Resolution: 300 lines per inch
Imaging options: Black and white text and line art
34 grey tones via dithering options
No color processing
Software controlled contrast settings
Limited WYSIWYG and editing

Production Dynamics: Sheet feed operation [maximum thickness .012"]
2.5 minute average for 8.5" x 11" image scan to disk

Storage Capacities: 225 Kb average file size for images in condensed format
Recommended VHS videotape capacity = 96 Mb
[Because it does not support multiple backups, Videotrax effectively limits tape capacity to the capacity of a single disk backup; in this case, approximately 30 Mb.]
130+ 8.5" x 11" image files per tape
Archive + working tape backup time = 90 minutes

Retrieval Throughput: Approximately 10 output pages per hour

Reproduction Capacities: Archive videotapes of digital data can be commercially reproduced in mass quantities with acceptable fidelity

Transmission Capabilities: Image files can be transmitted via telephone

available for use by UCLA/GSLIS faculty and staff and was used for various applications not connected with the study or with digital imaging.

The experimental system was built around an IBM PC-AT equipped with an EGA (Enhanced Graphics Adapter) card and medium-resolution monitor. In conjunction with image production software, the monitor provided WYSIWYG ("What you see is what you get") views of image areas equivalent to about 1 square inch of the bit map as well as stepped-down views of the complete image. On-line storage was provided initially by a standard 20-Mb hard disk, later upgraded to a 72-Mb disk partitioned into two 36-Mb drive.

Digitization capacity was provided by a Microtek MS-300A scanner controlled by a proprietary software package called Eyestar. This unit was capable of converting hardcopy materials to digital bit maps with a maximum resolution of 300 lines per inch in either "line art" mode (black and white) or "halftone" mode (34 gray-tone distinctions in 12 dithering patterns). Scanning sensitivity can be controlled from the keyboard for "brightness," "contrast," and "grain" in order to bring up faint elements on the original or drop out background discoloration. The MS-300A, which is no longer on the market, is a friction feed device that moves single sheets past its scanning aperture on rollers. It can accept single sheets of flat media (.012 inch maximum) in sizes up to 8.5- by 11-inches. This scanner, like most tabletop units, is only effective with black and white or shaded materials; color photos or illustrations drop out to black or white. For storage to disk, Eyestar has a routine that permits 1-way compression of image file data, allowing a space savings of 50 to 90 percent. Compressed files must be individually restored to full bit-map extension for editing and output.

The output device selected for the system was a Hewlett-Packard Laserjet Series II laser printer enhanced with 2 megabytes of additional random access memory to permit full-page (8.5- by 11-inch) bit-map processing. The additional memory is a normal enhancement option offered to support graphics-oriented desktop publishing. The Laserjet II has a 300-lpi resolution and accepts paper sizes up to 8.5- by 17-inches. A simple modification of parameter files packaged with Microtek's Eyestar software produced an effective printer output.

Off-line storage of digitized image files was provided by an Alpha Micro Systems Videotrax tape backup system. This unit permits high-density data storage and retrieval using VHS videotape cassettes processed on a modified videotape recorder. The Videotrax controller board and software installed in the microcomputer enable the user to transfer specific disk files to and from the tape. Data access is sequential--as with all tape-based systems--and data reliability is secured by redundancy (multiple copies of each data element).

Test materials were mainly published items (periodical issues, etc.) donated by the Pacific Southwest Regional Medical Library Service (PSRMLS). All test documents were broken down into individual sheets suitable for the sheet-feed processing required by the Microtek scanner. Materials were selected to provide tests of a range of bibli-

ographic phenomena: typographic text, line art illustration, discolored paper, photographic prints, and photocopied items.

The entire system fit comfortably on a single desktop with a separate stand for the laser printer. Archive tapes were housed in a single drawer under the printer stand. No special electrical wiring or environmental control conditions were required for system operation.

Experimental Profile

The study was not an "experiment" in the full scientific sense of the term; instead, it was an examination of a technical system whose characteristics and capabilities as a system had not been specified or fully understood--hence, an "experimental" system in the engineering sense. The approach taken was akin to that used by reviewers of microcomputer hardware and software: intensive use of system components in order to evaluate general characteristics of performance and cost. Performance tests were carried out by a single operator with a background in microcomputer applications and digitization techniques.

The experimental system constituted a single workstation capable of handling one operation at a time. Multitasking was unavailable in this configuration as an approach to system operation. The standard sequence of cooperation for the four main functions of production, storage, retrieval, and output developed during the course of the study were as follows:

Production

1. Materials selection
2. Materials preparation
3. Digitization
4. Test output and quality control
5. Redigitization (if necessary)

Storage

6. Storage of image files to hard disk
7. Cumulative storage of disk files to tape
8. Access database updates

Retrieval

9. Consultation of access databases
10. Image file selection
11. Restoration of files from tape to disk

Output (for each image file)

12. Bit map restoration (decompression)
13. Bit map submission to laser printer

This sequence of operations was found to be the most efficient within the constraints imposed by the system design and the operational constraints of the various components. Technically, each process could be carried out successfully by an operator familiar with the system. During the course of the study, approximately 1,000 images were digitized using the system; of that total, 331 were stored to videotape and recorded for retrieval in a pair of access databases built in dBase III to respond to known-item searches. Experience with this set of images constituted the dataset for operational testing.

After system installation and integration had been completed, it quickly became apparent that all components were technically capable of performing their required functions. Following a series of test scans to establish production parameters, the Microtek MS-300A successfully produced clean digital images from properly formatted test materials under control from software mounted on the IBM PC/AT. At a resolution of 300 lpi, test materials were legible down to the smallest type sizes and line art items were reproduced with acceptable fidelity. The quality of halftone illustrations was somewhat degraded from the originals, but was nonetheless generally better than that obtained from commercial photocopies. The Videotrax tape backup system accepted files stored on the PC hard disk and successfully copied them to VHS videotapes and restored them to the disk in response to selection and retrieval queries. With the installation of a simple printer driver into the scanner's Eyestar software, the laser printer began producing good quality reproductions of scanned images. At a very basic and undemanding level, the UCLA/GSLIS system began to look like a success and desktop digitization appeared to be an easily operationalized application. However, with systematic operational testing, the limitations of desktop digitization (when supported by minimalist generic components designed for other applications) began to dim the prospects and highlight the problems.

The primary operational variables for a computer system are "time" and "space." The time that system operations require--particularly the human component--is a critical factor in judging efficiency and cost effectiveness. The implicit opportunity costs of committing machines and people to a particular task are the range of things they might otherwise be doing with those minutes and hours. Similarly, kilobytes and megabytes required to store or process digital information and the storage/processing capacities of system components are limiting factors in production and retrieval/output performance. The time and space numbers found for the UCLA/GSLIS system were not encouraging for desktop digitization.

The fundamental unit of analysis for digitized image processing is the scanned image, a body of information more or less equivalent to a hardcopy page of text and graphics. In testing the UCLA/GSLIS system, it was convenient to use the figures for an 8.5- by 11-inch page ("full frame scan") as a benchmark. Although not completely elastic, time and space costs for scanning and processing smaller areas tended to require fewer resources.

In simulated production-run scanning, it was found that under the control of an experienced operator, the system required an average of approximately 2.5 minutes to scan a full-frame image and save the compressed file to the hard disk. Compressed files required an average of about 225 Kb of storage space. Images containing halftone or gray-tone graphics were substantially more "expensive," doubling the costs in production time and storage space.

Using the 2.5 minute average as a production time standard, it was possible to specify the tasks and estimate the productivity of a trained operator for a 2-hour work session, a time period that subjectively matched the tolerance of the operator for this kind of work. The following is a typical workflow for such a session.

Consult access database to determine sequential file number status	3 minutes
Dismount materials with guillotine paper cutter	1 "
Mark materials with sequential file numbers	5 "
Microcomputer system runup and preparation	3 "
Eyestar software preparation	3 "
Digitization and storage to disk (100 min. / 2.5 = 40 images)	100 "
Work break and/or system shutdown	5 "
TOTAL	120 minutes

By this standard, the throughput rate for image production with the UCLA/GSLIS system is about 20 images per operator/machine hour. Although it is difficult to make direct comparisons between production procedures used in this system and those used for other preservation methods, it is instructive to note that a 1986 study of preservation microfilming found an average operator/machine time requirement of 147 minutes to film a 240-page disbound book (Kantor, 1986); the same job on the UCLA/GSLIS system (assuming an 8.5- by 11-inch page format) would require about 720 minutes of straight production time and another 320 minutes for administrative time, including access database update and transfer of data to off-line storage. Clearly, whatever the intrinsic advantages the digital format may have, in a direct arithmetic comparison of production processes, desktop digitization is at a severe disadvantage in this system configuration. In fact, slow response times and throughput rates due to insufficiency of computing power or rigidities in component design are perhaps the salient drawbacks of the desktop approach to digital archiving. This is emphasized by the problems of mass storage and retrieval that are characteristic of graphics-oriented computer applications.

One of the more novel aspects of the system configuration was the use of a VHS videotape system for off-line storage of image files, although a dedicated imaging system by Kirsch Technologies, InfoStation, also takes this approach (Hessler, 1987). The Alpha Micro Systems Videotrax unit, a modified videotape recorder used at UCLA/GSLIS, interfaces with the PC-AT by means of a controller board and associated software. Alpha Micro Systems estimates that a T-120 VHS cassette can reliably store 96 megabytes of data in 2 hours of playing time using standard settings and

their recommended level of data redundancy. Videotrax, like most tape backup systems, was designed primarily to handle high-volume backups common in dynamic computing environments in which an entire hard disk is periodically dumped to tape as insurance against data loss. In such an application, rapid storage and retrieval of individual files is less important than density and reliability of data storage in a convenient and inexpensive medium. Consequently, file storage and retrieval on Videotrax is characterized by relatively slow operating speeds, low unit price for media per megabyte of storage, concerns about long-term reliability of media, and some rigidities and inefficiencies in system design.

The low cost and ready availability of the storage medium is, perhaps, the major attraction. The ability to store 96 megabytes of data on a high-quality videotape that markets for about $5 is difficult to beat. By comparison, 20 Mb disk cartridges for Bernoulli boxes cost about $100, while specially engineered tape backup systems for microcomputers require cartridges costing $30-$40. Alpha Micro Systems and other vendors have taken advantage of the mass popularity of the videotape cassette, the price of which has been brought down to that of the paperback book.

Videotapes are not specifically designed to serve as a storage medium for digital data, nor are videotape recorders constructed to operate as data-processing devices. To operate as a digital storage system, a videotape-based device must be modified to accept data in the digital rather than the analog format, must be capable of recognizing file structures and labeling data in order to perform search and retrieval operations, and must be reliable enough to provide the complete, bit-by-bit fidelity that digital data transfer requires. Videotrax offers adequate technical solutions to all of these problems, including most notably a system of saving multiple copies of each data block to overcome imperfections in the tape surface that could result in loss of data. A badly flawed tape can overcome even the most lavish use of data redundancy; after a series of backup and tape verification tests, however, it was found that only a seriously abused tape recorded at the minimum redundancy of four data copies and "extended play" speed would return the "hard error" messages that indicate irretrievable data loss. For both retrieval efficiency and data security reasons, the UCLA/GSLIS system was operationalized with a two-tape backup scheme: a "working" tape recorded at a minimal redundancy level (four copies) for retrieval and an "archival" tape recorded with enhanced redundancy (20 copies) to ensure preservation of the data set.

The Videotrax system of securing data by multiple copies proved to have serious operational significance. It was found that data security could be increased either by purchasing "high grade" videotapes or by increasing the number of data copies per block. Data capacity per tape could be increased either by reducing the number of data copies or by setting the videotape recorder to "long play" or "extended play" speed. Alpha Micro Systems advises users to record 10 copies of each data block on a tape running at standard speed; the result is a T-120 (2-hour) tape with a 96-Mb capacity and a reliability ratio in excess of 100:1. Maximum capacity for a T-120 tape could be achieved by setting minimum redundancy and recording at "extended play" speed; the result is a tape with a 720-Mb capacity and somewhat lower data reliability. Unfortunately, due to

rigidities in the Videotrax operating system which permit the storage of only one backup per tape, this extravagant data capacity could only be achieved if the unit was connected to an input device with similar capacity--a 720-Mb disk drive, for example. Such drives are not standard equipment on off-the-shelf microcomputer systems.

On the UCLA/GSLIS system, due to maximum hard disk capacity, a limit of 30 Mb per tape was found to be the standard backup. This limit was well below the rated capacity for a T-120 tape and put a hard ceiling on the density of image storage: approximately 130 full-frame images per tape. Although this capacity is equivalent to 85 360-Kb floppy disks, it is still disappointing when the requirements of image data processing are considered.

The time required to store and access data on a videotape unit also depends on the number of data copies recorded. On Videotrax, it takes approximately 45-50 minutes to store or retrieve 15 megabytes of image file data when an enhanced redundancy archive tape is the medium; on the other hand, the same data set can be recorded or restored in 12-15 minutes when a minimum redundancy tape is used. Hence, operating efficiency competes with storage reliability as a priority in deciding how to construct a storage tape. With sequential access, even placement of a desired data set on the tape is a factor in retrieval time. Although automatic fast-forwarding is supported, at minimum redundancy, there is approximately a 1-minute difference between the retrieval times of a sequence of image files at the beginning of a tape as opposed to a sequence at the end of a tape. With enhanced redundancy, the differential increases by a factor of four.

Having detailed so many disadvantages of videotape storage, it is refreshing to be able to point out a singular advantage: the capability of inexpensive mass commercial copying of data sets. Commercial copying of analog videotapes is a fairly well-developed specialty within the communications industry and, technically, the copying process is identical for both digital and analog formats. When test tapes were submitted to a vendor, it was found that data sets recorded at the enhanced redundancy level could be commercially reproduced on high-grade output tapes with satisfactory levels of reliability. The aggregate price for reproduction of 100 copies of a 15-Mb data set is between $250 and $300, plus the cost of the output media. At $5 per output tape, the total cost of duplication would be between $750 and $800. Larger data sets would tend to push the total cost figure closer to $1,000. Standard turnaround time for a job of this magnitude is 24-48 hours. Clearly, large-scale commercial copying of image data sets is technically feasible--and something of a bargain.

The ability of videotape to serve as an archival medium is more difficult to assess. Professionally produced analog videotapes have been held under optimal storage vault conditions for as long as 25 years and then played without appreciable loss of data. Tapes, as we have seen, are also readily reproducible and can achieve reliability through progressive redundancy. Nevertheless, digital data are more sensitive than analog signals, and magnetic media are notoriously susceptible to a variety of environmental ills that are difficult and sometimes expensive to control. It is possible to indicate techni-

ques that will maximize the reliability of a digital videotape--temperature control, for example--but not to estimate its average lifespan.

The slow response time of the Videotrax unit in addressing image files, combined with the need imposed both by Microtek's Eyestar software and by limitations in system RAM and disk capacity for sequential processing of image files, make the system difficult to use as a platform for a "demand printing" service--the only feasible user access venue. Files restored from tape to the hard disk must be individually reconverted from condensed format to full-bit map extension and then sent to the laser printer for output. The result is a retrieval-output throughput rate of approximately 10 pages per hour of operator/machine time. This is plainly inadequate for all but the most leisurely and intermittent of access needs.

The system was configured to operate as a single user workstation. An individual attempting to perform the four major processes (production, storage, retrieval, and output) would find it necessary to master the MS-DOS operating system, three application programs, laser printer control procedures, and physical manipulation of materials and media for the scanner and tape backup system. Some organizational and clerical ability is needed to maintain the archive and access databases, but simple accession number processing and the limitation of search capabilities to author and title keywords keep this requirement to a minimum. As with microform production, visual acuity and a feel for the medium is necessary to maximize image quality, but most scans can be effectively processed with standardized control settings and minimal editing. An individual with the necessary skills and aptitudes could be accurately described as a "technician" or "operator" and could clearly require substantial training to master all system functions. For an individual with a background in microcomputer operations, about 30-60 hours would be necessary for proficiency; for a computer novice, perhaps as much as 120 hours would be required.

In evaluating the UCLA/GSLIS experimental system and the prospects for desktop digitization as an approach to library preservation and access issues, it is important to recall that this is only one of a range of possible off-the-shelf hardware and software configurations. For any technical solution found to problems of digital image processing, more effective (but not necessarily more costly) solutions exist. In fact, the system was obsolescent when examined, and certain of its components are now quite obsolete. For example, the Microtek MS-300A is no longer on the market and its successor, the MS-300G, is a flatbed unit with other enhanced capabilities that is now offered at a lower list price than was the original model. Similarly, the WORM drive, now coming down to personal computer prices, seems likely to take over the mass storage niche filled by the Videotrax unit--or possibly in conjunction with it if videotape copying remains an attractive option. In any case, the selection of obsolescent components for the UCLA/GSLIS digitization system was not studied, but it did fall in nicely with the minimalism of the research approach. Yet, if we must conclude that minimalism in this instance did not pay off with a system powerful enough to support a realistic set of library needs--and I think we must, we must also remember the truism that, in the microcomputer world, today's top-of-the-line component is tomorrow's bargain base-

ment minimum. It is upon this, as much as anything else, that the prospects for so casual an application as desktop digitization rest.

Desktop digitization has not yet been born as an application of microcomputer technology in the library workplace. It is one of the potentialities inherent in a situation in which the opportunity costs of digitization are dropping as a result of the spread of appropriate technology and of the progressive socialization of professional library users to graphics-oriented microcomputer applications and products. The experimental system assembled at UCLA/GSLIS was only one example--a primitive one--of what it is now possible for almost any library to attempt. More powerful possibilities are already on the shelves--and even on some of the desktops.

References

Hessler, David. "InfoStation: A Lo-Cost Electronic Document Storage, Retrieval, and Transmission System," *Library Hi Tech*. 17 (1987).

Kantor, Paul B. *Costs of Preservation Microfilming at Research Libraries: A Study of Four Institutions.* (Washington, D.C.: Council on Library Resources, 1986) p. 2.

BRADFORD S. MILLER

Special Collections Librarian,
Northrop University, Los Angeles,
California

Bradford S. Miller received his M.L.S. from UCLA and is studying for his Ph.D. in the Graduate School of Library and Information Science at that institution. Prior to his appointment at Northrop University, Mr. Miller worked for 6 years as a typographic specialist for Autologic, Inc. During that time, he produced and edited digital typeforms using both scanner and vector graphics (CAD) technologies. Mr. Miller has a book in preparation for Ablex Press entitled "Microcomputer Graphics as a Library Resource."

CONCLUDING ADDRESS

ROBERT M. HAYES

Dean, Graduate School of Library and Information Science,
University of California at Los Angeles

Robert M. Hayes became Dean of the Graduate School of Library and Information Science at UCLA in academic year 1974-75 after 10 years of service as Professor in the School, the first 5 years of which were as Director of the Institute of Library Research. Dr. Hayes joined the faculty after 15 years of experience in government and industry with such organizations as the National Bureau of Standards, Hughes Aircraft, National Cash Register, and Magnavox Research Labs. In 1959, he founded a small consulting and research company, Advanced Information Systems, of which he was President until joining the UCLA faculty in 1964. That company developed some of the first generalized computer programs for file management and pioneered in research related to computer-based information retrieval. With Joseph Becker, Dr. Hayes formed the company Becker & Hayes, Inc., and served as its Vice President until becoming Dean.

Dr. Hayes received his Ph.D. in mathematics in 1952 from UCLA. He has had adjunct faculty appointments at American University, the University of Washington, the University of Illinois at Urbana, and the University of New South Wales in Sydney, Australia. He has been active in professional societies, having been President of the American Society for Information Science, President of the Information Science and Automation Division of the American Library Association, Vice President and Chairman of Section T of the American Society for the Advancement of Science, and Chairman of the Committee on Accreditation of the American Library Association.

Dr. Hayes served as a Presidential appointee, during 1979-80, on the Advisory Committee to the 1979 White House Conference on Library and In-

formation Service. He served as Chairman of the Public Sector/Private Sector Task Force of the National Commission on Libraries and Information Science, with primary responsibility for its report on interaction between the public sector and the private sector in the area of information services. He served as Chairman of the Planning Panel for the National Library of Medicine, concerned with the future of the NLM collection and its organization.

Dr. Hayes has published extensively, including coauthorship with Joseph Becker of two basic texts entitled "Information Storage and Retrieval: Tools, Elements, and Theories" and "Handbook of Data Processing for Libraries," both published by John Wiley and Sons, Inc. He has served as editor for several scholarly publications, including 15 years as editor of the "Information Sciences Series" of John Wiley, a series of monographs in the field that includes over 30 titles. For 10 years, he was Associate Editor of the "Journal of the Association for Computing Machinery."

Dr. Hayes has received several awards in recognition of contributions he has made to both librarianship and information science. Among them were appointments as National Lecturer by the Association for Computing Machinery and the American Society for Information Science, the Beta Phi Mu from the American Library Association for his contributions to library education, and an award from the UCLA Alumni Association for professional achievement, which he received in 1986.

Concluding Address

Robert M. Hayes
Graduate School of Library and Information Science,
University of California at Los Angeles

For very obvious reasons, I am reminded this afternoon of some 40 years ago when I started in the computer business. At that time, Ralph Shaw, then the Director of the Department of Agriculture Library, had been for several years at the forefront of developments in our field. The Rapid Selector, of course, was one of the great experiments we have had. I am reminded of the ensuing years in the history of the NAL, of Foster Morhardt who was such a good friend to Joe Becker and me. And now, Joe Howard, Sarah Thomas, and the NAL staff have brought a new vitalization of the historical position of NAL at the forefront of developments. I am proud to be associated with it in any way.

I feel a great personal debt to all of these people, but especially to Ralph Shaw. He served as the model for me of the ideal professional. I remember a meeting in Kansas City when I had given a talk that was somewhat equivocal, avoiding the harsh words to the effect that a development was not good. Ralph said, "Bob, will you have lunch with me?" We had lunch together, and he said, "Bob, it's your professional responsibility to say it and say it out." In the years since then, I have tried to meet that model, but not necessarily well.

This most recent association with NAL has been excellent from the standpoint of working together. In terms of accomplishing the objectives that Joe Howard visualized, however, the results have been less than successful. I warned Joe when we started that success was unlikely. As you saw from Brad Miller's talk this morning, I am a frugal person, so we chose the minimalist route. I'm not an admirer of John Cage, but minimalist I am, and that's not the way to produce dramatic technological developments.

Joe's concept was that of a magical waving of a wand across text, from wherever it came, a journal published in India or some handwritten script, with the text automatically being converted into the desired abstracts. With all due respect to the claims of Optiram, I think that vision is not yet here. Indeed, I would be interested in the means by which Optiram may have solved a very important and significant problem, namely, how do we produce an interactive system that effectively integrates scanning with the decision process of the human being. I think they have done that very, very well. I suspect, though, that they have a room full of, not elves perhaps, but at least persons sitting before CRTs and keying when it is appropriate to do so.

In any event, I owe a great debt to NAL, to Joe, and to Sarah. This program is certainly part of it. It seems to me to have been one of tremendous value and content. I was impressed with the degree of practicality, with the extent to which we were given facts, not wishes, and the extent to which we were given experiences, good and bad, not hopes.

The value to me was the confirmation we had in data, not only in the data themselves but in the qualitative character of them, of the fact that we are still dealing with a labor-intensive operation. Our experience at UCLA, with a minimalist approach, obviously highlights the degree to which OCR conversion is labor intensive, but I saw the evidence for it occurring throughout all of the presentations. The data on actual error rates, the extent to which they differ from the manufacturers' quoted rates, and the clear identification of the effects of different levels of error--these all make evident the critical point of decision. It's almost a flip-flop decision at about 95 percent accuracy; above that rate, one can make the OCR choice, below it the keyboarding choice.

We saw a range of contexts in the results presented: OCR applications in optical disk storage, other applications, and the conversion of images through scanning. Throughout it all, the predominant theme was the rate of change in the technology. The costs are coming down rapidly; the software capabilities are developing at a spectacular pace and we are learning how to use the technologies. Indeed, the fact that we are learning so much about use may well be the most important result we have obtained.

Normally, a keynote sets the stage for a conference, so that everyone can, during their talks, say, "As the keynoter said, ..." But the presentations have been so excellent, so much of a common fabric, so complete that a keynote that simply set the stage for them, particularly coming after the fact, hardly makes a great deal of sense. So this talk cannot be a keynote. Instead, I hope it may serve as a prelude to the future.

So I want to ask, where do we go from here? How do we build on the capability we are developing? As I have said, that may be the significant accomplishment achieved during this conference--the capability to deal effectively with the technology and its applications.

To answer those questions, I must turn from what I think has been the inward-looking focus evident in the papers to an outward-looking one. I raise the issue not of how we can apply this technology to the library, its operations and even its services, but of how we can apply it to the needs of the users. Therefore, I want to step back from the specifics of this conference and set them in a larger frame of reference.

In part, that larger frame of reference reflects a project that I have been involved in at UCLA for the last couple of years. Among other things, I've been engaged in an effort at strategic planning for information resources in the university. It is a project sponsored by the Council on Library Resources. The intent is not to serve as a model but to explore some of the avenues toward strategic planning. It places its emphasis, at least initially, not on the library and library decisions, but rather on university and faculty decisions.

The approach taken, therefore, has been to encourage faculty to identify and examine its information needs. A memo was sent to the 1,500 to 2,000 faculty members at UCLA, stating that money was available to them if they had interest and need in examining an information problem that they perceived as important during the coming

decade or so. To show you my frugality, $1,500 was the amount offered, but interestingly enough, I did get faculty who said they'd like to have some of that money. I have supported on the order of 30 to 40 such studies on campus at that level and perhaps half a dozen at about $15,000.

The result of the studies has been the identification across the campus of some generic needs. It is on one of them that I want to focus, since it is directly related to this conference and is paramount among the needs that have been identified. I'm personally convinced that it will represent one of the most important components of research progress over the coming decades. It is the digitized image. Perhaps it comes as no surprise to this audience that digitized images are important in faculty research, but I haven't seen that much attention paid to it in the library literature. In the remainder of my comments here I would like to discuss it. I will identify three categories of needs in an order of increasing complication and, I think, of increasing importance.

The first category of needs is that to which we have been paying attention for the past day and a half. It is the conversion of existing data for purposes of preservation and use. In fact, that need was part of the impetus for the strategic planning effort and it's worthwhile amplifying on it.

This need was part of the motivation for a campus committee of faculty and administration concerned with what we called "the library of the future." One of the key problems was that of dealing with the film archives on campus. It is a huge file--we had just received a donation of the Hearst Metrotone newsreels covering several decades, for example--and the problem was how do we handle this massive collection? How do we handle film that is deteriorating and, in many cases, is explosive? How do we provide access to it? Do we convert the film to another medium, and if so, to what and how?

In addition to the film archives, there are the problems in preservation of library print materials; conversion is one of the means of solution. And throughout the campus, in areas of key importance to faculty, there are catalogs and files that need to be converted. As an example of that, I think of the catalog and files on folklore and mythology--one of the finest collections of folk medicine anecdotes and data existent, but it's all on 5" by 8" cards. How do we put that in a form for access and use? During my visits to other universities for the purpose of examining preservation needs, I saw at Cornell their wonderful collection of maps, but it was in sadly deteriorating condition. How do we convert them?

These are all examples in the spirit and form of our discussions at this conference. They are all important and I don't want to underestimate that, but I do want to emphasize that they are only the tip of the entire scope of needs. Of greatest importance, even with respect just to these conversion needs, is the fact that they are not simply library needs; they are faculty needs which can be solved by effective collaboration between the library and the faculty.

Let me turn now to the second category: digitized images being created by programs and operations. There are programs and processes designed to produce images. An example is the use of CAD-CAM in engineering, in which the images are the equivalent of engineering drawings but with vastly greater utility. The development of algorithms and procedures for translating conceptual designs into images is pervasive throughout the university. The needs arise in architecture and urban planning. in the imaging of dance, in film and television, in cartooning, and in the entire revolution created by the Macintosh and its user interface. And these only begin to identify the examples; there are probably two to three times that throughout the campus.

Now, these examples represent controllable image files, perhaps not much different in character from those we have dealt with over the past 2 days. However, they do raise interesting problems in the relationships between the generating programs and the resulting images. Which of the images do we store and which do we simply regenerate from the algorithms? When do we make the algorithm the means for storage and retrieval and when do we treat the image as the basis for it?

Now, let me turn to the third and final category--the great monster. I think it is the area of greatest importance, the greatest magnitude, and one in which I hope the library community can and will play a vital role. It is the great explosion of digitized images generated from automatic observation of natural processes. Before starting the strategic planning effort, I was aware of this phenomenon, as I'm sure all of you are, but I hadn't brought the full import of it into focus. It was when I began to receive requests from all over campus that clearly exemplified this need that I began to realize how important it had become.

Just to illustrate, several departments from the School of Medicine independently raised the question of management of huge files of digitized images. From radiology, from neurology, and most interesting pediatrics, I received requests for use of that limited amount of money--$1,500--to examine their needs in digitized image filing and access. Consider radiology, encompassing X-ray and supersonic imaging; similarly, neurology with cat-scan imaging. In each case, dozens of frames can be generated each second. These are not being generated by algorithms but by the monitoring of natural processes, and the volume of data involved is unbelievably great.

The example in pediatrics is interesting as an illustration of the use that can be made of digitized images. They are taking images of a child's knee, for example, in order to perform digital surgery upon the image to see how the child's knee would then flex after the surgery. This kind of digital image processing is being done throughout medical research and practice.

A second major context for generating digitized images of natural phenomena is satellite data--the earth satellites and the space exploration probes--pouring in at rates that boggle the imagination. These kinds of data are vital in the space sciences. Such data are also vital in agriculture, in urban analysis, in water resource planning, and in a myriad of other technical and societal contexts. How do we manage these files?

In high-energy physics, in the use of the supercolliders, we will be generating digital images as essential bases for subsequent analysis. How do we store and retrieve these data?

The papers we have heard over the past day and a half have all been focused on the first of the three categories I've described--that of converting existing data files. Conversion of catalogs and documents, and conversion to alternative means for distribution--these are without question important.

But I do suggest that you consider the other two categories. I think the capabilities we are developing--our understanding of image processing, our experience with the realities of image storage and distribution--are applicable to them.

Let me raise just some of the things with which we might be concerned. The first is the professional librarian's expertise in identification, what we might call cataloging if you will. It is needed in both descriptive and subject aspects, for both management and access. It is evident, for example, that the files in radiology and neurology require the kinds of access tools that cataloging provides.

Beyond that, though, are the fascinating problems in content retrieval. As soon as we consider satellite data, for example, we must consider geographic areas, the identification of manmade objects within the image, and the retrieval of images containing specific elements. In other words, we face the counterpart of full-text retrieval from documents. The problems are many and fascinating. Their solution will take far more than my lifetime, but what a wonderful challenge to our profession!

It's not an impossible problem. Just to illustrate, let me present two techniques that are applicable. First are the methods of spectral analysis which provide a mathematical measurement of properties of the digitized image; such methods can be applied to image enhancement, for example, but they also provide a means for characterizing information content and for subsequent retrieval of documents with a desired spectral composition.

Second are the methods of "scene analysis," in which the digitized images are considered not in isolation but instead as parts of sequences. There are straightforward methods for determining a continuity across a sequence and points at which breaks or scene changes occur (for example, the rate of change in statistical properties from image to image in the sequence). That means we have the ability to retrieve "clumps" from a digitized image file; the clump constitutes a scene, and one of the images in it, in fact, could serve as a surrogate or representation of the clump as the basis for retrieval.

I see the role of the library as multiple. First, I think that the technical tools and capabilities of the librarian can be brought to bear on the problems involved with storage and retrieval of digitized images; that will require the augmentation of those technical tools, and that in itself is exciting. Second, I want the library to be the testing

ground, as we have been in the past, for the practical applicability of technologies in an operating environment; we have the ability to test, as the past 2 days have demonstrated with respect to the evaluation of scanning technologies as real, economic, and effective.

Third, I see the role of the library in acquisition and preservation of images of wide-ranging value. In this context, I see already in limited areas, and increasingly so in the future, the publication of digitized images. Libraries need to be prepared for acquisition of them. Obviously, we have this in the form of CD-ROM image databases, but I see it in other aspects as well. I look forward to publication in this form for the needs of medicine, agriculture, and urban planning.

This is not merely my pipe dream. A couple of weeks ago, as I was formulating my thoughts for this paper, there appeared an article in the *New York Times*, on October 30, entitled "Super-computer pictures solve the once insoluble." It is a superb description of the increasing importance of visual conceptualization in scientific research. Previously unsolvable problems have been rendered soluble by digitized imaging. It would have been fascinating to see what would have happened if Watson and Crick had used digitized processing and when the spiral would have been seen. I suspect it would have been encountered much earlier, because when you can take an image and easily twist and turn it and look at it from various perspectives, suddenly relationships are seen that otherwise would not have been recognized.

The crucial point in all of this is that you need the supercomputer to do the image processing, but you also must have the files of digitized images to be retrieved, processed, compared, and analyzed. You need to have the tools for organization and management of those files and for retrieval from them.

It is that combination of things that I want to bring to your attention. That's my keynote, if you will. It's a keynote, not for this conference, but for where we may go from it. I'm grateful for the opportunity to talk with you and to present this picture. I wish you all success with your own further steps. Thank you.