Library of Congress
Manuscript Digitization Demonstration Project
Final Report

October 1998

The Manuscript Digitization Demonstration Project was sponsored by the Library of Congress Preservation Directorate and was carried out in cooperation with the National Digital Library Program from 1994-97.

Executive Summary

Principal Investigators

Louis H. Sharpe II Picture Elements, Inc.

D. Michael Ott Picture Elements, Inc.

Contributing Author

Carl Fleischhauer National Digital Library Program
Library of Congress

Executive Summary

This following questions framed the Manuscript Digitization Demonstration Project: What type of image is best suited for the digitization of large manuscript collections, especially collections consisting mostly of twentieth century typescripts? What level of quality strikes the best balance between production economics and the requirements set by future uses of the images? Will the same type of image that offers high quality reformatting also provide efficient online access for researchers?

A project steering committee drawn from several Library of Congress units determined that the Federal Theatre Project documents selected for this activity were typical of large twentieth century manuscript collections at the Library of Congress. The normal preservation actions that would be applied to collections of routine manuscript documents include (1) rehousing the original documents to conserve them and (2) creating a preservation microfilm copy.

After discussion, the committee reached consensus that the importance of these manuscripts lies in their information value. A preservation microfilm would be judged successful if the documents were legible and if a researcher could gain a reasonable sense of what the documents looked like. But there would be no expectation that the microfilm image would offer a fully realized facsimile of the original.

During the project's Phase I, the steering committee viewed a number of sample digital images produced by Picture Elements, the project consultants, and selected two image types for testing. Grayscale and color images were selected for the highest quality reproduction, called *preservation-quality* in this project. The committee agreed to tolerate some aesthetic degradation of the images so long as legibility was not impaired and agreed that “lossy” JPEG compression could be applied to the preservation-quality images. For the *access-quality* images, the model of microfilming projects influenced the outcome and bitonal images were deemed useful as a supplement to the preservation-quality images. Bitonal images resemble the high-contrast images familiar to microfilm users and offer small file size (for ease of use in a computer network environment) and print efficiently and cleanly from a laser printer.

During Phase II, Picture Elements produced a test bed of 20,000 images representing two versions of each of 10,000 pages. The image specifications were:

- Preservation-quality images
  - Grayscale (8 bits per pixel) or color (24 bits per pixel)
  - 300 dpi
  - 10:1 JPEG compression
- Access-quality images (initial type; see discussion below)
  - Bitonal (1 bit per pixel)
  - 300 dpi

The Library found that the grayscale and tonal preservation-quality images were generally very satisfactory, that some bitonal access images were satisfactory, and that other bitonal access images were unsatisfactory. There are two reasons for dissatisfaction with some of the bitonal access images. First, some had lost legibility. This was largely related to the original documents themselves, many of which consisted of typed carbon copies on onionskin paper, marked up by a lead pencil. Such documents are nearly impossible to reduce to a clean bitonal image in which all marks are retained.

The second reason for dissatisfaction with the bitonal documents applied to the entire set and reflects the exigencies of *access* in the World Wide Web environment. World Wide Web browsers are not natively set up to accommodate TIFF-format, ITU T.6 (Group IV) compressed images, and the Library found that it needed to produce an additional set of access images for the World Wide Web. These were reduced scale tonal images in the GIF format, produced by running a batch conversion process on the previously scanned preservation-quality images.

The project's findings are summarized in Section 14.

Next Section | Previous Section | Contents

1. Background and purpose of the project

Introduction.

The Manuscript Digitization Demonstration Project was sponsored by the Library of Congress Preservation Office in cooperation with the National Digital Library Program (NDLP). This report includes copies of sample images created during the project's Phase I, which extended through 1995.[1] During 1996, Phase II of the project created a testbed of 10,000 images of manuscript items from the Federal Theatre Project collection in the Library's Music Division. These images are now online as a part of that collection; selected examples have been referenced and made accessible in later sections of this report.

Background. The Library of Congress is developing its capabilities for providing computerized access to its collections. In part, this means wrestling with practicalities of production and identifying and testing a broad range of tools and techniques. In part, it also means investigating the ramifications of digitization as it pertains to preservation, understood to include both the conservation of the original item and the conversion of originals through preservation reformatting.

Preservation reformatting refers to the copying of items as a safeguard against loss or damage, i.e., insurance that the world's heritage will be kept alive for future generations. Today, most preservation reformatting consists of microfilming, although other types of copies are also made. Two features are of special concern to those responsible for carrying out preservation reformatting: the faithfulness of the copy and its longevity. This demonstration project was concerned with the former, i.e., image quality. Other parallel projects are investigating longevity issues.[2]

The Library commissioned the Manuscript Digitization Demonstration Project because it believes that certain classes of manuscript documents lend themselves to the creation of digital copies that are faithful to the originals in a reasonably efficient manner. The Library was cognizant of the work being carried out by the Cornell University Library regarding printed matter[3], and saw that manuscripts would make for a useful demonstration project at the Library of Congress.

A key issue for the Library is finding the most judicious balance between conserving precious original documents--protecting them from damage--and achieving a reasonably rapid rate of conversion. The outcomes of this project are expected to assist the Library in designing models for further conversion applications for the Library's collections.

Manuscript collections. The manuscript holdings of the Library of Congress include extensive papers of individuals and organizations, many from nineteenth and twentieth century America. Since the Library's digitization efforts are initially focused on its American holdings, this demonstration project emphasizes the physical types of documents found in these papers collections. The specific test documents were selected from the Federal Theatre Project collection held by the Music Division. The Federal Theatre Project (FTP) was a New Deal effort that employed out-of-work playwrights, actors, directors and stagehands to produce and perform plays in many American cities during the latter years of the Great Depression.

For the purposes of this project, a manuscript page was defined as a separate handwritten or typed sheet of paper, generally at A size or legal size, i.e., from 8.5x11 inches to 8.5x14 inches. The test documents include scripts, administrative files, and surveys of theater genres commissioned by the FTP.

During Phase I, a set of documents was used to produce a variety of sample images for study. Examples of these images illustrate this report and are accessible from Appendix A. A portion of the sample set represented paper in good condition with reasonably clear, dark writing on a reasonably light background. The other portion of the preservation research sample included documents that represent typical scanning problems:

- a mix of colors or pencil and ink,
- low contrast and carbon copies of typed materials in which the edges of the character imprint are soft,
- documents that have extraneous markings or print-through.

The Document Digitization Evaluation Committee. The Manuscript Digitization Demonstration Project was carried out by Picture Elements, Inc., working in close relationship with a special Document Digitization Evaluation Committee. This committee was made up of Library of Congress staff members (listed here alphabetically) representing various units with an interest in digitization.

- **Ardith Bausenbach**
  Automation Planning and Liaison Office, Library Services
- **Julio Berrios**
  Photoduplication Service
- **Lynn Brooks**
  Information Technology Services
- **Paul Chestnut**
  Manuscript Division
- **Carl Fleischhauer**

National Digital Library Program; project planner and contracting officer's technical representative
- **Nick Kozura**
  Law Library
- **Basil Manns**
  Preservation Research and Testing Office
- **Betsy Parker**
  Prints and Photographs Division
- **Ann Seibert**
  Conservation Office
- **Leo Settler**
  Automation Planning and Liaison Office
- **Tamara Swora**
  National Digital Library Program; project planner and contracting officer's technical representative
- **Peter Waters**
  Conservation Office
- **Walter Zvonchenko**
  Music Division

The committee met on a regular basis during Phase I. At these meetings, Picture Elements representatives reported their survey findings, presented sample images, conducted tours of sites at which special scanners could be examined, and led the discussions that ultimately resulted in the findings and proposals provided in this document.

The activities of Phase II are reported in Sections 12 and 13. The project's findings are summarized in Section 14.

Next Section | Contents

2. Two surveys of the field

Survey of imaging practices. The Manuscript Digitization Demonstration Project began with a survey of imaging practices at selected libraries, archives, and commercial sites. Picture Elements staff carried out the survey in January 1995, seeking information about projects in which manuscripts or manuscript-like materials were being digitized. The survey questionnaire is provided in Appendix C.

Picture Elements contacted archives and libraries through several Internet listserv E-mail discussion forums relevant to libraries. Nearly all returned questionnaires were from organizations responding to the posting on the IMAGELIB listserv. Organizations not responding to the survey, but known (from their publications or by other means), to have projects were contacted by telephone.

Responses to the survey of practices. Formal responses to the survey were received from the first five organizations listed below. Picture Elements supplemented the formal survey with interviews conducted by telephone or in person; this step added responses from two additional organizations.

These responses represented the state of the projects as of early 1995. Since significant changes took place between 1995 and 1997, when this report reached completion, and since a number of new projects have been launched in the intervening years, the following listing is limited to a brief categorization.

University of Georgia Libraries. Paper document scanning project.

***Rutgers University: Thomas Edison Papers*.** Project in planning stage.

***University of Maryland*.** Project in planning stage.

***University of Maine*.** Project converting wide range of materials, including some manuscripts.

New York University. Project in planning stage.

***Cornell University*.** Book scanning project.

Virginia State Library. Historical record scanning project.

Summary of survey of imaging practices. The early 1995 survey uncovered relatively little activity in the digital capture of manuscript documents. There were a few small-scale but no large-scale projects. There was much more activity relating to printed matter, notably the project at the Cornell University Libraries. It is also worth noting that most large-scale efforts in the planning stages focused on printed matter. Where manuscript scanning was being tested or under way, workers reported a mix of binary image production and the production of grayscale or color images.

During the survey period, a published report described a relevant project not covered by the survey. The report detailed the planned approach for an Advanced Papyrological Information System (APIS), based upon the previous work of Duke University and the University of Michigan in the scanning of papyri, a highly challenging subclass of the larger class of manuscripts. This group's recommendations included the use of full 32 bit color at 600 dpi. Archival images were to be stored without compression while access images would be produced using JPEG compression.

The papyrus project has been described by principal investigator Roger Bagnall in an article titled *Digital Imaging of Papyri* in the CPA Newsletter #83, October 1995 and in the report *Digital Imaging of Papyri*, published by the Council on Library and Information Resources (September 1995, ISBN 1-887334-44-0). Additional information on APIS is available from the following sources:

- American Society of Papyrologists
- Papyrology Home Page
- APIS Grant Proposal to the National Endowment for the Humanities
- University of Michigan Papyrus Collection

Although no formal survey of private industry practice was undertaken, Picture Elements relied on their familiarity with this field and reported that digital image systems used by business firms (the field of "office document imaging") generally create binary images. Most commercial operations are scanning forms, office correspondence, and other high-contrast materials with a clean background. Grayscale or color images are produced in exceptional cases such as when photographs must be captured, e.g., for insurance companies, or for the scanning of airline tickets (with their characteristic rich colors and low-contrast, fuzzy features produced by the imprinting approaches used), or with bank checks.

Survey of hardware and software. The consultants also surveyed manufacturers of scanning hardware and image enhancement software.

The survey of available image enhancement software is covered in Appendices D, E and F.

The survey of available scanning hardware for fragile documents and for high throughput is covered in Appendices G, H and I. A presentation on Aspects of High-Speed Scanning is presented in Appendix B.

Next Section | Previous Section | Contents

3. Committee Discussion

Preservation-quality and access-quality images. At the first meeting of the Evaluation Committee, the Picture Elements consultants and the Library of Congress project leaders (Fleischhauer and Swora) sketched the project's key premises and solicited the responses of committee members. A distinction was made between *preservation copies* and *preservation-quality images*. The distinction hinges on the fact that the longevity of a digital image depends upon organizational, procedural, and financial commitments; to achieve full status as a *preservation copy*, a digital image must be kept alive for the long term.[4] This project is dedicated to the development of specifications for images that, if longevity can be promised, will serve the goals of preservation, i.e., will serve as reasonable substitutes in the event that the original item is lost or deteriorates.

Foreseeing that high-quality digital images appropriate for preservation might be large and unwieldy over a computer network, the project's organizers also sought to develop specifications for *access-quality images*. Such images would be lower in either spatial resolution ("dots per inch") or tonal resolution ("bits per pixel") or both, and derived, if possible, from the preservation-quality images. Lower-resolution images--whose digital files should be smaller in extent (bytes)--can be more easily handled in computer systems. The project sought to identify images that, although less faithful to the original than preservation-quality images, offer high legibility and good service to researchers.

Nuances of the term *preservation*. The committee's response to the project outline included discussion of the word *preservation*. When original documents are retained and conserved in the manner of most Library of Congress manuscripts, some members asked, when are reformatted copies (whether microfilm or digital) for preservation rather than for access? The discussion did not provide a concrete answer but proceeded to raise two additional related questions:

- Given the importance of conservation when original items are retained, is it not as important to reformat without damage as it is to produce the highest quality reproductions?
- Why make perfect-facsimile reproductions of routine documents when researchers only need to obtain the information the documents contain, especially if high quality facsimiles are more costly to produce and to store?

Although there was no formal poll of the committee, the discussion of these two questions appeared to reach a certain consensus. If an original document is retained, and especially if it is not a "treasure" (1) the stakes are lower regarding the quality of the copy and (2) the copy must be judged (at least in part) in terms of the access service it provides, e.g., legibility. Some committee members distinguished between producing a copy of a retained manuscript document (lower stakes) and the reformatting of a brittle book destined to be discarded (higher stakes).

In response, other committee members argued that the group need not be so shy of the adjective *preservation* in this context. The typical microfilm of a manuscript collection does not offer a perfect facsimile of the original documents but is nevertheless called a preservation copy.

The committee pointed out that no discussion of a preservation reformatting could omit due consideration of *conservation*. Especially if the original item is to be retained, it must not be damaged in any way by the reformatting process. Conservation treatment may precede or follow scanning but ought to be part of any reformatting plan. Some members said that it would be better to suffer an inferior image than to injure the original.

The committee also noted that if a digital reformatting project provides good *access-quality images* (for general access) and extremely faithful *preservation-quality images* (for scholarly use and as a source for future access-quality images), it reduces the need for continued physical handling of the original. Thus one handling (at conversion time) can obviate a hundred handlings by patrons in ensuing years.

Document look and feel. The committee discussed the degree to which an image need replicate the look and feel of an original manuscript (or other archival) artifact. Although most members agreed that Library of Congress treasures, e.g., drafts of the Gettysburg address, warranted a kind of museum-quality facsimile, there was less consensus that such treatment was warranted for routine documents, especially the kind of twentieth-century typescripts that form the greatest portion of the Federal Theater Project collection.

One manuscript curator pointed to decades of successful use of microfilm by researchers, stating that "most historians seek the information in the document and are not passionate about the look and feel of the paper." Since most Library of Congress manuscript collections are conserved, researchers who need to see such elements as paper watermarks or the direction of fold lines can arrange to examine the originals. Prints and Photographs Division staff reported that their division had been very satisfied with the use of access-quality electronic images; their preservation copy is typically a large-format negative or color transparency.

The committee also noted that the Library does occasionally microfilm and discard a manuscript collection, e.g., a current project to reformat a large accumulation of unpublished copyright-deposit playscripts dating from the first half of the twentieth century. No special effort is being made to increase the quality of the microfilm in this instance. Further, some members said, the existence of one million or more pages of routine typescript would weigh heavily against making an extraordinary effort to produce a museum-quality digital facsimile of each page.

Binary and tonal digital images. Playing the role of devil's advocate, the Picture Elements consultant pointed out that one might interpret the argument against the need to produce near-perfect facsimiles as an indication that binary digital images would be satisfactory for preservation. In a binary image, only one bit per pixel is retained, representing either black or white. Such images are frequently used in office-automation imaging and tend to resemble the familiar (and well accepted) appearance of document photocopies. Did the committee, the consultant asked, view a binary image as an idealized form of the original paper document? There is some justification for this view: the thresholding operation that creates a binary image from a grayscale image attempts to sort foreground markings which are important (turning them full black) from extraneous, background information (turning it white).

High-quality tonal (grayscale or color) images, however, were favored in remarks made by several committee members and the consultants. Some committee members pointed out that it is not always clear what constitutes *information* in the case of a manuscript document. Penciled marginalia, stricken-out first drafts, and coffee stains are regular features of manuscript documents and, in some cases, are part of what invests them with historical value. Those features contain tonal information, or at least require that a copy image provide tonal distinctions in order to keep them perceptible in the image. In an elaboration of the argument in favor of tonal images, the consultant pointed out that it was not always the case that binary images could be more efficiently produced than high-quality tonal images.

Thresholding: the crucial aspect of image binarization. The consultant reported that despite decades of development, the most difficult aspect of binary imaging remains *thresholding*, the determination of the setting which determines whether a given "stroke" or mark on a document is translated into black (or not, and rendered as white, thus becoming indistinguishable from the background). Even the most advanced thresholding techniques, which use the edge information inherent in an image, still face a fundamental difficulty: in order to render the information as black and all other features of the document white, a judgment must be made as to what constitutes information for a given document. This is a difficult, but soluble, problem when a clear set of objective characteristics can be defined, e.g. all the red ink is information, but when a subjective element is introduced, where different researchers may be looking for different information, a general thresholding solution cannot exist.

The consultant cited an example from work carried out by Picture Elements in the field of bank check imaging. Personal checks often contain colorful pictorial scenes across the face of the document. To aid the legibility of the *information*--the payee, the amount, the signature--one might wish for the scenic background to drop out, i.e., to be turned white. Yet if a processing system does this, the customer may state that this image is not of one of their checks, for it has few of the familiar features they use to make that judgment. Thus the pictorial feature may be seen at different times and to different users as either information or as noise.

If the scanning equipment operator must make judgments at scanning time about which features of a manuscript are information, production will proceed slowly and results will be uneven, varying with the individual operator, who would have to be more skilled and thus more highly paid. The approach would be more prone to mistakes, which would result in costly re-scanning.

Another onerous and costly labor burden ensues from the use of binary images: the need for something approaching 100 percent inspection after scanning. If a low-contrast, but significant piece of information (such as a marginal note) is missing, it could render the image useless for some purposes. This form of image "failure" is the more worrisome in that the user of such an image may have no warning that information is missing. For the same reason, to do an adequate job, the visual inspection person would ideally compare each screen or printout image with the paper original, an ungainly and very expensive process.

Argument for tonal images. Regarding the efficient production of tonal images, Picture Elements argued that this image type preserves subtle shadings without requiring irreversible, skilled judgments at scanning time. Thus production can proceed in a cost-effective manner and with more uniform quality. Although grayscale and color imaging may require more of the hardware and software that processes the larger images, the reduced reliance on the application of operator judgment can reduce the cost of producing high-quality images.

There was some discussion of file size as a consideration, especially given that tonal images will be much larger than bitonal images. The consultants argued that the predominant cost in a conversion project is the labor, which runs in some multiple of $0.10 per page. Storage costs run currently in the range of multiples of $0.01 or $0.001 per page. Even if a compressed grayscale or color image were to be four times the size of a compressed binary image, its storage cost would be dwarfed by the cost of labor to scan and inspect it. The consultant also pointed out that, for the last twenty years, storage costs have been halved approximately every three years and there is every reason to expect this pattern to continue.

Some members of the committee, however, counter-argued that this analysis did not take into account the complexity and high cost of managing a server-based storage system through, say, two or three cycles of obsolescence, data migration, and backup. A fully realized storage system, they said, would be very costly.

The consultant noted that the production of tonal images would support a transition to future digital formats and standards. The arrival of new image types may warrant the migration of images captured today into new formats that are enhanced, compressed, or stored in different ways. If new images are to be produced from existing digital files (rather than by rescanning the originals), this will be more successfully carried from a high-quality image than from a limited-tone or low resolution image. Rescanning the originals is a very costly alternative that places documents at additional risk and may not always be an option.

Scanning microfilm. The consultant reviewed the argument advanced by some librarians that capture on microfilm will also permit scanning to meet future digital developments. In this model, the film represents the preservation copy and derived digital images offer online access. This approach carries an increased cost, the consultant said, due to the need to scan the film and--if the master negative is to be protected--the need to produce a copy of the master for scanning. These analog-image generations will result in a digital image that will be inferior to that produced by scanning the paper original.

Perfect reproduction not a *sine qua non*. This general discussion of preservation and preservation-quality images yielded some helpful principles for the deliberations that followed:

- Perfection of reproduction (i.e. image quality) was not a solitary goal, but one key consideration in developing a strategy for creating digital reproductions of a given collection. For typical manuscript-document collections, it is not a sine qua non to have utterly perfect images.
- A number of practical matters--safety of the originals, efficiency of production, cost--are equally important co-considerations.

Topics not covered. The committee noted that at least two aspects of digital production were not included in this demonstration project. The curator of the Federal Theatre Project collection pointed out that the project's focus on separate-sheet items excluded the many bound items in the collection. Others on the committee expressed regret at the exclusion of bound materials, stating that the handling of such items presented special problems for custodians. The project planners responded that the handling of bound materials represented a complex problem of its own and this topic had been excluded in an effort to make the current project more manageable.

Additional discussion noted another exclusion: printed halftones. Such pictorial elements--typically the reproductions of photographs in books and periodicals--are ubiquitous in printed matter but less frequently encountered in manuscript collections. Printed halftones present their own thorny set of scanning complications and also represent a problem of their own, significant enough to require a separate project. In contrast, hand-drawn sketches are more common in manuscript holdings but since they typically involve the same marking devices (pens, pencils, charcoal, and the like) used for writing, these pictorial elements need not be separately discussed or studied as imaged elements.

Next Section | Previous Section | Contents

4. Sample documents and initial image capture

Selecting samples. The opening round of discussions set the stage for selecting sample documents and and producing test images. The small number of responses from the library and archive survey led the committee to charge the consultants to be guided by the general trend of the committee's discussions and by their own technical experiences.

The Music Division curator worked with the consultants to select 35 documents from the Federal Theatre Project collection. The samples were intended to (1) represent the range of imaging difficulties present in the collection and (2) include the boundaries of that range. i.e., the more difficult items. In addition, the documents were selected to include typical small features in order to assist in the determination of levels of spatial and tonal resolution. The set included documents with some of the following characteristics:

- typewriting in black and red ribbon
- pencil strike-outs over typewriting
- red rubber stamps
- red pen underlining
- extremely dark or stained paper
- penciled handwriting
- library book request slips with print-through
- playbill cover with a printed halftone
- thin onionskin
- paper mimeographed copies

Although the committee's discussions had provided the guideline that facsimiles need not reach a pinnacle of perfection, the samples suggested that high-quality grayscale or color images would be required to retain the information in these documents. But since the survey had indicated that bitonal images at a high spatial resolution were often proposed as a preferred approach (usually for printed matter), Picture Elements agreed that some test images of this type should also be produced.

Initial capture. Picture Elements began with "raw" capture, producing 35 uncompressed images to serve as image source for the various manipulations to come. Grayscale capture of 8-bits-per-pixel at 300 dpi was employed for 25 images. Color capture of 24-bits-per-pixel, also at 300 dpi, was used for 10 documents.

These images were then reprocessed in a series of experiments which compared varying spatial resolutions and varying levels of compression, using the JPEG (Joint Photographic Experts Group) compression algorithm. The processed samples would permit comparisons of grayscale and binary renditions of the images and would demonstrate the adequacy of deriving binary access images from JPEG-compressed preservation images as opposed to deriving them from the raw (uncompressed) original images. The actual digital files were delivered to the Library as digital image files on a recordable CD-ROM disk. The 24 images deemed to represent the most distinctive examples (10 of them also in color) accompany this report in Appendix A.

For the committee's deliberations, the digital image files were documented in a set of high quality printouts produced on a photographic-quality printer (Primera Pro from Fargo). Copies of these printouts are available in the offices of the National Digital Library Program at the Library of Congress.

The experiments, and the committee's reactions to them, formed the basis of the set of choices for the image formats for the Federal Theatre Project. The committee guided Picture Elements's work in this area through an iterative set of changes. These changes, and the pointed comparisons which resulted from them, led to the final recommendations, which were placed before the committee for approval.

Next Section | Previous Section | Contents

5. Preservation-quality images: types and resolution

Note: *This and following sections are illustrated with selected images from the project; Appendix A includes the sample-image set. Each sample is presented in the format employed during the committee's comparative examination, including many TIFF images. In order to view the images, readers of this report must equip themselves with suitable* image-viewing software. *The grayscale images have a tonal resolution of 8 bits per pixel while the color images are 24 bits deep. For display of the full image quality, the viewer's display monitor adapter must accommodate these levels of tonal resolution.*

Tonal versus binary images. The committee agreed with the consultants that, generally speaking, the tonal images reproduced the documents with greater fidelity than binary images. Some binary images offered equal but not superior legibility. Binary images provide significantly smaller file size and easier printability but the committee felt that these advantages were more significant in the consideration of options for access images (see Section 9 below).

Marks or strokes with different densities (thickness, darkness) could be distinguished better in the tonal images than in the binary. For example, in a typewritten script with pencil strike-overs, the underlying typed words could be seen more clearly in the grayscale images than in the binary. A high degree of verisimilitude was also noted for a handwritten document. Greater legibility was noted in the grayscale version of a printed-and-written call slip.

Note: *The grayscale examples reproduced here have been contrast stretched and compressed with the JPEG algorithm; comparisons of uncompressed and compressed images are provided in Section 8 below. These are the full-resolution images. Appendix A offers more network-ready alternatives, including thumbnails.*

- Full sheet marked-up script sample PD1
  - Grayscale 300 dpi, JPEG, 577 KB
  - Binary 300 dpi, TIFF G4, 73 KB
- Full sheet marked-up script sample PD2
  - Grayscale 300 dpi, JPEG, 602 KB
  - Binary 300 dpi, TIFF G4, 84 KB
- Full sheet handwritten sample SS3
  - Grayscale 300 dpi, JPEG, 622 KB
  - Binary 300 dpi, TIFF G4, 127 KB
- Cutout from call slip sample NY1
  - Grayscale 300 dpi, JPEG, 15KB (small cutout)
  - Binary 300 dpi, TIFF G4, 7 KB (small cutout)

Color images indicated both the difference in color and the actual color in the document with red-ribbon and black-ribbon typing. Grayscale images successfully indicated the differences between red and black, albeit without a specific indication of the color. In the binary images, the distinction was lost.

- Cutout from sample SS6
  - Grayscale, 300 dpi, TIFF uncompressed, 1026 KB (small cutout)
  - Binary, 300 dpi, TIFF G4, 10 KB (small cutout)
  - Color, 300 dpi, JPEG, 98 KB (small cutout)

Based on these and other similar examples, the committee endorsed the consultant's proposal to produce tonal images when the testbed was created during the project's Phase II and to capture in color those documents having significant color content.

Levels of spatial resolution. The tonal images were presented at three levels of spatial resolution. Although the collection is too large and varied to permit an exhaustive survey, the selected documents were thought to include the smallest typical features, thus offering a reasonable test of spatial resolution. Upon examination of the images, the committee was surprised to see that the differences between the levels of resolution were not terribly significant. Although 300 dpi provided better detail than 200 dpi, few felt that increasing the spatial resolution of the tonal images to 600 dpi produced noticeable improvement for Federal Theatre Project collection documents.

Note: *The 600 dpi sample was scanned at an optical resolution of 400 dpi and interpolated to yield 600 dpi. The 200 and 300 dpi samples were averaged from the 600 dpi sample.*

- Cutout from call slip sample NY1
  - 600 dpi, TIFF uncompressed, 1026 KB (small cutout)
  - 300 dpi, TIFF uncompressed, 257 KB (small cutout)
  - 200 dpi, uncompressed, 114 KB (small cutout)
- Cutout from marked-up script sample PD1
  - 600 dpi, TIFF uncompressed, 1026 KB (small cutout)
  - 300 dpi, TIFF uncompressed, 257 KB (small cutout)
  - 200 dpi, TIFF uncompressed, 114 KB (small cutout)

The consultants pointed out that spatial resolution beyond the minimum necessary is costly (increasing the uncompressed image size as the square of the increase) and offers minimal gain. This is because that minimum necessary resolution preserves all the required features to assure perceived fidelity. When spatial resolution is such that one or two full pixels fall across the width of the thinnest stroke on the manuscript, that stroke will be preserved. At standard viewing distances, further increases in, for example, the accuracy of edge placement (which would come with increasing resolution) would not be seen.

The consultants also noted that an argument could be made that future processing might benefit from the increased resolution. It is worth noting, however, that in cases where a scanner delivers data at a spatial resolution above its native optical resolution (as determined by the spacing and number of the photosites in its image sensor), the future processing could accomplish this same function. Many current scanners tend to be limited to 300 or 400 dpi optical resolution; thus one could wait until a later moment to interpolate the image information to higher resolution. Successful future processing for increasing resolution (by interpolation), however, would depend upon keeping the artifacts produced by compression to a minimum; see Section 8 on compression and JPEG artifacts.

Bitonal images have a more pressing need of spatial resolution beyond that indicated by the 1 to 2 pixel per stroke dictum. When a character (whose stroke thickness varies in a smooth way, reflecting the aesthetic intent of the typeface designer) is represented by binary pixels, subtle changes in the original stroke thickness are rendered as quantum jumps in thickness: from 1 to 2 pixels (a 100 percent increase), or from 2 to 3 pixels (a 50 percent increase). Moving beyond 1 to 2 pixels per stroke lessens this effect, hence the requirement for 600 dpi in preservation bitonal images of text suggested in the widely-quoted Cornell work.

Tonal images instead lessen this quantum effect by subtly "graying out" pixels which are only partially on the stroke; they do this without an increase in spatial resolution. While this "anti-aliasing" technique has been in use in phototypesetting and vector computer graphics for 30 years, imaging designers have recently re-discovered it. When used to scale down a high spatial resolution binary image to a lower resolution tonal image for screen display, it is often called "scale to gray."

Next Section | Previous Section | Contents

6. Preservation-quality images: enhancement

Enhancement options. The consultants introduced the topic of enhancement by enumerating four algorithms that are especially relevant to manuscript and other document images. They recommended that contrast stretching be considered for this project and offered descriptions of the other three. Although some committee members were at first uncomfortable with the term enhancement, suggesting as it does tampering or distortion, most members agreed that the examples of contrast stretching improved legibility with little or no adverse effect.

The consultants argued that reformatting projects should create a record of the processes that were applied when images are created and should report those processes to users. The scanning software written by Picture Elements for this project automatically generates a readable-text log file that includes a detailed record of the scanning process.

Contrast stretching. Contrast stretching in its simplest definition is "making paper lighter and writing darker." In the extreme example presented to the committee, the enhancement rescued a document that was illegible when scanned with the system's default settings. The sample document was a mimeograph print on dark red paper.

- Dark ink, dark paper example captured with system default settings, 300 dpi
  - Grayscale without contrast stretching, TIFF uncompressed, 257 KB
  - Grayscale with contrast stretching, TIFF uncompressed, 257 KB

The consultants reported that in the production phase of this project, contrast stretching would be accomplished in an automated mode by applying a look-up table to each pixel's brightness value to increase the spacing between brightnesses for those ranges of brightnesses more likely to contain information. This can be termed an enhancement, and while in many cases it is only imperfectly reversible, the nature of the modification can, in principle, be stored with the image data as a form of metadata about the provenance of the image.

Depending on the file format used, it may not be necessary to modify every pixel in performing this operation (which is very slow). Some file formats permit a brightness correction curve to be stored in the file which is only used to brighten the image at display time, often using only the video display's palette (which is very fast, and just as effective). This approach is fully reversible, since a future technician could simply drop the correction curve from the file to remove the enhancement. Since this option does not exist for JPEG-compressed files, it was not considered for this project.

In addition, each scanner has a characteristic way in which its inherent brightness curve is imperfect. With test charts, this can be established and documented. In this context, it is worth noting that avoiding any brightening of an image out of a sense that this increases its fidelity is an imperfect strategy; every scanner is introducing its own brightness distortion. If no hardware support exists within the scanner or its interface electronics for the histogram operation which is needed in this brightening process (as is currently typical), the creation of this data in software can slow down scanning cycle times somewhat, regardless of which brightening approach is used.

There was discussion (but there were no examples) of the use of contrast stretching for what might be considered to be an aesthetic purpose, counteracting an effect sometimes seen in tonal document images where paper may be rendered in a middle gray tone. Some brightening of the paper is a natural side effect of (the recommended) objective contrast enhancement method, differing from a subjective one only in the exact goal of the operation. An objective, repeatable contrast enhancement operation will always improve the aesthetics of the image, but in striving not to lose information, may not whiten the paper to the extent a human operator might choose. To a certain degree, this variant form of contrast stretching can be produced in imaging software by adjusting the brightness and contrast settings.

It is worth noting that contrast stretching after JPEG compression is likely to increase the visibility of whatever JPEG compression artifacts are present. The consultants recommended a procedure (and developed an application program) that performed the contrast stretching before JPEG compression.

Deskewing. This algorithm removes skew caused by poor parallelism of the document's edges to the scanning axes. In a platen-type (flatbed) scanning operation, this should not be necessary if the operator uses due care in placing the document and in not closing the scanner lid too quickly.

This operation would be of critical importance in a high-speed scanner (of the type described later in this report) where originals are hand-placed onto a staging platform (with alignment aids, more or less) and then slid onto a moving vacuum belt. Considerably more skew could be expected in that situation relative to a flatbed scanning operation.

Deskewing can be performed on either grayscale or binary data. Much superior results are achieved when using grayscale data, however, this operation is so computer-intensive that hardware assistance is required to accomplish it in reasonable times. Grayscale deskewing has only recently become available in some scanner hardware. Picture Elements has drafted a white paper on this topic.

Some automated software tools do exist for processing binary images. Deskewing in the binary produces considerably poorer results as

compared to grayscale deskewing (when viewed at higher magnifications), since the jagged, stair-stepped jumps seen in scanned straight lines in a skewed binary image are actually increased in number rather than decreased by binary deskewing. While the visual impression is improved at standard viewing distance by the application of binary deskewing, it nonetheless has an impact on the formation of the fine details. For this reason, the consultants did not recommend binary deskewing and no examples were prepared for the committee.

Despeckling of binary images. Despeckling is the operation of removing unwanted small pieces of black (and sometimes white) in the binary image. If the document backgrounds are relatively clean, this might not be necessary. In general, the consultants noted, despeckling is a useful step and many scanners incorporate it natively. For those that do not, a software enhancement post-process can be used to remove speckles. The thresholding technology used by Picture Elements natively includes a despeckler for up to 4 by 4 black pixels and up to 3 by 3 white pixels. Since many of the manuscripts have noisy backgrounds, this enhancement was recommended for use on all binary images.

Display-time smoothing of JPEG-image backgrounds. An enhancement called AC prediction can be performed on JPEG-compressed images. At high compression levels, JPEG images can be prone to "blockiness" in the flat regions of the background where subtle shading changes occur. AC prediction can reduce the perceptibility of the 8x8 blocks in a JPEG image by causing them to match their neighboring blocks better. As with contrast stretching, this can be applied at display time and can create a more pleasing image for the viewer. The consultants' recommendation was to press display-software designers to include this feature in typical viewing software.

Next Section | Previous Section | Contents

7. Preservation images: grayscale and color

Grayscale images. The committee accepted the Picture Elements proposal to use 8-bit-per-pixel capture for grayscale images. The capture of 8-bit grayscale information (providing 256 shades of gray) is sufficient, since studies have shown that a person can only distinguish 32 to 64 shades of gray on a display screen. Furthermore, since the source documents are primarily modal in nature (characterized by dark markings from a limited set of relatively uniform darknesses on a relatively uniform light background) a lesser number shades would probably be sufficient. JPEG (although only the lossless version, not the widely implemented baseline JPEG used in this project) supports up to 12-bit luminances, which would be more useful in scientific or pictorial-image applications.

Color images. The committee discussed but was unable to fully resolve the question of when images should be captured in color. The software used to view JPEG image files generally supports both, so there is no technical impediment to having a mixture of color and grayscale images within the same collection. The argument against color is that is typically takes extra time to capture each color image and color images require more storage space.

The consultants' initial proposal regarding color was that color information be captured whenever color was incorporated in the document by its creator(s), what one consultant called *intentional* color. For example, the Federal Theatre Project includes typed playscripts in which a production company--in effect, a kind of second-echelon creator of the archived document--underlined elements in red. This, the consultants argued, warranted color capture. In contrast, yellowing paper "just happened," and the consultants suggested that color capture was not warranted.

The committee response to this proposed principle was mixed. Some agreed, while others stated their belief that elements like the red underlining need not be captured in color. One curator pointed out that many similar examples existed in microfilmed collections and that there was no evidence that the lack of color on the film had impeded historical research. Other committee members, however, advanced an argument that would have the opposite effect and warrant color capture in a far greater number of cases. These members stated that the yellowing of paper, for example, is significant to understanding the artifact and thus should be captured. But this argument did not prevail and the consultants were instructed to produce the testbed images using their rule for "intentional color" to determine when to capture color.

There was no particular discussion of the relative value of so-called true color, generally understood to mean the capture of 24 bits of data for each pixel, versus the capture of reduced color information, e.g., 8 bits per pixel. The consultants reported that 8-bit color is often called paletted color, because an appropriate palette of colors is selected for each individual image. This process is called color quantization. A rich set of algorithms or methods exist for choosing this palette, but each introduces a different distortion. True color avoids this decision-making process (the creation of the palette), which could be error-prone. The argument is parallel to the one that holds that thresholding a grayscale into a binary images increases the risk of losing information. In fact, thresholding may be viewed as an extreme type of color quantization.

A strong argument can be made that a limited palette is well suited to the manuscript case, where few distinctions of shading exist as compared to, say, a color photograph. Unfortunately, no sophisticated compression algorithm (based on the human visual model) exists for these types of images. JPEG compensates considerably for the surfeit of color information in a 24 bit image; it retains the chrominance components of the image at only half the spatial resolution of the luminance (grayscale) information (so-called 4:2:2 color). In addition, JPEG quantizes the chrominance information more coarsely than it does the luminance information.

Interestingly, a JPEG color image is typically only 20 percent larger on document-type images than a JPEG grayscale image of the same item. This argues strongly for including occasional color in the collection. Mitigating against this view is the observation that many scanners take three times as long to capture a color as a grayscale scan.

Although it is possible that 8 bits or fewer of color information would suffice for many manuscript documents, the widespread use of 24-bit color in the capture of pictorial matter led this project to adopt a true-color approach.

A section on color images is found in Appendix A.

Next Section | Previous Section | Contents

8. Compression

Advantage of compressing images. One of the most interesting discussions pertained to image compression. This topic, of course, takes on great importance once a decision to create tonal images has been reached. An 8.5x11-inch document captured at 300 dpi and 8 bits per pixel creates a file of over 8 million bytes before it is compressed. An uncompressed color image of that document comprises about 25 million bytes.[5]

Lossy compression and JPEG. The most widely used compression algorithms for tonal images are lossy. This means that in addition to removing redundancy from the image in a reversible way, simplifications are introduced into the image's representation that exploit the weaknesses of the human visual system. The goal is to introduce losses that are visually imperceptible under standard viewing conditions.

The JPEG (Joint Photographic Experts Group) standard is a joint international standard of the ISO and ITU-T (formerly CCITT) standards organizations. JPEG compression includes a wide variety of techniques, but of particular interest is the widely-implemented baseline DCT (discrete cosine transform) algorithm. This permits a wide range of trade-offs of image quality versus compressed file size.

JPEG quality setting. The amount of JPEG compression is variable and can be set by the user at a desired level. Most compression software (or software-hardware packages) ask users to set the "quality" at a certain numerical value; the amount of compression actually delivered will vary from image to image, depending upon the image's characteristics. In addition, the chrominance components of a color image (which are also subsampled at half the spatial resolution during the JPEG compression process) are compressed even more strongly than the grayscale component. Thus, at the same quality setting, the compressed 24-bit file will be reduced in size by a greater degree than a comparable 8-bit grayscale file.

Depending on their intended use, images compressed with the JPEG algorithm by factors of as much as 25 to 1 or 30 to 1 can still be very useful, although artifacts created by the process may be visible. These include blockiness in the image, especially visible in "flat" areas of even tonality, and "echoing" or "ringing"--a visible shadow that echoes the sharp edge between dark and light areas, e.g., on a typed or written mark. When compression ratios are lowered to the order of 10:1, the introduction of artifacts is minimal.

- Compression comparison, cutout section of a typed and marked-up document
  - Uncompressed grayscale, 300 dpi, TIFF 257 KB (small cutout)
  - JPEG compressed grayscale examples, 300 dpi, JPEG (small cutouts)
    - Quality 20, 28 KB (9:1)
    - Quality 10, 19 KB (14:1)
    - Quality 6, 15 KB (18:1)

- Compression comparison, cutout section of a call slip
  - Uncompressed grayscale, 300 dpi, TIFF 257 KB (small cutout)
  - JPEG compressed grayscale examples, 300 dpi, JPEG (small cutouts)
    - Quality 20, 23 KB (9:1)
    - Quality 10, 15 KB (14:1)
    - Quality 6, 11 KB (18:1)

- Color compression comparison, cutout section of a two-color typed document
  - Uncompressed color, 300 dpi, TIFF 257 KB(full sheet)
  - JPEG compressed color examples, 300 dpi, JPEG (small cutouts)
    - Quality 40, 234 KB (13:1)
    - Quality 20, 152 KB (20:1)
    - Quality 10, 98 KB (32:1)

The consultants recommended using a quality setting that provides, on average, 10:1 to 20:1 compression for grayscale images and higher for color. Compression of 10:1 was produced by "quality level" 20 in the system used for the preliminary samples; other systems may require different numerical settings. This setting, the consultants said, would reduce both 8 megabyte grayscale images and 25 megabyte color images to 1 megabyte or less.

Lossy and lossless compression. Some committee members expressed reservations about the use of lossy compression algorithms. As had been the case when considering image enhancement, lossy compression suggested distortion or degradation of the image. If an archivist were considering an image for preservation, ought the archived form not perfectly represent the captured bitstream, i.e., be stored uncompressed or with lossless compression? Other members and the consultants referred back to the general principle that perfect facsimiles need not be a *sine qua non* for routine manuscript documents, especially when reckoned against the storage savings afforded by modest levels of compression. The committee's consideration of this trade-off recognized the need of archives and libraries to produce very large numbers of images of documents that have only moderate artifactual value. For the Federal Theatre Project's hundreds of thousands of typescript pages, it was not necessary to have museum-quality reproduction.

The discussion then turned to lossless compression. Some committee members pointed out that algorithms for binary images were lossless and asked whether this might not be a reason to reconsider the provisional decision to name tonal images as the preservation-quality choice. Others asked about lossless algorithms for multitonal images. e.g., lossless JPEG or LZW (Lempel Ziv Welch).

The consultants pointed out that scanners that produce binary images capture grayscale information and then apply thresholding that discards seven-eighths of the information. By definition, this is a very lossy process and the "lossless" compression algorithm is applied *after* this lossy thresholding has occurred. JPEG compression introduces loss at an earlier stage in the process and succeeds in preserving most of the tonal content of the original. The consultants argued that introduction of loss, whether through thresholding in a binary image or through the compression algorithm of a grayscale or color algorithm, was appropriate. Much of the wealth of data in an image is either redundant (hence derivable for a given pixel from its neighboring pixels), imperceptible to people (because the finest of details are not seen well by the human eye), or is simply noise. Many scanners produce data wherein noise dominates the bottom bit or two of the data--oddly enough, the eye often finds this pleasing--but it would be difficult to argue that this noise is an inherent part of the image that cannot be lost.

The consultants reported that lossless JPEG--which is not widely implemented--and LZW have varying performance but often produce around a 2:1 compression ratio. These algorithms could be used, but one might ask if the slight improvement in quality from lossy JPEG merits five to ten times the file size? For certain items, like pictorial works and top treasures, this might be warranted. For other items, like routine documents, it probably is not.

It is worth noting that LZW is patented and requires a license. The family of compression algorithms called "Zip" compression is another alternative with somewhat better performance and no licensing difficulties.

It is also worth noting that the JPEG committee is now actively evaluating several possible replacement algorithms for the current lossless approach, owing to universal disappointment with its performance.

At the end of the discussion, the committee authorized Picture Elements to compress the test-bed images with the JPEG algorithm, applying a quality level that would produce an average compression of 10:1 for the grayscale examples.

Next Section | Previous Section | Contents

9. Defining and deriving access-quality images

Binary versus tonal images for access. The discussion of the relative merits of binary and tonal images for access began with a consideration of computer-storage and file-handling issues. Much future access to documents will be offered via computer networks which, for the current period at least, have limited abilities to move large quantities of data quickly. The consultants noted that smaller, more efficient files could be made either way: binary at moderate resolution (e.g., 200 dpi) or JPEG files with reduced spatial resolution (e.g., 100 or 150 dpi) and/or increased compression (e.g., 30:1).

The discussion turned to the likely actions of researchers who desired the images. Most library and archive researchers, committee members asserted, require printed copies; "printability" is a very important feature to be considered. The printing requirement tends to reward binary images since most present-day laser printers accommodate such images more readily than tonal images.

With this guidance, the consultants produced an array of sample images:

- Typed playscript
  - 600 dpi grayscale source image (derived from a color scan), uncompressed TIFF, 1026 KB
- Typed playscript binary access images (uncompressed)
  - 600 dpi, 25 KB
  - 400 dpi, 15 KB
- Typed playscript
  - 300 dpi grayscale source image (derived from a color scan), uncompressed TIFF, 257 KB
- Typed playscript, binary access images (uncompressed)
  - 300 dpi, 10 KB
  - 200 dpi, 6 KB

The binary examples were compared with grayscale images with both high levels of JPEG compression and, in two examples, reduced spatial resolution. The goal was to create JPEG compressed files that were comparably sized to the proposed binary access images.

Examples of reduced spatial resolution, extreme JPEG compression with visible artifacts produced for the discussion of options for access images (see Section 9 below).

- Typed playscript, uncompressed TIFF 257 KB
  - 300 dpi grayscale source image
- Typed playscript, JPEG compressed access images
  - 300 dpi, C=100, 13 KB
  - 200 dpi, C=100, 8 KB
  - 100 dpi, C=100, 3 KB

Binary preferred. For access purposes, most committee members indicated a preference for the 300 dpi binary image compressed using Group 4 compression, formally called ITU-T Recommendation T.6. (ITU-T is the international FAX standards organization formerly known as CCITT.) For the types of content seen in the FTP collection (typewritten and handwritten letters, no small point sizes), no noticeable improvement over the 300 dpi image was seen in binary images with a resolution of 400 or 600 dpi (interpolated from the 300 dpi grayscale source image). The consultants noted that the file size for the 300 dpi example is smaller than the 400 and 600 dpi examples in direct proportion to the resolution and not in the ratio of the squares of the numbers, due to the way Group 4 compression works. The 300 dpi images print well and reasonably quickly on existing laser printers, many of which are not capable of 400 or 600 dpi printing. The grayscale images that were heavily compressed with the JPEG algorithm were not favored because of the challenge at print time and the visibility of JPEG image artifacts.

The access images are stored in TIFF (Tagged Image File Format) version 5.0 files. Although a de facto ("industry," not formal) standard promulgated by the Aldus (now Adobe) Corporation, the TIFF family of formats is in widespread use and employs a publicly disclosed set of tags to identify various parameters of the image in the file.

GIF and PDF formats. The consultants commented on the GIF format, an alternate way to produce grayscale images and one that is well supported in the World Wide Web environment. GIF images are widely used and employ the proprietary (patented) LZW compression method, which performs relatively poorly on natural images as compared to non-noisy computer-generated graphics (where long strings of identical values are common). GIF images on the WWW are commonly used for navigational purposes to give the user a sense of the content of a larger image before committing to a long download.

The committee also discussed the use of the Portable Document Format (PDF, a proprietary format developed by the Adobe Corporation) for access. PDF images can be viewed in a software called Acrobat; Adobe distributes a read-only version of Acrobat for the WWW at no cost to users. Since a PDF file can contain multiple page images, once it is opened, a viewer can page from one page to the next.

PDF has promise as an access format. It is based on PostScript, which is a presentation-control language (as opposed to an archival

bit-mapped image format). Although PDF version 1.0 did not support binary data without conversion to printable ASCII (gibberish) characters, version 1.1 (current at the time of the demonstration project) permits the inclusion of binary data and natively supports both Group 4 compression and JPEG compression in addition to formatted text with specified fonts. In addition to concerns about the proprietary format, the key drawback to PDF for this demonstration project was the capability of the then-current version of the Acrobat read-only software called Acroread. In Acroread version 2.0, the entire PDF file must be downloaded before the first page is displayed. This was changed in the 3.0 and later releases, which permit incremental downloading (via "byteserving" code on the server) and viewing in a style more compatible with the WWW environment.

In the end, the desire to accommodate printing and to minimize the reliance on proprietary formats led the committee to decide that the testbed's access images should take the form of binary images in the TIFF format with Group 4 compression. When production was under way and as the time neared for actually presenting the Federal Theatre Project collection on the World Wide Web, however, the need for "screen" or "display" access images became more evident. As will be reported in Section 13 below, the Library produced GIF images of the document pages and the online presentation features both screen-access GIF images and printer-access TIFF images.

Next Section | Previous Section | Contents

10. Hardware for rapid scanning

Rapid-scanning systems for discussion only. The Manuscript Digitization Demonstration Project included an investigation of hardware and software, with a special attempt to identify options for rapid throughput scanning of manuscript documents. The small scale of the Phase II testbed scanning activity and the high cost of many automated devices precluded the actual use of a rapid-throughput scanner in this project and thus the description of this option must be treated as provisional and preliminary. Readers should note that the survey of scanner manufacturers--which covers both non-automated and automated scanners--is covered in Appendices G, H, and I.

An analytic overview of scanner technology is provided in Appendix B, which reproduces a set of overhead slides used in presentations to the committee and to the staff of the Library's National Digital Library Program by Picture Elements. This presentation discussed the component parts of scanners and why particular types of scanners are suited to different collections. The discussion was broad-ranging and covered not only scanners for manuscripts, but also touched on equipment for brittle books, microfilm, photographs, maps, and card catalogs.

The greatest risk to paper in a high speed or automated scanner comes from the mechanical system for handling the sheets. This mechanical system includes a paper feeder, paper transport, and stacker.

Paper feeders. Automatic feeders, because of their need to perform the "de-doubling" operation and because of their tendency to turn documents around corners with high curvatures (usually to preserve stacking order or to flip for imaging the back side) are notoriously injurious to brittle documents. An alternative is for a person to hand place the pages onto a paper transport, thus avoiding all the autofeed problems. This process may introduce skew but, with sophisticated systems, skew can be detected and corrected in the early grayscale processing of the image (using bi-linear interpolation), prior to the binarization stage--without the introduction of any jaggedness.

Paper transport. The most gentle class of transport mechanisms are those that use vacuum belt transports and move documents in a straight path. As of 1995, BancTec, Photomatrix, Electrocom Automation and Image Trak produced scanners with straight-path vacuum-belt design. By using two digital cameras, no flipping of the document is required as would occur when an autofeeder is attached to a flatbed scanner. Members of the Document Digitization Evaluation Committee viewed and tested documents on both the BancTec and Photomatrix scanners and were impressed by their gentle handling.

Paper stacker. The stacker is typically a simple box with appropriate angles, padding, and static-reduction techniques to allow safe accumulation of the pages. The one drawback of the straight paper path is that the stacker ends up containing the book or manuscript in inverted page order. Most of the scanners described in the survey have an "inverter" option, which corrects the stacking order by flipping the document before placing it on the stack. This is not desirable for brittle sheets and nullifies the advantages of the otherwise straight paper path. This inverted stacking order can be fixed by an operator during a subsequent QA step on the original, while checking the original for lost, damaged, or out of order pages. Alternatively, the operator could "deal from the bottom of the deck" when placing the document pages onto the transport. As is indicated in the preceding description, the consultants and committee agreed that straight-path, vacuum-belt scanners showed great promise for handling brittle or delicate materials without damage. The document is held to the moving belt by a vacuum applied through small holes beneath the belt, thus holding the sheet on the belt. No rollers are used in this design and the paper follows a straight-line path. An example of such a transport is illustrated in Figure 10.1.

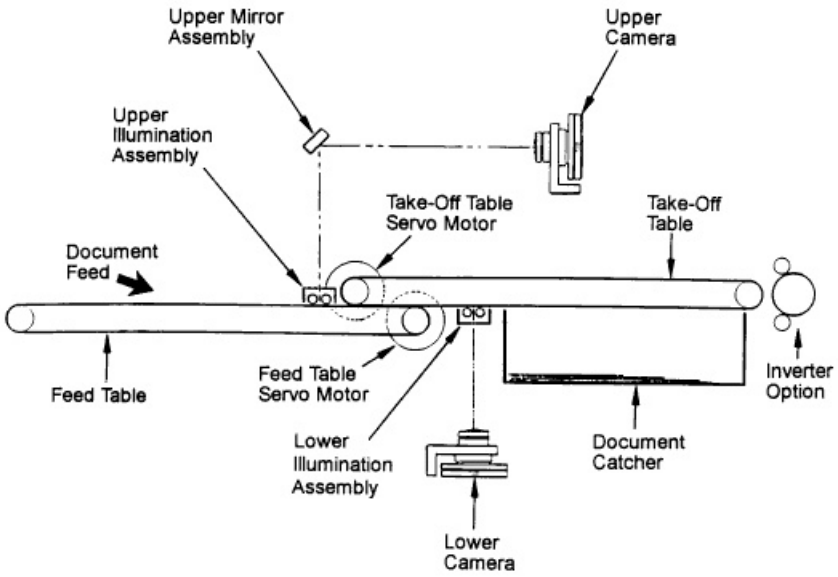


Figure 10.1. Vacuum belt transport having straight paper path. DocuScan DS-4500 Scanner Transport (Side View). Copyright 1995, BancTec. Used with permission.

Next Section | Previous Section | Contents

11. Future Image Formats and Compression

Future formats for discussion only. This project was focussed on finding the most appropriate ways to scan a testbed of 10,000 Federal Theatre Project images with readily available, broadly applicable, reasonably standard tools and techniques. Thus the consultants and the Document Digitization Evaluation Committee removed from immediate consideration several new technologies but note that these should receive more study in the future.

Object wrapper file formats. Several new file format registration schemes are developing at this time. One interesting format is Bento, developed at Apple Computer. Bento provides a descriptive metadata wrapper around objects of all sorts which unambiguously defines their type. One key type attribute for an image or other multimedia object is the software needed to view, decompress or activate it. New standards for downloading executable content through a network, like Java, could allow the (platform-independent) viewer to be stored with the object, mitigating the preservation worry about data orphaned by obsolete platforms, but raising parallel concerns about the long-term executability of the stored program. As run-anywhere languages such as Java evolve and change, it will be essential for future Java Virtual Machines to know to what version of the language an ancient viewing program adheres.

Wavelet compression. Wavelet compression is a technique based on a class of pulse-like functions, unlike the cosine functions that form the basis for the JPEG baseline compression algorithm. Printed matter involves foreground information which was formed by dragging or impressing a dark, narrow marking device on a light background, producing narrow, pulse-like strokes. Because wavelets more closely resemble the types of data seen in manuscripts and other printed matter, a compression scheme based on them could be expected to offer better compression. Wavelet compression typically offers a lower level of artifacts at a given compression ratio than JPEG does. Unfortunately, no internationally sanctioned standard existed at the time the committee was meeting and thus it did not consider the use of wavelet compression in this project.

Next Section | Previous Section | Contents

12. Phase II production plan

Image types. The plan for Phase II, the demonstration project's production phase, reflected the Phase I discussions and called for the capture of the following image types:

- **Preservation-quality images**
  Spatial resolution: 300 dpi
  Tonal resolution: 8-bit grayscale or 24-bit color
  Enhancement: contrast stretching
  Format : JFIF
  Compression algorithm: JPEG
  Compression ratio: 10:1 to 20:1

- **Access-quality images (emphasis on printability)**
  Spatial resolution: 300 dpi
  Tonal resolution: binary (1-bit)
  Enhancement: edge thresholding, despeckling (4 by 4 black and 2 by 2 white speckles removed)
  Format : TIFF
  Compression algorithm: CCITT Group 4

For the record, in a post-project activity carried out by the Library, a second set of access images was produced:

- **Access-quality images (emphasis on display access)**
  Spatial resolution: 60-100 dpi
  Tonal resolution: 4-bit
  Format : GIF

Producing the preservation images. The initial image capture of the 10,000 testbed images was carried out in an intermittent activity through 1996. The scanning took place in the Music Division at the Library of Congress. At scan time, each sheet was placed on the platen of an HP ScanJet 2cx flatbed scanner and captured at 300 dpi. If the image was captured in grayscale, it was scanned at 8-bits-per-pixel. If the operator saw any intentional color on the document, it was scanned in color at 24 bits. Backs of sheets were scanned if they contained significant (or potentially significant) marks of any kind.

Approximately 20 percent of the images were captured in color. Understanding *capture* to include finding the next sheet, placing it on the platen, scanning, removing the sheet and refiling it, and examining the image on screen, the average time to capture a grayscale image was 1 to 2 minutes and 3 to 4 minutes for a color image. It is worth noting the greater amount of time required to capture color; this could be a factor to consider when weighing the pros and cons of color capture for a large-scale production project. With the percentages of grayscale and color images indicated, and allowing for various administrative and other activities, the operator was generally able to capture approximately 200 to 300 images in an eight-hour day.

Contrast stretching. The SCANJPEG program developed to control scanning performed contrast stretching under operator guidance. After scanning and before compression, while the entire luminance image is in memory (after RGB to YCbCr conversion for color scans), a gray level histogram was calculated. It was then shown to the operator as a graph. The mouse is used to control the dark and light cutting points where no significant number of pixels have grayscale values.

In most of the scanning, both the dark and light cut points were set. The software then created a look-up table which altered each original grayscale value to stretch the used grayscale values over the full 0 to 255 range, thus brightening the image and increasing its contrast. The cut points were written into the log file for the batch. A screen image immediately showed the brightened image to the operator as it was compressed and written to disk.

Late in the scanning, it was decided that a more correct procedure was to only stretch at the white end of the scale, since a seemingly insignificant number of darker pixels (too small to register on the graph) might actually play a role in rendering tonal distinctions. Aggressive clipping at the dark end could partially eliminate some of those distinctions, slightly lowering quality.

Each image was compressed at approximately a 10:1 to 20:1 ratio using the JPEG baseline algorithm. The compression took place in software on the scanning workstation. The compressed image was then written to the workstation's hard disk. Following JPEG specifications, the color images were first converted from RGB (red-green-blue) to YCbCr (luminance, chrominance-blue, chrominance-red). These stored images constituted the set of preservation-quality images.

In batches, the preservation-quality images were recorded on writable CD-ROM disks for shipment to the contractor's facility in California. The directory naming structure followed specifications developed by the Library for the Federal Theatre Project collection curatorial staff. The names for individual files represented exposures or "pages" and these consisted of an incrementing sequence of numbers generated by the scanning software.

Producing the access images. When the batches of preservation-quality images arrived in California, they were processed to create the set of binary access images. This process begins by decompressing the JPEG preservation image in the case of the grayscale examples or, for the color images by decompressing only the luminance or Y portion of the image). Next, a sophisticated thresholding algorithm was used to create a binary access image at 300 dpi. Finally, the binary access images were compressed via the CCITT Group 4 algorithm in software and written to CD-ROM disks for delivery to the Library.

The **edge thresholding algorithm** is based on techniques developed by Picture Elements. It has been implemented in both software and high-speed hardware for integration into scanners by manufacturers. It produces very detailed and clear binary images by finding significant edges within the image and then placing black/white boundaries in the binary images at those precise edge locations. This produces highly accurate binary images with stroke widths exactly matching those in the original grayscale images. A variation incorporated into the algorithm also reproduces broad, fuzzy lines (such as those produced on copies by carbon paper) with high sensitivity.

Iterative thresholding. Picture Elements first attempted to threshold the images with a single setting, but some of the images had low contrast information, while others had higher contrast "noise" (especially the ones with onionskin paper), requiring either manual inspection and re-thresholding or automatic analysis and re-thresholding. Picture Elements chose the latter approach.

Two approaches were used to identify images with excessive noise. The first was based on the observation that images with noise tend to have a large number of very small specks and a fewer number of larger specks. Small specks can be removed from the image via speckle deletion, but large specks can be as large as fine print, so they cannot be speckle-deleted without risking information loss and therefore are more of a problem. Reducing the thresholding sensitivity tends to reduce both sizes of speckle, with the larger specks being eliminated first, so the goal is to determine how much the sensitivity needs to be reduced to make a clean image. Luckily, images don't tend to have both low contrast information and noisy backgrounds at the same time, so the highest sensitivity with acceptably low noise tends to produce the best binary image possible for the page.

If the number of small specks is low, then there are probably very few large specks in the image, and a high sensitivity can be used to pick up any low contrast information. An easy way of estimating the number of small specks in the image is to threshold the image twice, once with speckle deletion enabled and once with it disabled and comparing the compressed file sizes. If the ratio of the file sizes is close to 1 (e.g. 1.05), then there must have been a low number of small specks in the image, so there are probably very few large specks in the image. A large ratio of file sizes (e.g. 1.5 to 1) indicates that there is a lot of noise in the image and a less sensitive threshold is needed to produce a clean image.

The noisy images also had very large compressed file sizes. A manuscript page will typically have less than 4000 typewritten characters, but a noisy image can also have 100,000 noise specks. A clean image can compress to 20 KB (kilobytes), but a noisy image can be over 100 KB, with the average at around 50 KB. Decreasing the threshold sensitivity rarely causes a large percentage of the characters to drop out, but it can cause all of the noise to disappear. Therefore a test is made to see if the compressed file size is over 100 KB. A second test is to see the ratio of the file sizes between slightly differing threshold sensitivities. If the ratio is small, then reducing the sensitivity probably doesn't eliminate much noise (e.g. the noise is already gone), but could be dropping out some information, so the more sensitive threshold will be used.

Hardware thresholding and image compression allow lots of alternative settings to be automatically evaluated, starting with the most sensitive and stopping when it is determined that a reasonably clean thresholded image has been produced. In the algorithm employed, a total of 8 threshold sensitivities are available (each with and without speckle deletion), with the noisiest documents requiring the most iterations.

While the first images were processed using a software routine, most of the images were binarized using the same algorithm implemented in hardware. In hardware, the process took from less than one to a few seconds per image.

This type of thresholding algorithm is complex and too slow to run at commercially viable speed in software-only implementations (even without iteration). Hardware implementations of related (though non-iterative) algorithms are now seen in binary scanners from BancTec, IBML, Bell & Howell and Kodak.

Next Section | Previous Section | Contents

13. Phase II image examination

Image review. In early 1997, a subset of the captured images was loaded on the World Wide Web as one part of the Library's online Federal Theatre Project collection. In its final form, the online presentation will be framed by a finding aid that describes the content of the entire collection, only a portion of which has been digitized. Links to the digital reproductions retrieve both the document images created by Picture Elements in this demonstration project and a set of pictorial images made under other auspices. The Library invites end-users who consult the online Federal Theatre Project collection to inspect the images and forward their comments to the National Digital Library Program collections information email site (ndlpcoll@loc.gov).

The images created during this demonstration project received a preliminary examination at the Library upon receipt from Picture Elements. This examination was carried out by staff in the National Digital Library Program and the Music Division, custodians of the Federal Theatre Project collection.

Lack of objective measuring tools. When Library staff evaluated the Picture Elements images, they employed "informed subjective judgement." The examiners were struck by a factor that the consultants had not elucidated during Phase I: the seeming absence of objective measuring tools. Several staff members had experience in microform production and were aware of the use of targets and densitometry to measure microfilm's spatial resolution, tonal resolution, and consistency of color.

The Picture Elements staff explained that three test chart scans were made at the start of each day's scanning, two black and white (the IEEE-167A fax chart and the AIIM MS-44 test chart No. 2) and one color (the RIT Process Ink Gamut chart). Since each set of test scans did not correspond to a single directory of delivered images, but rather to a single day, these test scans were written to a single disk at the end of scanning. Appendix J contains one scan of each of these test charts. The scans of the test charts were intended as a reference to be available to future researchers into the captured images, however, rather than as an aid to the review process.

In the post-scanning discussion of measurement tools, the consultants described an approach used to measure the performance of scientific and military optical systems: Modulation Transfer Function or MTF. This method is better suited for the characterization of systems used to produce tonal images than methods based on measurements of bitonal grid patterns. Work by the Mitre Corporation originally performed to assist the FBI's characterization of fingerprint scanners has raised the interest of the document imaging community in MTF measurement. The calibrated, continuous-tone target suitable for MTF measurement and the corollary software that would have to be written or adapted to make the measurements was not budgeted for this project.

In a separate consultancy, James Reilly and Franziska Frey of the Image Permanence Institute (IPI) reported to the Library that, indeed, there are not ready means for lay workers to objectively measure tonal digital images. (Properly equipped engineers can measure aspects of images using laboratory equipment.) The IPI report is *Recommendations for the Evaluation of Digital Images Produced from Photographic, Micrographic, and Various Paper Formats.* In a follow-on consultancy, Reilly and Frey will develop an approach and a toolset for the Library to use in judging grayscale digital images.

Preservation-quality images: spatial and tonal resolution. The preliminary examination indicated that the selection of 300 dpi as the level of spatial resolution appeared to be satisfactory. There were no instances in which significant features on original documents were lost. The examination, however, raised some questions about the clarity of some images and the discussion of this matter highlighted the importance of tonal resolution (the quantity and distinguishability of tones) in the images. For binary images, spatial resolution is the key factor for capture of fine detail; with tonal images *both* spatial and tonal resolution are important.

Preservation-quality images: distribution of tones and image clarity. The following two images are examples of the type that seemed to show loss of clarity when compared to the original documents. In these digital image, for example, the opening in several of the *a* characters was filled or partially filled. In the paper originals, the *a* was open or relatively open. In addition, the overall appearance of the digital image was dark, in the words of one Library staff member, "as if a color slide had been underexposed by a half stop."

- Carbon copy on onion skin paper example A (646 KB)
- Carbon copy on onion skin paper example B (618 KB)

As these and similar images were examined, Library staff asked whether capturing at a setting that yielded dark-looking images might not have contributed to the loss of clarity, just as excessive inking in letterpress printing might have closed up openings in a letter like *a*. The person who examined the greatest number of images observed that, for relatively "clean" documents, the darker image tone caused a number of stray or irrelevant marks to become visible, leading to the related question of whether these would have benefited from a "lighter touch."

- Clean document with blurred typed keystrokes (819 KB)

Prompted by these samples, the Library team asked, "Should the contrast stretching applied to the images have produced a lighter value for the paper relative to the density of the strokes?" and "Would lighter values overall have provided a better reproduction of the letter *a*?"

Picture Elements offered a two-part response. First, an observer should never expect a reproduction to be identical to an original. And the perception of the original can be influenced by subtle viewing factors. For example, changing the angle of the original paper page in the light, he said, can make a significant difference in one's ability to read poor-quality information. In addition, it is always possible--as is the case with microform--that some researchers will have to examine some original documents in order to answer all of their questions.

Second, Picture Elements reminded the Library that in Phase I, the discussion of contrast stretching had focussed on the rescue of illegible documents, not aesthetic cleanup. The sample that had received the committee's attention was a sheet of dark red paper printed with black ink. The reason to stretch contrast is to avoid leaving significant portions of the tonal range unused in the case of images whose initial representation occupies only a small portion of that range. The algorithm Picture Elements used for contrast stretching makes the *lightest* pixels in the image the highest value of white but does not raise all of the lighter pixels to this same value. If all lighter values were raised then the highest values would be "clipped," i.e., rendered the same as lower values, and some information about tonality in the paper texture would be lost. The Picture Elements contrast-stretching algorithm used to capture the preservation-quality images permits a hardware implementation that will analyze the image histogram during scanning and make adjustments without human intervention. In order to capture **faint** keystrokes, it is necessary to use dark settings. As a result, the setting may not yield the most aesthetically pleasing tonal archival images.

The Library had noted that some aesthetic improvement was seen when the two images were opened in a graphic-arts software and the brightness and contrast were increased. Picture Elements responded that, when this was done in a manner that clipped nearly 50 percent of the pixels, distracting dark patches remained on the background of the page. In order to achieve an aesthetically pleasing effect, 90 percent of the pixels had to be clipped, which would be ill-advised during the production of preservation-quality images, while it might be entirely appropriate (and still an option) for access-quality images.

Finally, Picture Elements noted that the contrast stretching provided in these two examples was sufficient to permit the binarization algorithm to work when the access images were produced (see discussion of access images below).

- Binary access image, carbon copy on onion skin paper example A (39 KB)

The preservation-quality images were produced in a manner that would minimize the loss of the image content of the original. That is why contrast stretching was limited to a "semi-reversible" amount, with the lightest pixels mapped to full white, but with no saturation (where a range of relatively light pixels would be mapped irretrievably to full white).

Improving the visual quality of the originals (e.g. reducing the visibility of "stray or irrelevant marks") would introduce some loss, which should not be done in the preservation image. Any lossy processing should be done in the production of some form of access image. The goal of the preservation-quality image, Picture Elements asserted, is to only have to scan the original once, picking up all of the information that is technically feasible. Any intentional loss of visible information should be incorporated in the production of some class of access images intended for a certain use.

Preservation-quality images: color and color fidelity. The examination of images did not return in a systematic way to the most intriguing aspect of color reproduction: when is it needed? The image reviewers were drawn to color reproduction: having chromatic cues on the page makes it easier to spot the script lines the stage director underlined in red, but no one has stated that this determination could not have been made from a monochrome image (like the binary access image of the same page). But the reviewers did not encounter grayscale examples where they wished that capture had been in color. The Library looks forward to hearing comments on this question from researchers who use the collection.

The examination of color images identified some examples with a greenish cast that had not been present in the paper originals. In others, red values did not closely resemble the shades of red in the originals.

- Color image with color-value shifted red markings (608 KB)

After receiving the Library's review, Picture Elements re-examined this images. The consultant first isolated a small portion of the "overscan" at the bottom. This area reproduces the white color of the scanner lid and not the particular shade of the paper. The red, green, and blue (RGB) values for the overscan-area sample were R=220, G=231, and B=215 (average=222, delta +/- 4 percent). For a perfect white, the RGB values should be the same and should come closer to 255; an evenly weighted 222 should look like a color-free light gray. But the difference in RGB values, although slight, favored green and gave the image a slight but noticeable color cast.

This observation prompted Picture Elements to review the color test images made for Phase I of the project. This review revealed that some of the images appeared to have an subtle increasing green shift as one moved from the top to the bottom of the page.

- Color image with greenish cast from Phase I investigation
- Color image with greenish cast from Phase I investigation

Picture Elements was unable to determine the cause for the color shift, speculating that the lamp might be changing color very slightly during the scan. In an echo of laundry detergent advertisements, the consultants noted that it is very difficult to obtain white whites

when producing color images of documents from a CCD sensor consisting to three separate color channels.

It is worth noting that the challenges associated with the maintenance of color fidelity in digital systems have been reported in literature associated with the printing industry, which today relies heavily on computerized design and production. Several recent trade journal articles have highlighted the need for better color management systems and the IPI report on objective measures for digital images skirts the issue of color.[6] Thus there was no expectation that this demonstration project would wrestle with color fidelity in a conclusive way. Color capture was addressed to explore the question of when the added capture time and storage requirements for color would be warranted in a manuscript digitization project.

Preservation-quality images: black background sheet to reduce print-through. Early in the production phase, Picture Elements staff noted that many documents were on onion-skin or other thin paper and that writing on the backs of these sheets was visible in the scans.

This effect was increased by the white lid of the scanner lid; light passed through the sheet and bounced back from the lid and again through the paper, thus making the back-side writing more visible. In an attempt to reduce the visibility of the print-through, black paper was placed between the sheet and scanner. This action, however, had the effect of heightening the visibility of the onion-skin paper's texture: the "valleys" in the paper were darkened where they pressed against the black sheet, increasing the difference in tone from the paper-texture "hills." This increase in the visible paper texture was slightly bothersome in the tonal preservation-quality images and led to problems when the binarization algorithm was applied. In some circumstances, the texture could be mistaken for meaningful marks on the sheet. After these effects were observed, use of the black backing sheet ceased.

- Example with white backing; note visibility of rubber stamp on reverse at bottom center of the image (745 KB)
- Example with black backing sheet; note reduction in rubber stamp visibility and increase in paper texture (including watermark in original paper) (1.9 MB)

Preservation-quality images: reproducing the entire sheet. The scanner employed in this project had an active area of 8 1/2x14 inches. This meant that it was not possible to scan beyond the left and right edges of 8 1/2x11-inch typing-paper-size documents or beyond any edge of legal-size documents. Thus the project could not adopt the preservation-microfilming convention of showing that the entire items has been captured by displaying the full sheet of paper.

Access images: clean appearance and printability. In some cases, the lack of clarity in the characters as rendered on onionskin paper carbon copies robs the binary images of legibility. But for many of these images, when they are printed on a laser printer, the paper copies present a very clean appearance. In contrast, the printed copies of the tonal images have an overall gray cast produced by the effect of halftoning at print time; the background tonal values ("the paper") printed with a light pattern of dots. (At the time of these inspections, the Library did not have at hand a special printer capable of rendering tonal images on paper.)

Had a different posture been taken at the outset of the project, the production of a variety of special-purpose access copies could have been planned for. For instance, in addition to a binary print access image, the production of the following derivative image types is possible (thanks largely to the high fidelity of the tonal preservation-quality image):

- a tonal image designed to print more cleanly;
- a tonal image designed for casual network-constrained screen access (perhaps using JBIG 2, 3, or 4 bit grayscale), with whiter whites;
- a binary image specially tuned to the lightening/darkening characteristics (TRC or tonal reproduction curve) of a specific manufacturer's print engine;
- a tonal image brightened and color twisted to match the specific gamma and color profile of a specific manufacturer's monitor or color printer.

Picture Elements agreed that multiple access images are desirable, and that their particulars will change with time and amongst users. In the future, they will be built instantly on demand by software associated with online delivery, using the high-quality preservation image as a source.

Access images: information loss. Some information was lost in the binarization process. The collection's curator offered the strongest statements on this topic, pointing out that some dim or subtle writings, including erased pencilled notes, were visible in the tonal images and not in the binary derivatives. If he were a researcher consulting the collection online, the curator said, he would wish to have access to the preservation-quality images. A similar benefit would follow from using tonal access-quality images. Thresholding has an inherent tradeoff between producing a clean, speckle-reduced image and losing the faintest markings. Reducing speckles is necessary for two reasons: because the tonal reproduction curve (TRC) of printers gives undue weight to small black specks (due to toner spread), and because the T.6 compression algorithm performs extremely poorly when they are present.

- Sample A, grayscale image (479 KB)
- Sample A, binary image (21 KB)
- Sample B, color image (645 KB)
- Sample B, binary image (65 KB)

In the post-production discussion of these two examples, Picture Elements described the binarization process and how it made use of "iterative thresholding," the automated process by which the thresholding sensitivity is selected. The writing in sample A (image 0003a.jpg) was lost, the consultant said, due to the iterative thresholding choosing a sensitivity that was low enough to clean up the noise in the image. An examination of the log that was produced when this image was processed shows that the iterative thresholder evaluated four decreasing sensitivities that produced compressed image sizes of 44 KB (kilobytes), 30 KB, 21 KB, and 18 KB respectively. The ratio of the difference in file sizes between 30 KB and 21 KB was large enough that it was likely to be caused by noise, but the difference between 21 KB and 18 KB was small enough that it was not worth the potential information loss associated with the smaller image, so the 21 KByte image was automatically chosen. The tradeoff with the automatic iterative thresholding algorithm is between having a larger average file size (and noisier images) versus accepting some dropout when there is a lot of noise in the image that has the same contrast as some of the significant information.

Picture Elements pointed out that a key benefit of iterative thresholding is that it is completely automatic. Putting a person in the loop (who could better discriminate between information and noise) would probably have improved this image but at the expense of more noise and a larger file size and increased labor costs. In a high volume production line, however, the differences under discussion here might not have been noticed even in such a manual approach, especially if the operator was less than fully attentive.

Picture Elements reported that sample B (image 0010a.jpg) represented an example in which the iterative thresholding produced an optimum result by showing a hint of the erased writing. This is preferable, he said, to picking up all of the erased writing which would fail to signal that the erasure had occurred. It is true that if someone needs to investigate the erasure, the tonal image will need to be examined. In this case, the iterative thresholding process evaluated two decreasing sensitivities that produced compressed file sizes of 71 KB and 60 KB. The ratio of compressed file sizes was small enough that it was unlikely that the image had excessive noise, resulting in the sensitivity that produced a 71 KB image file being chosen: this picked up all of the writing and the "hint" of erased writing.

Access images: the WWW environment. As this project proceeded and as the time neared for presenting the Federal Theatre Project collection on the World Wide Web, the Library saw that the binary access images would not serve for "screen" or "display" access and navigation. Meanwhile, during 1996, a number of libraries began demonstrating effective presentations of document image sets in WWW browser software; the Library of Congress soon followed suit.[7] These presentations exploit tonal images in the GIF format, structured to permit end users to page through a multi-page document. The same approach can be used with JPEG files, including progessive JPEG.

One portion of the Library's online presentation of the Federal Theatre Project collection includes both the preservation-quality and access-quality document images created by Picture Elements during this demonstration project. In order to offer paging navigation, the Picture Elements images have been placed in browser-based paging displays like those cited above. As end-users page through a GIF-image paging set, they can call up the Picture Elements-produced images. In effect, there will be two access images: *print-access* images in the TIFF format with Group 4 compression (produced by Picture Elements during the demonstration project) and tonal *screen-access* images in the GIF format (produced by the Library after completion of the demonstration project).

In order to work well on today's display screens and to be small enough for easy transmission on the Internet, the spatial and tonal resolution of the GIF images has been reduced from the originals. The GIF images are at about 60 dpi with a tonality of 4 bits per pixel.

Access images: types will change over time. At the end of the project, the project planners and consultants reflected on the apparent need to produce a set of three images for each Federal Theatre Project collection document. Tonal preservation-quality images provide the most faithful reproduction of the original document but are very large and thus cumbersome to view and print. Binary printing-access images offer clean printouts and are conveniently small but may suffer some information loss. Tonal display-access images open easily in WWW browser software and can be used to navigate through a paging set but their reduced resolution means they offer an imperfect reproduction of the original document.

Archiving the highest-quality image is sensible and necessary. But are the two access image types necessary? Certainly, changes in technology will lead to changed practice. Some online projects have demonstrated ways to create GIF images for end users on the fly; adoption of this practice would eliminate the need to produce and store such files. Greater network bandwidth and faster computers in the future will make it easier to display the preservation-quality images and reduce the need to have separate display-access images on hand. Improved printing options may make it easier for end-users sitting at desktop computers to to print a tonal image thus reducing the need for binary access images. Other more advanced technologies, further in the future, may make it possible for institutions to produce many types of derivative images on the fly to meet the end-user's immediate needs by dynamically processing the archival version of the images.

Next Section | Previous Section | Contents

14. Project findings

The findings of the Manuscript Digitization Demonstration Project may be summarized as follows:

- Tonal (grayscale and color) images are necessary for preservation-quality reproduction of many manuscripts.
- A spatial resolution of 300 dpi for tonal images is sufficient to capture all information in twentieth century materials like typescripts.
- Some documents benefit from color reproduction but ambiguity remains about when color reproduction is necessary to serve researchers.
- For preservation-quality images of routine documents--especially in the setting of a high-volume digitization effort--the *benefit* of reduced file size resulting from the modest application of lossy compression (at around 10:1 for JPEG) can be argued to outweigh the *cost* of the slight image degradation that results.
- Access images can be produced by reprocessing the preservation-quality images.
- If end users are to be provided with easy navigation in the World Wide Web, "browser-capable" tonal display-access images must be provided.
- If end users are to be provided with clean printouts from a laser printer, separate binary print-access images should be considered; laser printouts from the reduced-size tonal display-access images will be less legible.
- Objective tools for quality review are needed.

The Library welcomes comments on the Manuscript Digitization Demonstration Project, especially responses from researchers who use the testbed images. Comments may be sent to ndlpcoll@loc.gov. The Library looks forward to carrying out additional projects that will further refine the issues described here and hopes that others will also continue to investigate these and related matters. The Library's staff stand ready to consult with interested parties undertaking such investigations.

Previous Section | Contents

References

Preserving Digital Information. Report of the Task Force on Archiving of Digital Information. Commissioned by The Commission on Preservation and Access and
The Research Libraries Group, Inc. May 20, 1996. URL: http://www.rlg.org/ArchTF/index.html.

Brown, C. Wayne and Barry Shepard. *Graphics File Formats: Reference and Guide.* Boulder: Manning Publications,1994.

Fraser, Bruce. "Color Under Control," in *Adobe Magazine*, September/October 1995, pp. 41-45.

Hou, H. S. *Digital Document Processing.* New York: Wiley, 1983.

Kay, David C., *Graphics File Formats, 2nd ed.* New York: Windcrest/McGraw-Hill, 1995.

Kenney, Anne R. and Stephen Chapman, *Tutorial. Digital Resolution Requirement for Replacing Text-Based Material: Methods for Benchmarking Image Quality*. Commission on Preservation and Access, 1995.

Kenney, Anne R. and Stephen Chapman, *Digital Imaging for Libraries and Archives*. Department of Preservation and Conservation, Cornell University Library, 1996.

Kenny, Anne R. and Lynne K. Personius, *Joint Study in Digital Preservation. Report: Phase I*. Cornell / Xerox / Commission on Preservation and Access January 1990-December 199l. URL: http://www.clir.org/cpa/reports/joint

Pennebaker, William B. and Joan L. Mitchell. *JPEG Still Image Data Compression Standard.* New York: Van Nostrand Reinhold, 1992.

Rabbani, Majid and Paul W. Jones. *Digital Image Compression Techniques.* Bellingham: SPIE Optical Engineering Press, 1991.

Reilly, James and Franziska Frey. *Recommendations for the Evaluation of Digital Images Produced from Photographic, Micrographic, and Various Paper Formats.* Report from the Image Permanence Institute, Rochester, NY, to the Library of Congress, 1995. URL: http://lcweb2.loc.gov/ammem/ipirpt.html.

Rodney, Andrew. "Desktop Color Management," in *Photo-Electronic Imaging*, January 1997, pp. 40-42.

Spielman, Frankie R. and Louis H. Sharpe. *Raster Graphics: A Tutorial and Implementation Guide.* NISTIR 5108. Gaithersburg: National Institute for Standards and Technology, January 1993.

Strong, William S. *The Copyright Book: a Practical Guide, 4th ed.* Cambridge: MIT Press, 1993.

Ulichney, Robert. *Digital Halftoning.* Cambridge. Mass. : MIT Press, 1987.

Written, Ian. H., Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* New York: Van Nostrand Reinhold, 1994.

Contents

Appendix A
Sample Images

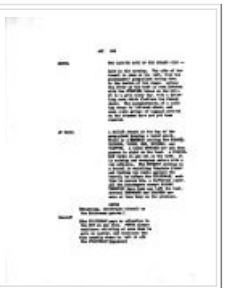
A.1 Phase I Sample Visual Catalog

Here are all the Phase I Sample Items shown in the 300 dpi, contrast stretched, JPEG compressed rendition used for Phase II scanning. Progressive JPEG thumbnails are shown. When a thumbnail is clicked, a progressive JPEG screen width image is shown.

A detailed analysis of each Sample Item (typically a manuscript page) is given in the section referenced beneath its corresponding thumbnail. Access to the full size JPEG version of each item is available from there.

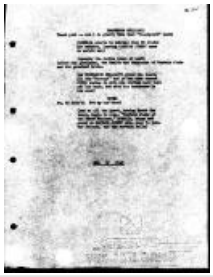
A.1.1 **Sample Item AP1**

A.1.2 **Sample Item CJ1**



A.1.3 **Sample Item CJ2**

A.1.4 **Sample Item CJ3**



A.1.5 **Sample Item CJ4**

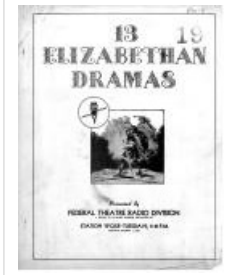
A.1.6 **Sample Item DJ1**

A.1.7 **Sample Item DM1**

A.1.8 **Sample Item DM2**

A.1.9 **Sample Item DM3**

A.1.10 **Sample Item DM4**



A.1.11 **Sample Item ED1**

A.1.12 **Sample Item ED2**

A.1.13 **Sample Item LS1**

A.1.14 **Sample Item LS2**

A.1.15 **Sample Item NY1**

A.1.16 **Sample Item NY2**

A.1.17 **Sample Item OS1**

A.1.18 **Sample Item PD1**

A.1.19 **Sample Item PD2**

A.1.20 **Sample Item SS2**

A.1.21 **Sample Item SS3**

A.1.22 **Sample Item SS4**

A.1.23 **Sample Item SS5**

A.1.24 **Sample Item SS6**

Contents

Aspects of High-Speed Scanning

Presentation
for the
National Digital Library Program

18 May 1995

Lou Sharpe
Picture Elements, Inc.

303-444-6767
lsharpe@picturel.com
http://www.picturel.com

Click here to start

Table of Contents

Contents

Anatomy of a High Speed Scanner

- **Feeder**
- **Transport**
- **Camera**
- **Illumination**
- **Stacker**
- **Interfaces**
- **Accessories**

Previous slide Next slide Back to first slide

Feeders

- **De-doubler**
- **Roller or Belt**
- **Vacuum or not**

Previous slide Next slide Back to first slide

Transport Types

- **Roller**
- **Belt**
- **Vacuum Belt**
- **None: flatbed (paper stationary)**
- **None: digital camera (paper stationary)**
- **None: book cradle type**

Previous slide Next slide Back to first slide

Camera Types

- ## Line-scan
  - Any length, fixed width
  - Easily can “stitch” camera for large widths
- ## Area scan
  - Fixed width, fixed aspect ratio
  - Awkward to match to item format
  - “Tiling” of multiple scans re-composes poorly
  - More lens distortion than line scan

Previous slide Next slide Back to first slide

Camera Data Quality

- **Tonal range - gray scales: 8, 10 or 12 bit**
- **Tonal linearity**
- **Number of photosites: optical resolution**
- **Good lens, flat field**
- **Low noise**
- **Good thresholding for binary imaging**

Previous slide Next slide Back to first slide

Illumination

- **Flat across field**
- **No streaks or dark spots**
- **Good color rendition: white light or green**
- **Can be incandescent or fluorescent**

Previous slide Next slide Back to first slide

Stackers

- **Can preserve or reverse original order**
- **Multiple pocket**

Previous slide Next slide Back to first slide

Classes of Scanner Types

- **Page Scanners**
- **Large Document Scanners**
- **Microfilm Scanners**
- **Digital Cameras**
- **Book Scanners**
- **Specialty Scanners**

Previous slide Next slide Back to first slide

Page Scanners

- ## Flat beds
  - Hand-placed (gentle, slow)
  - Good alignment, but must flip page over
  - High accuracy: no slipping, paper not driven
  - Often have sheet-feed option
  - Examples: Fujitsu, HP, Microtek, Ricoh, Epson
- ## Roller Drive
  - Paper driven by rollers
  - Simple mechanical design
  - Short paper paths
  - Prone to skew
  - Example: Ricoh 510/520

Previous slide Next slide Back to first slide

Page Scanners 2

- ## Belt Drive
  - Often return document to user
  - Good for production, but not for most fragile
  - Examples: Bell and Howell
- ## Vacuum Belt Drive
  - Can have return path or straight path
  - Straight best for fragile documents
  - Examples: TDC 4000 series, ElectroCom, ImageTrak Photomatrix, ScanOptics

Previous slide Next slide Back to first slide

Page Scanners 3

- ## Diagram of Vacuum Belt Drive

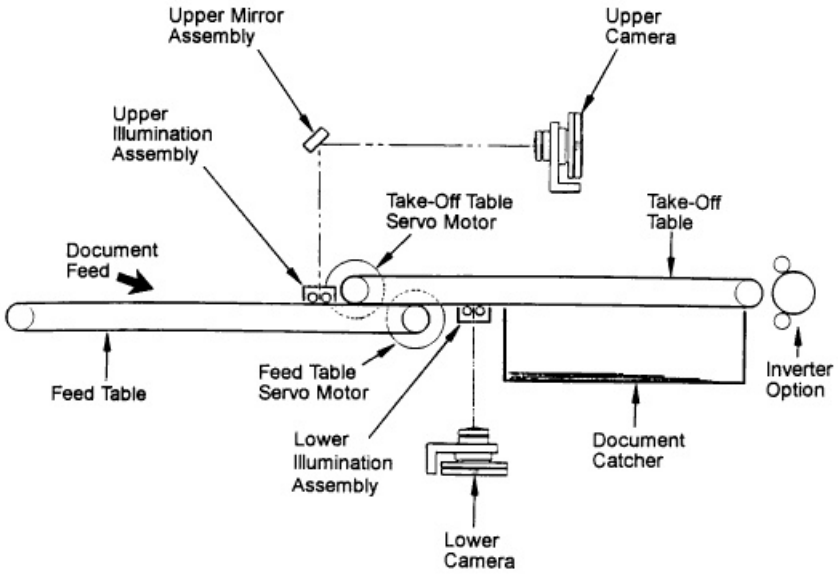


Diagram of BancTec 4500 Transport courtesy of BancTec, Inc.

Previous slide Next slide Back to first slide

Book Scanners

- **Minolta DS-3000**
- **Needs work to produce JPEG**
- **Needs work to produce 600 dpi binary**

Previous slide Next slide Back to first slide

Overhead-Style Digital Cameras

- **“Camera on a stick”**
- **Distortion problems**
- **Fixed number of pixels**
- **Expensive, especially if no bad pixels**
- **Examples: Kontron, Kodak, AVX, Dalsa, Phase One**

Previous slide Next slide Back to first slide

Large Document Scanners

- **Maps, engineering drawings, posters**
- **Examples: Vidar, Scangraphics**

Previous slide Next slide Back to first slide

Specialty Scanners

- ## Sheet film
  - originally designed for X-ray or aerial photo scanning
  - Vidar, Vexcel
- ## Microfiche and Microfilm
- ## Cinematic Film
  - Kodak Cineon
- ## Card scanners
  - card catalogs
  - idea: adapt a bank check scanner

Previous slide Next slide Back to first slide

Advanced Features

- **Grayscale output with JPEG compressed data**
- **Grayscale and binary simultaneously**
- **Auto-duplex**
  - ignore blank backs
- **Automatic image quality assurance**
  - check: skew, dog-ears, blank, upside-down, speckles
- **Automatic skew detect and correct**
  - best quality if done in grayscale

Previous slide Next slide Back to first slide

Input Subsystems

- New trend for production scanning architectures
- Permit fast, flexible processing

- **Cornerstone InputAccel**
- **Kofax Ascend**
- **Intrafed**

Previous slide Next slide Back to first slide

Thorny Issues

- **Exception Item Handling**
- **Indexing**
- **Document Preparation**
- **Quality Control**

Previous slide Next slide Back to first slide

Scan, Then COM vs. Film, Then Scan

- **Generational loss in each step**
- **COM output is “perfect film”**
- **COM can handle grayscale and binary**
- **If paper is available, don’t scan the film**
- **Microfilm scanners best used when paper gone**

Previous slide Next slide Back to first slide

In-House Competing Service Bureaus

- **Sublet space yearly to pool of bidders**
- **Award 5 each year, recompete yearly**
- **Give adjacent space within LC to awardees**
- **LC groups bring individual jobs to any/all for bids**
- **Each can specialize their equipment/procedures**
  - e.g. large documents, microfilm, etc.
- **Low risk to collections**
- **Quick correction of failures: see early results**
- **Strong competition, low prices**

Previous slide Back to first slide

Appendix C
Image Formats Survey Questionnaire

Library of Congress
Preservation Office

Digital Imaging for Manuscript Preservation
A Survey of the Field

14 December 1994

NOTE: Please return completed questionnaires directly to the contact person listed at the end of this message, not as a reply to the mailing list from which you received it. Replies are requested by 31 January 1995. Partial submissions will be gladly accepted. Please note the offer of useful utilities (carrots!) down in the Instructions section.

Purpose

The Library of Congress wishes to learn about existing practice among archives, libraries and the wider commercial marketplace for digital imaging of documents, especially the types of documents found in manuscript collections. This information, together with technical studies undertaken by the Library, will be used to develop approaches for future digital conversion efforts.

General Background

As the Library of Congress continues to develop its capabilities for providing computerized access to its collections, it must address a wide array of issues as well as identifying and testing a broad range of tools and techniques, especially those which will assure that digital imaging can be used successfully within the institution's preservation programs.

In order to address some of those preservation issues in the particular case of manuscript and document collections, the Library has engaged Picture Elements, Inc., to carry out one or more surveys in order to determine and/or identify:

- the most appropriate image formats for manuscript conversion projects
- available software and hardware tools for image enhancement
- available software and hardware tools for efficient throughput

This survey is a part of that effort. A later demonstration conversion project is also planned.

Detailed Background

The focus of this demonstration project will be on documents consisting of unbound, separate handwritten or typed sheets of 8.5 inch by 14 inch or smaller paper -- what might be considered to be typical manuscript documents.

A key issue for the Library is finding the most judicious balance between **conserving** precious original documents--protecting them from damage--and achieving a reasonably rapid rate of **conversion**. The outcomes of this project are expected to assist the Library in designing models for further conversion applications for the Library's collections.

The Library foresees the need for at least two types of images that reproduce typical manuscript-collection documents. One image, proposed for consideration as a potential digital preservation-quality image, will have high quality and offer a faithful copy of the original.

The Library also seeks to create smaller-sized images in addition to the preservation-quality image. These will be used in end-user retrieval systems, especially those accessed via computer networks, including Internet. Smaller-sized or access-quality images can be more easily handled in such systems. The Library would like to identify a practical level of quality that, although less faithful than the preservation-quality image, offers high legibility and good service to researchers.

Instructions

Please complete this questionnaire if your organization has or is planning a project involving preservation of manuscript or other primarily textual documents using digital imaging.

It may be that you have no such project, but have opinions or policies on this topic. Or, you may find a questionnaire format confining. You may not have time to address the entire set of questions. In these cases, please feel free to provide any information with a form and content you feel appropriate, using the questionnaire as a guide to our issues of interest. Comments may be inserted in-line into the questionnaire or attached. When presented with a list, multiple answers will often be appropriate.

You may reply in any format by contacting Lou Sharpe of Picture Elements, Inc. directly at **lsharpe@picturel.com** or by phone at 303-444-6767 or by fax at 303-415-1392.

Please complete the General Questions section below. Then proceed to special questions for Archivists and special questions for Technologists. You need not answer every question; feel free to offer some replies in both sections.

In exchange for your returned questionnaire, we would like to offer you two useful public domain utilities for checking the format of image files. TIFFLOOK dumps TIFF files and JPEGINFO dumps JPEG Interchange Format (JPG) or JPEG File Interchange Format (JFIF) files. Please

indicate your desire for further information on these utilities on your returned questionnaire or obtain them from the Picturel Elements web site at: http://www.picturel.com.

General Questions

1. Contact Information

Name _____
Organization _____
Address _____
_____
_____
Email address _____
Phone _____

2. Name of Project or Department

_____

3. Nature of Your Organization

archive ___
library ___
commercial company ___
government agency ___
other (please specify) _____

4. Materials Being Digitized

4a. physical form
loose pages ___
bound volumes ___
other (please specify) _____

4b. content types
typewritten ___
handwritten ___
engravings ___
lithographs ___
other (please specify) _____

Archivist or Curatorial Questions

5. Do you create digital images of

manuscript papers ___
printed matter ___
handwritten items ___
other types of documents ___
(please specify)? _____

6. Do you consider these copies to be for

access ___
preservation surrogate ___
document delivery ___
republication ___
transcriptions ___
optical character recognition ___
a mix of the above ___
other (please specify)? _____

7. What is the total number of images scanned in your project to date?

8. If you create digital images with preservation as a goal, do you discard or retain the original paper item?

discard ___
retain ___
other (please specify)? _____

9. Regarding microfilming of the items being digitized, do you

microfilm in parallel ___
scan from microfilm ___
output digital image to an
electron beam film recorder ___
other (please specify)? _____

10. What are your approaches to retrieval?

catalog ___
non-bibliographic database ___
directory ___
register ___
SGML-tagged register ___
searchable full texts ___
other (please specify) _____

11. Do you use the image file header content for searching and retrieval?

12. Do you use any special approaches to protect or authenticate images?

encryption ___
authentication ___
watermarks ___
hidden watermarks ___
other (please specify) _____

13. Regarding the rapid and efficient capture of images:

Do you use a sheet-feed or other device? ___

Do you use a book-edge or other special scanner? ___

What is the approximate number of images that your
conversion facility can capture per hour or day? _____

How many staff and scanners provide this total throughput? _____

14. Have any of your documents been damaged during the capture process? Please give details, if possible.

15. Do you capture any items while they are sleeved in Mylar?

16. Have you been forced into workarounds by special problems, for example:

thin paper bleedthrough requiring special image processing,
image quality problems forcing transcription?

Technical Imaging Questions

17. Do you create more than one type of digital image, e.g., a preservation image and an access image? Why? How do they differ in terms of the below three sections (image characteristics, compression techniques, file formats)?

Image characteristics used

18. Please provide technical information on the image types you create, including:

spatial resolution as delivered (dots per inch or millimeter)

actual optical resolution (dots per inch or millimeter)

tonal-depth resolution (number of shades or colors or bits per pixel)

In this regard, are you aware of whether the scanning subsystem converts from the actual optical resolution to the delivered resolution? Do you know what technique is used for this process (for example pixel replication/deletion or linear interpolation)?

Compression techniques used

19. Please indicate the compression techniques used.

CCITT T.6/Group 4 ___
CCITT T.4/Group 3 ___
JPEG ___
JBIG ___
LZW ___
other (please specify) _____

File formats used

20. Please indicate the image file formats used.

TIFF vs. 6.00 ___
TIFF vs. 5.00 ___
ODA or ANSI/AIIM MS-53 ___

JPEG Interchange Format ___
JFIF ___
other (please specify) _____

File header or trailer information fields

21. For the named file formats, what header fields or tags are used? Please provide a list, giving the tag or element number. Can you provide a text dump of one of your files? Do they conform with any identifiable subsets of file formats (e.g. TIFF Class B or RFC 1314)?

22. Do you place identifying information in the header, e.g., a code number for the image, the name of your organization, or a title or subject term?

Scanners used

23. What primary scanner was used for capture (manufacturer and model)? Was it modified or customized in any fashion?

24. Was more than one scanner type used? Are any documents routed to a specialized scanner having different capture characteristics?

Special image processing used

25. What image processing or image enhancement approaches have you found helpful?

26. Do you apply de-skewing or border cropping techniques to your images?

27. Does your approach result in bitonal (one-bit-per-pixel) images?

28. Does your system employ special forms of threshholding, density control, or contrast and brightness management?

Database issues

29. How do you link images to the retrieval tool? Do you link directly by pathname/filename, use a look-up table, use an identifier in the header, or some other approach?

30. What file or directory naming conventions do you use? Are these techniques used to link images to other records?

31. What indexing means is used to link documents and image files to bibliographic records or other search tools?

32. Do you do more or less indexing work for materials being scanned as compared to traditional materials?

Access issues

33. What are the intended uses of your images? Are they for preservation only, for screen access over local area networks, for wide area network access, or for local printing?

Storage

34. Does your institution have an approach for the preservation of the digital data represented by the images? Please provide a brief statement.

35. Is there a policy on migration of data to newer media as time progresses?

36. Is there a policy on the monitoring of error correction rates or for random sampling of seldom used collections?

37. What media are used?

38. What is the average image size? If both preservation and access images are stored, please indicate the average size for each type.

Standards

39. To what extent are standards issues key to your approach to digital imaging? Do you believe de facto or de jure standards should be used?

Quality assurance

40. What level of quality assurance is used?

41. Is visual inspection used? On what percentage of scans?

42. Is automatic quality assurance used?

Document preparation

43. How do you prepare documents for scanning?

44. Do you separate different material types or keep them together in the workflow?

45. Are any special steps taken in the physical preparation for scanning, such as

disbinding ___
guillotining ___
fastener removal ___
other (please specify)? _____

46. How much time does each of these steps consume?

Futures

47. What breakthroughs in imaging technology would help you most?

good microfilm scanner ___
high-end book scanner ___
face-up book cradle ___
page turning device ___
high-speed input subsystem ___
automatic quality assurance ___
preservation file format ___
other (please specify) _____

Contact Information

For further information, to convey answers verbally, or to discuss any of the questions in more detail, please contact Picture Elements directly.

Louis H. Sharpe, II
Picture Elements, Inc.
410 22nd Street
Boulder, CO 80302
303-444-6767
303-415-1392 fax
lsharpe@picturel.com

Contents

Appendix D
Image Enhancement Questionnaire

This appendix shows the image enhancement survey questionnaire which was circulated.

Survey of Image Enhancement Tools for Manuscript Collections

This survey is divided into three sections: General Information; Interactive Image Processing Tools; and Non-interactive (Automatic) Image Processing Tools. Due to the diversity of tools being surveyed, many of the questions may not be applicable to your product (answer N/A). If you don't have time to fill out the survey form, please call me instead, or send me your standard product literature.

GENERAL INFORMATION

Primary Contact _____ Telephone _____

Company Name _____

Address _____

_____

_____

Does your company currently manufacture image processing tools (hardware or software) suitable for use in the enhancement (quality improvement and/or file size reduction) of scanned manuscripts?

___ Yes ___ No Comments _____

If selected by the Library Committee, will it be possible to arrange a demonstration of your products at the Library of Congress?

___ Yes ___ No Comments _____

Does your company offer a demo disk or evaluation copy?

___ Yes (___ PC, ___ Macintosh, ___ SUN, ___ Other Unix) ___ No

Could a video tape of a demonstration of your products be provided as an alternate form of demonstration?

___ Yes ___ No Comments _____

What particular strengths do your company's products have that are suited to the Library's needs?

_____

Any other information or suggestions?

_____

INTERACTIVE IMAGE PROCESSING TOOLS

Which of your products would you recommend for the interactive enhancement of scanned manuscript images at the Library of Congress?

(For multiple recommendations, use duplicate copies of this form).

Product _____ Description _____ Price _____

What computer environment is this product compatible with?

___ PC, DOS

___ PC, Windows

___ Macintosh

___ Unix, _____

___ Other, _____

Is this a hardware product or a software product?

___ Hardware, _____ Bus

___ Software

Can the input image come directly from a scanner?

___ Yes, supported scanners are: _____

Can the input image come from a file?

___ Yes, supported file formats are: _____

Can the input image be a compressed file (e.g. JPEG)?

___ Yes, supported compression formats are: _____

What input pixel depths are supported?

___ 24-bit color ___ 8-bit monochrome ___ Binary ___ Other _____

Time to read in a letter size (8.5" x 11 ") image

___ 24-bit color, 300 dpi, uncompressed _____

___ 24-bit color, 300 dpi, JPEG compressed _____

___ 8-bit monochrome, 300 dpi, uncompressed _____

___ 8-bit monochrome, 300 dpi, JPEG compressed _____

___ Binary, 600 dpi, uncompressed _____

___ Binary, 600 dpi, Group 4 compressed _____

Recommended hardware/software configuration to achieve these throughputs:

_____

_____

_____

Are there any maximum limits to the image size? (e.g. 64K x 64K pixels)

_____

What image enhancement operations are supported?

___ Grayscale Analysis

___ Histogram ___ Average ___ Min/Max ___ Standard deviation

___ Automatic brightness or contrast adjustment (please describe)

_____

Time to process an 8.5" x 11", 300 dpi, 24-bit image _____

___ Color space conversion _____

___ Automatic deskew

___ Based on measured edges of page

___ Based on location of information in the image

___ Automatic cropping

___ Based on measured edges of page

___ Based on location of information in the image

___ Convolutions (linear filtering):

___ Edge enhance ___ Sharpen ___ Average ___ Other _____

Time to process an 8.5" x 11", 300 dpi, 24-bit image _____

___ Morphological operations (non-linear filtering):

___ Erosion ___ Dilation ___ Other _____

Time to process an 8.5" x 11", 300 dpi, 8-bit image _____

___ Page Segmentation (text/photo discrimination)

Time to process an 8.5" x 11", 300 dpi, 8-bit image _____

___ Thresholding and/or halftoning

___ Fixed threshold _____

___ Global adaptive threshold _____

___ Local adaptive threshold _____

___ Edge enhanced threshold _____

___ Halftoning _____

___ Error diffusion _____

___ Other _____

Time to process an 8.5" x 11", 300 dpi, 8-bit image _____

___ Speckle Deletion

Size range for black specks _____

Size range for white specks _____

___ 2-D Transforms _____

___ Resampling (resolution conversion)

___ Nearest neighbor ___ Interpolation ___ Other _____

___ Other image processing operations (please describe) _____

_____

_____

_____

Can the output image go to a file?

___ Yes, supported file formats are: _____

Can the output image be compressed (e.g. JPEG)?

___ Yes, supported compression formats are: _____

What output pixel depths are supported?

___ 24-bit color ___ 8-bit monochrome ___ Binary ___ Other _____

Time to write out a letter size (8.5" x 11 ") image

___ 24-bit color, 300 dpi, uncompressed _____

___ 24-bit color, 300 dpi, JPEG compressed _____

___ 8-bit monochrome, 300 dpi, uncompressed _____

___ 8-bit monochrome, 300 dpi, JPEG compressed _____

___ Binary, 600 dpi, uncompressed _____

___ Binary, 600 dpi, Group 4 compressed _____

Recommended hardware/software configuration to achieve these throughputs:

_____

_____

_____

Is a standard user interface style supported (e.g. MOTIF or MS Windows)?

___ Yes, _____

What maximum display resolution is supported? _____ Horiz. x _____ Vert.

Is scale-to-gray supported for viewing higher resolution images?

___ Yes, _____

Are any printers supported?

___ Yes, _____

What features of your image processing tools make them particularly well suited for use in the enhancement (quality improvement and/or file size reduction) of scanned manuscripts?

_____

_____

Are there any libraries currently using your image processing tools?

_____

_____

Any other comments:

_____

_____

_____

_____

_____

_____

_____

_____

NON-INTERACTIVE (AUTOMATIC) IMAGE PROCESSING TOOLS

Which of your products would you recommend for the automatic enhancement of scanned manuscript images at the Library of Congress? (For multiple recommendations, use duplicate copies of this form).

Product _____ Description _____ Price _____

What computer environment is this product compatible with?

___ PC, DOS

___ PC, Windows

___ Macintosh

___ Unix, _____

___ Other, _____

Is this a hardware product or a software product?

___ Hardware, _____ Bus

___ Software

Is this product stand-alone or is it a module (or library) for use with

other software?

___ Stand alone, please describe _____

___ Module, operates with: _____

Can the input image come directly from a scanner?

___ Yes, supported scanners are: _____

Can the input image come from a file?

___ Yes, supported file formats are: _____

Can the input image be a compressed file (e.g. JPEG)?

___ Yes, supported compression formats are: _____

What input pixel depths are supported?

___ 24-bit color ___ 8-bit monochrome ___ Binary ___ Other _____

Time to read in a letter size (8.5" x 11 ") image

___ 24-bit color, 300 dpi, uncompressed _____

___ 24-bit color, 300 dpi, JPEG compressed _____

___ 8-bit monochrome, 300 dpi, uncompressed _____

___ 8-bit monochrome, 300 dpi, JPEG compressed _____

___ Binary, 600 dpi, uncompressed _____

___ Binary, 600 dpi, Group 4 compressed _____

Recommended hardware/software configuration to achieve these throughputs:

_____

_____

_____

Are there any maximum limits to the image size? (e.g. 64K x 64K pixels)

_____

What image enhancement operations are supported?

___ Grayscale Analysis

___ Histogram ___ Average ___ Min/Max ___ Standard deviation

___ Automatic brightness or contrast adjustment (please describe)

_____

Time to process an 8.5" x 11", 300 dpi, 24-bit image _____

___ Color space conversion _____

___ Automatic deskew

___ Based on measured edges of page

___ Based on location of information in the image

___ Automatic cropping

___ Based on measured edges of page

___ Based on location of information in the image

___ Convolutions (linear filtering):

___ Edge enhance ___ Sharpen ___ Average ___ Other _____

Time to process an 8.5" x 11", 300 dpi, 24-bit image _____

___ Morphological operations (non-linear filtering):

___ Erosion ___ Dilation ___ Other _____

Time to process an 8.5" x 11", 300 dpi, 8-bit image _____

___ Page Segmentation (text/photo discrimination)

Time to process an 8.5" x 11", 300 dpi, 8-bit image _____

___ Thresholding and/or halftoning

___ Fixed threshold _____

___ Global adaptive threshold _____

___ Local adaptive threshold _____

___ Edge enhanced threshold _____

___ Halftoning _____

___ Error diffusion _____

___ Other _____

Time to process an 8.5" x 11", 300 dpi, 8-bit image _____

___ Speckle Deletion

Size range for black specks _____

Size range for white specks _____

___ 2-D Transforms _____

___ Resampling (resolution conversion)

___ Nearest neighbor ___ Interpolation ___ Other _____

___ Other image processing operations (please describe) _____

_____

_____

_____

Can the output image go to a file?

___ Yes, supported file formats are: _____

Can the output image be compressed (e.g. JPEG)?

___ Yes, supported compression formats are: _____

What output pixel depths are supported?

___ 24-bit color ___ 8-bit monochrome ___ Binary ___ Other _____

Time to write out a letter size (8.5" x 11 ") image

___ 24-bit color, 300 dpi, uncompressed _____

___ 24-bit color, 300 dpi, JPEG compressed _____

___ 8-bit monochrome, 300 dpi, uncompressed _____

___ 8-bit monochrome, 300 dpi, JPEG compressed _____

___ Binary, 600 dpi, uncompressed _____

___ Binary, 600 dpi, Group 4 compressed _____

Recommended hardware/software configuration to achieve these throughputs:

_____

_____

_____

What modes of automatic operation are supported?

___ Integrated into another automatic process (e.g. scanning), please describe

_____

___ Batch processing of a list of files _____

___ DLL (Dynamically Linked Library) _____

What features of your image processing tools make them particularly well suited for use in the enhancement (quality improvement and/or file size reduction) of scanned manuscripts?

_____

_____

Are there any libraries currently using your image processing tools?

_____

_____

Any other comments:

_____

_____

_____

_____

_____

_____

_____

_____

Contents

Appendix E
Image Enhancement Survey Responses

The address and phone number of each of the companies which were contacted is listed below. The following **category codes** indicate the nature of their products:

Code - Type of Product

E - Image **E**diting or Viewing
A - Image **A**nalysis or Machine Vision
C - Image **C**ompression
B - **B**inary Image Cleanup for OCR
H - **H**ardware, Accelerators
S - Optimized **S**canning
O - **O**ther

Company Category Codes
Status of Questionnaire

AccelGraphics **O, H**
2630 Walsh Avenue
Santa Clara, CA 95051
(408) 727-6126
Literature received

Accusoft **E, C, B**
2 Westbourough Bus. Pk.
Westbourough, MA 01581
(800) 525-3577
Survey response,
Literature received

Adaptive Solutions **E, H**
1400 NW Compton Dr.
Beaverton, OR 97006
(503) 690-1238
Survey response,
Literature received

Adobe Systems (see Aldus) **E**
1545 Charlson Road
Mountain View, CA 94039
(415) 961-4400
Phone response,
sending Literature

Alacrity Systems **O**
43 Newburg Road
Hackettstown NJ 07840
(908) 813-2400
Literature received

Aldus (see Adobe) **E**
411 First Street South
Seattle, WA 98104
(206) 622-5500

Applied Silicon **O, H**
220-2427 Holly Lane
Ottawa Ontario K1V 7P2
(613) 738-2434

Literature received

Audre, Incorporated **O**
10915 Technology Place
San Diego, CA 92127
Literature received

Automatix (Acuity) **A, H**
755 Middlesex Tpke
Billercia MA 01821
(603) 577-5830
Literature received

Aware, Inc. **C**
One Memorial Drive
Cambridge, MA 02142
(617) 577-1700
Literature received

Cimmetry Systems, Inc. **E**
1430 Mass Avenue
Cambridge, MA 02138
(514) 735-3219
Literature received

COGNEX **A, H**
15 Crawford Street
Needham, MA 02194
(617) 449-6030
Literature received

Coreco **A, H**
6969 Trans-Canada Hwy
St Laurent QU H4T 1V8
(514) 333-1301
Survey response,
Literature received

Corel Systems **E**
1600 Carling Ave.
Ottawa, KIZ 8R7
(613) 728-8200
Literature received

Data Translation **A, H**
100 Locke Drive
Marlboro, MA 01752
(508) 525-8528
Literature received, demo disk

DECOMP **C**
2528 West Greenbrier
Anaheim, CA 92801
(714) 952-3238
Survey response, product

Literature, demo disk

Delta Point **E**
2 Harris Court
Suite B-1
Monterey, CA 93940
(408) 648-4000
Literature received

Edsoft, Inc. **C**
387 West Center
Orem, UT 84058
(801) 221-2728
Literature received

Electronic Imagery **E**
100 Park Central Blvd
Pompano, FL 33064
(305) 968-7100
Survey response,
Literature received,
sending demo disk

Genesys **E, S**
Four North Park Drive
Suite 400
Hunt Valey, MD 21030
(410) 785-0000
Literature received

Graftek Imaging **A**
240 Oral School Road
Mystic, CT 06355
(800) 959-3011
Literature received

Handmade Software **E, C**
48820 Kato Road
Suite 110B
Fremont, CA 94538
(800) 358-3588
Literature received

Hypersoft, Inc. **E, C**
169 South River Road
Suite 11
Bedford, NH 03110
(603) 647-7880
Literature and demo disk received

ImageFast **O**
7926 Jones Branch Dr.
McLean, VA 22102
(703) 893-1934
Literature received

Image Machines **E**
590 Herndon Pkwy.
Herndon, VA 22070
(703) 709-7475
Literature and demo disk received

Image Systems Inc. **E**
2575 Aero Park Drive
Traverse City, MI 49684
(518) 283-8783
Literature received

Imaging Technology **A, H**
55 Middlesex Tpke
Bedford, MA 01730
(617) 275-2700
Literature received

Incat Systems **E**
1684 Dell Ave
Campbell, CA 95008
(408) 379-2400
Survey response,
sending demo disk

Informative Graphics **E**
706 E. Bell Road
Suite 207
Phoenix, AZ 85022
(602) 971-6061
Literature received

Inlite Research **B**
355 West Olive Avenue
Suite 211
Sunnyvale, CA 94086
(408) 737-7092
Literature received
and demo disk

Inset Systems **E, C**
71 Commerce Drive
Brookfield, CT 06804
(203) 740-2400
Literature received

Intrafed **E, S**
6903 Rockledge Drive
11th Floor
Bethesda, MD 20817
Literature received

ISTR, Incorporated **E**
360 Delaware Avenue
Suite 300
Buffalo, New York 14202

Literature received

Kofax Image Products **E, S, H**
3 Jenner Street
Irvine, CA 92718
(714) 727-1733
Literature received

Laser Master **E, H**
6900 Shady Oak Road
Eden Prarie MN 55344
(612) 944-9021
Literature received

LEAD Technologies **E, C**
8701 Mallard Creek Rd
Charlotte, NC 28262
(800) 637-4699
Literature received

MathSoft **E, A**
1700 Westlake Ave N
Suite 500
Seattle, WA 98109
(800) 569-0123
Literature received

Math Works Inc. **E, A**
24 Prime Park Way
Natick, MA 01760
(508) 653-1415
Literature received,
demo disk available
for 1 month loan

MATROX Electronics **A, H**
1055 St. Regis Blvd.
Dorval, H9P 2T4
(514) 685-2630
Literature received

Media Cybernetics **E**
8484 Georgia Ave.
Silver Spring, MD 20910
(800) 992-4256
Literature received

Optimas Corporation **E, A**
18911 North Creek Parkway
Suite 101
Bothell, WA 98011

Pegasus **E, C**
4350 W Cypress St.#908
Tampa, FL 33607
(813) 875-7575
Literature and demo disk received

Pixel Translations **E, S**
3031 Tisch Way
Suite 310
San Jose, CA 95128
(408) 985-6600
Literature and demo disk received

Pixelworks **E, C, H**
7 Park Ave.
Hudson, NH 03051
(603) 880-1322
Literature received

Seabreeze Engineering **E, O**
119 Commerce Way
Suite E
Sanford, FL 32771
(407) 321-2096
Literature received

Seaport Imaging **E, B, H**
1340 Saratoga-Sunnyvale
Suite 104
San Jose, CA 95129
(408) 366-6400
Literature received

Sequoia Data Corp. **B**
433 Airport Blvd.
Suite 414
Burlingame, CA
(415) 696-8750
Literature received

Softkey International Inc. **E**
450 Franklin Road
Suite 100
Marietta, GA 30067
Literature received

Storm Technology **E, C**
1861 Landings Drive
Mountain View CA 94043
Survey response,
Literature,
Mac demo disk

Tardis Systems **E**
P.O.Box 1251
Los Alimos, NM 87544
(505) 662-9401
Literature received

TIS America, Inc. **B**
25 Mall Road
Suite 300

Burlingon MA 01803
(617) 270-0685
Survey response,
Literature received

TMS Inc. **E, C**
P.O. Box 1358
Stillwater, OK 74067
(405) 377-0880
Literature and demo disk received

U-Lead Corporation **E**
970 West 190th Street
Suite 520
Torrance, CA 90502
(310) 523-9393
Literature received

UniSoft Imaging **E, C**
4606 N. Britton Road
Stillwater, OK 74075
(405) 624-9257
Literature and demo disk received

uTech (MuTech) **A, C, H**
800 W. Cummings Park
Suite 3800
Woburn, MA 01801
(617) 935-1770
Literature received

Wolfram Research **E, A**
17 West Street
Marblehead MA 01945
(617) 639-4423
Literature and demo disk received

Xing Technology **C**
1540 W. Branch St.
Arroyo Grande CA 93420
(800) 294-6448
Literature received

Xionics **E, S, C, H**
Two Corporation Way
Peabody, MA 01960
(508) 531-6666
Literature received

Zenographics **E**
Literature received

Contents

Appendix F
Image Enhancement Survey Mailing List

Accusoft
2 Westborough Bus. Pk.
Westborough, MA 01581

Adaptive Solutions
1400 NW Compton Dr.
Beaverton, OR 97006

Adobe Systems
1545 Charlson Road
Mountain View, CA 94039

Alacrity Systems
43 Newburg Road
Hackettstown NJ 07840

Aldus
411 First Street South
Seattle, WA 98104

Applied Silicon
220-2427 Holly Lane
Ottawa Ontario K1V 7P2

Ariel
433 River Road
Highland Park, NJ 08901

Audre, Incorporated
10915 Technology Place
San Diego, CA 92127

Automatix
755 Middlesex Tpke
Billercia MA 01821

Black Ice
113 Route 122
Amherst, NH 03031

Cimmetry Systems, Inc.
1430 Mass Avenue
Cambridge, MA 02138

Cimage
3885 Research Park Dr.
Ann Arbor, MI 48108

Computer Presentations
1117 Cypress Street
Cincinatti, OH 45206

Coreco
6969 Trans-Canada Hwy
St Laurent QU H4T 1V8

Corel Systems
1600 Carling Ave.
Ottawa, KIZ 8R7

COSMOS Imaging Systems

33971 Selva Road
Suite 160
Monarch Beach CA 92677

Cognex
15 Crawford Street
Needham, MA 02194

Datacube
300 Rosewood Dr.
Danvers, MA 01923

DECOMP
2528 West Greenbrier
Anaheim, CA 92801
Electrnic Imagery
100 Park Central Blvd
Pompano, FL 33064

Handmade Software
48820 Kato Road
Suite 110B
Fremont, CA 94538

Hypersoft, Inc.
169 South River Road
Suite 11
Bedford, NH 03110

Image Access
543 NW 77th Street
Boca Raton, FL 33487

Image Business Systems
417 Fifth Ave.
New York, NY 10016

ImageFast
7926 Jones Branch Dr.
McLean, VA 22102

Image Machines
590 Herndon Pkwy.
Herndon, VA 22070

Image Systems Inc.
2575 Aero Park Drive
Traverse City, MI 49684

Imaging Technology
55 Middlesex Tpke
Bedford, MA 01730

Incat Systems
1684 Dell Ave
Campbell, CA 95008

Information Tech.
3520 W Hallandale
Pembroke Park FL 33023

Informative Graphics
706 East Bell Road
Suite 207

Phoenix, AZ 85022

Infotronic
8834 N Capitan of Tx Hwy
Suite 200
Austin, TX 78757

Inlite Research
355 West Olive Avenue
Suite 211
Sunnyvale, CA 94086

Inset Systems
71 Commerce Drive
Brookfield, CT 06804

Kofax Image Products
3 Jenner Street
Irvine, CA 92718

Kubota Graphics
2630 Walsh Ave.
Santa Clara, CA 95051

Laser Master
6900 Shady Oak Road
Eden Prarie MN 55344

LEAD Technologies
8701 Mallard Creek Rd
Charlotte, NC 28262

Light Source Inc.
17 E Sir Frances Drake
Larkspur, CA 94939

Mathematica
402 S. Kentucky Ave.
Lakeland, FL 33801

Math Works Inc.
24 Prime Park Way
Natick, MA 01760

MATROX Electronics
1055 St. Regis Blvd.
Dorval, H9P 2T4

Media Cybernetics
8484 Georgia Ave.
Silver Spring, MD 20910

Micrografx
1303 Arapaho Road
Richardson, TX 75081

Optibase
P.O. Box 809030
Dallas, TX 75380

Pegasus
4350 W Cypress St.
#908
Tampa, FL 33607

Pixelworks
7 Park Ave.
Hudson, NH 03051

Pixel Translations
3031 Tisch Way
Suite 310
San Jose, CA 95128

Power Pixel
2312 Walsh Ave.
Santa Clara, CA 95051

Research Development
300 Pearl Street
Suite 200
Buffalo, NY 14202

Seaport Imaging
1340 Saratoga-Sunnyvale
Suite 104
San Jose, CA 95129

Sequoia Data Corp.
433 Airport Blvd.
Suite 414
Burlingame, CA

Staude Imaging
2837 Walnut Blvd.
Walnut Creek, CA 94596

Storm Technology
1861 Landings Drive
Mountain View CA 94043

Sunrise Imaging
42701 Lawrence Pkwy.
Fremont, CA 94538

Telephoto Communication
11722-D Sorrento Valley
San Diego, CA 92121

TMS Inc.
P.O. Box 1358
Stillwater, OK 74067

Top Image Systems
25 Mail Road
Suite 300
Burlington, MA 01803

U-Lead Corporation
970 West 190th Street
Suite 520
Torrance, CA 90502

VI&C Technology
54 Middlesex Turnpike
Bedford, MA 10730

Video Image & Compress

2221 Rosecrans Ave.
Suite 221
El Segundo, CA 90245

Wolfram Research
17 West Street
Marblehead MA 01945

Xing Technology
1540 W. Branch St.
Arroyo Grande CA 93420

Xionics
Two Corporation Way
Peabody, MA 01960

ZSoft
450 Franklin Road
Suite 100
Marietta GA, 30067

Contents

Appendix G
Scanner Survey Questionnaire

This appendix shows the scanner survey questionnaire which was circulated. It includes both the questionnaire on scanners optimized for fragile document handling and the questionnaire for scanners optimized for efficient throughput.

Survey of Document Scanners for Manuscript Collections

This survey is divided into three sections: General Information; Scanners Optimized for Fragile Document Scanning; and Scanners Optimized for Efficient Document Scanning. If you don't have time to fill out the survey form, please call me instead, or send your standard product literature.

GENERAL INFORMATION

Primary Contact _____ Telephone _____
Company Name _____
Address _____
_____
_____

Does your company currently manufacture document scanners suitable for use in scanning manuscripts or other Library collections?

Yes ___ No ___ Comments _____

If selected by the Library Committee, will it be possible to arrange a demonstration of your products either at the Library of Congress or at an offsite location in the greater Washington, DC, area?

Yes ___ No ___ Comments _____

Could a video tape of a demonstration of your products be provided as an alternate form of demonstration?

Yes ___ No ___ Comments _____

What particular strengths do your company's products have that are suited to the Library's needs?

_____
_____

Any other information or suggestions?

_____
_____
_____
_____

SCANNERS OPTIMIZED FOR FRAGILE DOCUMENT SCANNING

Which of your scanners would you recommend for use with valuable, fragile documents? (For multiple recommendations, use duplicate copies of this form).

Model _____ Description _____ Price _____

What feed methods are available on this scanner?

___ Manual, flat-bed
___ Manual, overhead
___ Automatic, straight paper path, please describe _____
_____

___ Automatic, curved paper path, please describe _____
_____
__ Other, please describe _____
_____

What is the document size range for the scanner?

Minimum Height _____ in. Maximum Height _____ in.
Minimum Width _____ in. Maximum Width _____ in.

What is the Optical Resolution of the scanner? (from photosite spacing)

Horizontal dpi _____ Vertical dpi _____

What is the finest resolution pattern that can be "resolved" by the scanner?

IEEE Facsimile Test Chart Std. 167A-1987 _____ line pairs
or ANSI/AIIM Scanner Test Target MS44 _____ line pairs

What is the Interpolated Resolution range for the scanner?

Minimum: Horizontal dpi _____ Vertical dpi _____
Maximum: Horizontal dpi _____ Vertical dpi _____

Scanning throughput in pages/minute for letter size (8.5" x 11") pages:

___ 24-bit Color, 600 dpi _____
___ 24-bit Color, 300 dpi _____
___ Color, Other, describe _____
___ 8-bit Monochrome, 600 dpi _____
___ 8-bit Monochrome, 300 dpi _____
___ Monochrome, Other, describe _____
___ Binary, 1200 dpi _____
___ Binary, 600 dpi _____
___ Binary, 300 dpi _____
___ Binary, Other, describe _____

Recommended hardware/software configuration for achieving these throughputs?

_____
_____
_____
_____

What Methods of Thresholding are supported? (Please describe)

___ Fixed Threshold _____
___ Global Adaptive Threshold _____
___ Local Adaptive Threshold _____
___ Edge Threshold _____
___ Halftoning _____
___ Error Diffusion _____
___ Text/Photo Discrimination _____
___ Other _____

Are simultaneous Color/Grayscale and Binary outputs available?

Yes ___ No ___ Explanation _____

What Automatic Image Enhancement Operations are supported?

___ Brightness Adjustment _____
___ Contrast Adjustment _____
___ Edge Enhancement _____
___ Despeckle _____
___ Deskew _____
___ Rotation _____
___ Cropping _____

What is the user interface for setting the scanner's thresholding, enhancement, and scan area parameters? (e.g. Front panel, or through Windows)

_____

Is a preview scan produced (or required)?

_____

Can compression be performed by the scanner (or its recommended interface)?

___ Group 3 _____
___ Group 4 _____
___ JBIG _____
___ JPEG _____
___ Other _____

What output interfaces are supported?

___ Video _____
___ Computer Bus (e.g. PCI) _____
___ SCSI _____
___ Network (e.g. Token Ring) _____

What features of this scanner make it particularly well suited for scanning fragile manuscripts?

_____
_____
_____

Are there any other libraries currently using this scanner for scanning manuscripts?

_____
_____
_____

Any Other Comments

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

_____

SCANNERS OPTIMIZED FOR EFFICIENT DOCUMENT SCANNING

Which of your scanners would you recommend for scanning of less fragile documents? (For multiple recommendations, use duplicate copies of this form).

Model _____ Description _____ Price _____

What feed methods are available on this scanner?

___ Manual, flat-bed
___ Manual, overhead
___ Automatic, straight paper path, please describe _____
_____
___ Automatic, curved paper path, please describe _____
_____
__ Other, please describe _____
_____

What is the document size range for the scanner?

Minimum Height _____ in. Maximum Height _____ in.
Minimum Width _____ in. Maximum Width _____ in.

What is the Optical Resolution of the scanner? (from photosite spacing)

Horizontal dpi _____ Vertical dpi _____

What is the finest resolution pattern that can be "resolved" by the scanner?

IEEE Facsimile Test Chart Std. 167A-1987 _____ line pairs
or ANSI/AIIM Scanner Test Target MS44 _____ line pairs

What is the Interpolated Resolution range for the scanner?

Minimum: Horizontal dpi _____ Vertical dpi _____
Maximum: Horizontal dpi _____ Vertical dpi _____

Scanning throughput in pages/minute for letter size (8.5" x 11") pages:

___ 24-bit Color, 600 dpi _____
___ 24-bit Color, 300 dpi _____
___ Color, Other, describe _____
___ 8-bit Monochrome, 600 dpi _____
___ 8-bit Monochrome, 300 dpi _____
___ Monochrome, Other, describe _____
___ Binary, 1200 dpi _____
___ Binary, 600 dpi _____
___ Binary, 300 dpi _____
___ Binary, Other, describe _____

Recommended hardware/software configuration for achieving these throughputs?

_____
_____
_____
_____

What Methods of Thresholding are supported? (Please describe)

___ Fixed Threshold _____

___ Global Adaptive Threshold _____
___ Local Adaptive Threshold _____
___ Edge Threshold _____
___ Halftoning _____
___ Error Diffusion _____
___ Text/Photo Discrimination _____
___ Other _____

Are simultaneous Color, Grayscale and/or Binary outputs available?

Yes ___ No ___ Explanation _____

What Automatic Image Enhancement Operations are supported?

___ Brightness Adjustment _____
___ Contrast Adjustment _____
___ Edge Enhancement _____
___ Despeckle _____
___ Deskew _____
___ Rotation _____
___ Cropping _____

What is the user interface for setting the scanner's thresholding, enhancement, and scan area parameters? (e.g. Front panel, or through Windows)

_____

Is a preview scan produced (or required)?

_____

Can compression be performed by the scanner (or its recommended interface)?

___ Group 3 _____
___ Group 4 _____
___ JBIG _____
___ JPEG _____
___ Other _____

What output interfaces are supported?

___ Video _____
___ Computer Bus (e.g. PCI) _____
___ SCSI _____
___ Network (e.g. Token Ring) _____

What features of this scanner make it particularly well suited for scanning less fragile manuscripts efficiently?

_____
_____
_____

Are there any other libraries currently using this scanner for scanning manuscripts?

_____
_____
_____

Any Other Comments

_____
_____
_____
_____

Contents

Appendix H
Scanner Survey Responses

The address and phone number of each of the companies contacted is listed below. The following "category codes" indicate the nature of their product:

Code - Type of Product

P - **P**lanetary Scanner
F - **F**latbed Scanner
S - **S**heet Fed Scanner
M - Film (or **M**icrofilm) Scanner
C - **C**olor Fax
O - **O**ther

Company, Category Codes
Status of Questionnaire

AGFA Corporation **F**
200 Ballardvale St.
Wilmingon, MA 01887
(508) 658-5600
Literature received

Amitech Corporation **M**
5501 Backlick Road
Suite 200
Springfield, VA 22151
(703) 256-2020
Literature received

ANAtech **S**
10499 Bradford Road
Littleton, Colorado 80127
(303) 973-6722
Literature received

Axiom Research, Inc. **O**
1304 East Eighth Street
Tucson, AZ 85719
(602) 791-2864
Literature received

Bell & Howell Co. **S**
6800 McCormick Road....
Chicago, IL 60645-2797..
(800) SCAN-494
Survey response,
Literature received

CalComp **S**
14555 N. 82nd Street
Scottsdale, AZ 85260
(602) 948-6540
Literature received

Canon USA Inc. **F**
One Canon Plaza
Lake Success NY, 11042
(714) 438-3000
Literature received

Cognitronics Imaging **S**
4780 Mission Gorge Pl.
San Diego, CA 92120
(619) 265-3434
Literature received

Contex **F**
Literature received

Dicomed, Inc. **M**
12270 Nicollet Avenue
Burnsville, MN 55337
(612) 895-3000
Literature received

ECRM **F**
554 Clark Road
Tewksbury, MA 01876
(800) 537-ECRM
Literature received

Eastman Kodak Company **S**
Office Imaging
901 Elmgrove Road
Rochester, NY 14653
(602) 759-0716
Literature received

Envisions **F**
822 Mahler Rd.
Burlingame, CA 94010
(415) 692-9061
Literature received

Epson America **F**
P.O Box 2842
Torrance, CA 90509-2842
(310) 782-0770
Literature received

FICUS Systems Inc. **F**
460 Wildwood Avenue
Woburn MA 01801
(617) 938-8799
Literature received

Fujitsu Computer Products **F,S**
2904 Orchard Parkway
San Jose, CA 95134
(408) 432-6333
Literature received

Hewlett Packard Co. **F**
16399 West Bernardo Dr.
San Diego, CA 92127
(800) 752-0900
Literature received

Houston Fearless 76 **M**
203 West Artesia Blvd.
Compton, CA 90220
(310) 605-0755
Literature received

Howtek **O**
21 Park Avenue
Hudson, NH 03051
(603) 882-5200
Literature received

IBM **F**
PO Box 100
Mail Drop 1335, Bldg.1
Somers, NY 10589
(704) 594-3659
Literature received

Ideal Scanners **F,S**
11810 Parklawn Drive
Rockville, MD 20852
(301) 468-0123
Literature received

Image Access **F,S**
543 NW 77th Street
Boca Raton, FL 33487
(407) 995-8334
Literature received

Image Graphics Inc. **O**
917 Bridgeport Ave.
Shelton, CT 06484
(203) 926-0100
Literature received

Imapro **F,M**
2400 St. Laurent Blvd.
Ottawa, Ontario
Canada, K1G 5A4
(613) 738-3000
Literature received

Improvision **S**
5901 Christie Avenue
Suite 502
Emeryville, CA 94608
Survey response,
Literature received

LA CIE Limited **F**
8700 SW Creekside Pl.
Beaverton, OR 97008
(503) 520-9000
Literature received

Laser Today International (LTI) **C**
1924 Old Middlefield Way
Mountain View, CA 94043
(415) 961-3015
Literature received

Lenzar Electro Optics **F,M**
1006 West 15th Street
Riveria Beach, FL 33404
Survey response,
Literature received

Mekel Engineering **M**
777 S Penarth Ave
Walnut, CA 91789
(909) 594-5158
Literature received

Minolta Corporation **P**
Document Imaging Div.
101 Williams Drive
Ramsey, NJ 07446
Survey response,
Literature received

Our Business Machine (OBM) **S,C**
12901 Ramona Blvd.
Suite J
Irwindale, CA 91706
(818) 337-9614
Literature received

OCE Graphics USA **M**
5450 N. Cumberland...
Chicago, IL 60656...
(312) 714-8500
Literature received

Paradigm Group **S,M**
3 Corporate Park Drive
Suite 270
Irvine, CA 92714
(714) 251-9410
Literature received

Pentax Technologies **F,S**
100 Technolog Dr.
Broomfield, CO 80021
(303) 460-1600

Literature received

Photomatrix **S**
5700 Buckingham Pkwy
Culver City, CA 90230
(213) 417-3800
Survey response,
Literature received

Photometrics **O**
3440 E. Britannia Dr.
Tuscon, AZ 85706
(602) 889-9933
Literature received

Pixel Craft (Xerox) **F**
17950 Sunmeadow #4504
Dallas, TX 75252
(800) 933-0300
Literature received

Ricoh Corporation **F,S**
3001 Orchard Pkwy
San Jose, CA 95134
(408) 432-8800
Literature received

Scan-Optics Inc. **S**
22 Prestige Park Circle
East Hartford, CT 06108
(203) 289-6001
Literature received

Scan Vantage **S,M**
P.O. Box 7067
Overland Park, KS 66207
Literature received

Scitex **F**
8 Oak Park Drive
Bedford, MA 01730
(617) 275-5150
Literature received

Sharp Electronics **F**....
Sharp Plaza
Mahwah, NJ 07675
(800) 892-9204
Literature received

SPARK **F,M**
1939 Waukegan Road
Glenview, IL 60025
(708) 998-6640
Literature received

Staude Imaging **S,M**
937 Meadowvale Court
Martinez, CA 94553
(510) 229-5588
Literature received

Sunrise Imaging Inc. **M**
42701 Lawrence Place
Fremont, CA 94538
(510) 657-6250
Survey response,
Literature received

Tangent Engineering, Inc. **F,S**
5776 Stoneridge Mall Road
Suite 135
Pleasanton, CA 94588
510) 227-0712
Literature received

Terminal Data Corp. (BancTec TDC) **S**
5898 Condor Drive
Moorpark, CA 93021
(805) 529-1500
Survey response,
Literature received

TEC America **F**
2710 Lakeview Court
Fremont, CA 94538
(510) 651-5333
Literature received

Ultima International (Artec) **F**
3358 Gateway Boulevard
Fremont, CA 94538
(510) 659-1217
Literature received

UMAX Technologies **F**
3353 Gateway Blvd
Fremont, CA 94538
(510) 651-8883
Literature received

Unisys Corporation **O**
P.O. Box 500
Blue Bell, PA 19424
(215) 986-4011
Literature received

VIDAR Systems Inc. **S**
520 D. Herndon Pkwy.
Herndon, VA 22070

(800) 471-SCAN
Literature received

Visionshape **S**
1434 West Taft Ave.
Orange, CA 92665
(714) 282-2668
Literature received

Contents

Appendix I
Scanner Survey Mailing List

Abaton
48431 Milmont Drive
Fremont, CA 94538

Advanced Vision Res.
2201 Qume Drive
San Jose, CA 95131

ANA Tech Corporation
10499 Bradford Road
Littleton, CO 80127

AGFA Corporation
200 Ballardvale St.
Wilmingon, MA 01887

Bell & Howell Co.
6800 McCormick Road
Chicago, IL 60645-2797

CalComp
14555 N. 82nd Street
Scottsdale, AZ 85260

Canon USA Inc.
One Canon Plaza
Lake Success NY, 11042

Cardamation Company
P.O. Box 1276
Lionville, PA 19341

Chinon America
615 Hawii Ave.
Torrance, CA 90503

Cognitronics Imaging
4780 Mission Gorge Pl.
San Diego, CA 92120

DEST
1015 East Brokaw Road
San Jose, CA 95131

Du Pont Imaging
65 Harristown Road
Glen Rock, NJ 07452

Eastman Kodak Company
Office Imaging
901 Elmgrove Road
Rochester, NY 14653

Ektron Applied Imaging
23 Crosby Drive
Bedford, MA 01730

Envisions
822 Mahler Rd.
Burlingame, CA 94010

Epson America
P.O Box 2842
Torrance, CA 90509-2842

FORA Inc.
3081 North First St.
San Jose, CA 95134

Fuji Photo Film
555 Taxter Road
Elmsford, NY 10523

Fujitsu Computer Products
2904 Orchard Parkway
San Jose, CA 95134

Hewlett Packard Co.
16399 West Bernardo Dr.
San Diego, CA 92127

Hitachi America
Computer Division
2000 Sierra Point
Brisbane, CA 94005

Houston Fearless 76
203 West Artesia Blvd.
Compton, CA 90220

Howtek
21 Park Avenue
Hudson, NH 03051

HSD Microcomputer
1350 Pear Ave., Ste. C
Mountain View, CA 94043

IBM
PO Box 100
Mail Drop 1335, Bldg.1
Somers, NY 10589

Ideal Scanners
11810 Parklawn Drive
Rockville, MD 20852

Image Graphics Inc.
917 Bridgeport Ave.
Shelton, CT 06484

Improvision
5901 Christie Avenue
Suite 502
Emeryville, CA 94608

Integrated Scanners
15740 Dooley Road
Dallas, TX 75244

Konica Imaging USA
71 Charles Street
Glen Cove, NY 11542

LA CIE Limited

8700 SW Creekside Pl.
Beaverton, OR 97008

Lenzar Electro Optics
1006 West 15th Street
Riveria Beach, FL 33404

Logitech Inc.
6505 Kaiser Drive
Fremont, CA 94555

Marstek
15225 Alton Parkway
Irvine, CA 92718

Mekel Engineering
777 S Penarth Ave
Walnut, CA 91789

Microseal Corp.
2000 Lewis Avenue
Zion, IL 60099

MicroTek
680 Knox Street
Torrance, CA 90502

Minolta Corporation
Document Imaging Div.
101 Williams Drive
Ramsey, NJ 07446

Nikon Elect. Imaging
1300 Walt Whitman Rd.
Melville, NY 11747

NISCA Incorporated
1919 Old Denton Road
Suite 104
Carrolton, TX 75006

OA Data Company
144-50 32nd Ave.
Flushing, NY 11354

OCE Graphics USA
5450 N. Cumberland
Chicago, IL 60656

Okidata
532 Fellowship Road
Mt. Laurel, NJ 08054

OMRON Corporation
10201 Torre Ave.
Cupertino, CA 95014

Optical Laser
5862 Bolsa Ave.
Huntington Beach, CA 92649

Optronics Specialty
8954 Comanche Ave.
Chatsworth, CA 91311

Our Business Machine
12901 Ramona Blvd.
Suite J
Irvindale, CA 91706

Panasonic Imaging
Two Panasonic Way
Secaucus, NJ 07094

Pentax Technologies
100 Technolog Dr.
Broomfield, CO 80021

Photomatrix
5700 Buckingham Pkwy.
Culver City, CA 90230

Photometrics
3440 E. Britannia Dr.
Tuscon, AZ 85706

Poloroid Elect. Image
549 Technology Square
Cambridge, MA 02194

Prime Option Inc.
2341 West 205th Street
Torrance, CA 90501

Regent Peripherals
901 Rainier Ave.
Renton, WA 98055

Ricoh Corporation
3001 Orchard Pkwy.
San Jose, CA 95134

Scan-Optics Inc.
22 Prestige Park Circle
East Hartford, CT 06108

Scangraphics
700 Abbott Drive
Broomall, PA 19008

Scan Vantage
P.O. Box 7067
Overland Park, KS 66207

Scitex
8 Oak Park Drive
Bedford, MA 01730

Seiko Elect. Imaging
1130 Ringwood Court
San Jose, CA 95131

Sharp Electronics
Sharp Plaza
Mahwah, NJ 07675

Skan Technologies
401 Church Street

Cedarhurst, NY 11516

SONY Imaging
One Sony Drive
Park Ridge, NJ 07656

Summagraph
8500 Cameron Road
Austin, TX 78754

Sunrise Imaging Inc.
42701 Lawrence Place
Fremont, CA 94538

TEC America
2710 Lakeview Court
Fremont, CA 94538

Terminal Data Corp.
5898 Condor Drive
Moorpark, CA 93021

The Complete PC
830 Hillview Ct #150
Milpitas, CA 95035

Toshiba Elect. Imaging
9740 Irvine Blvd.
Irvine, CA 92648

Truvel Corp.
8943 Fullbright Ave.
Chatsworth, CA 91311

UMAX Technologies
3353 Gateway Blvd.
Fremont, CA 94538

Unisys Corporation
P.O. Box 500
Blue Bell, PA 19424

VEMCO
305 S. Acatia Street
San Dimas, CA 91773

VIDAR Systems Inc.
520 D. Herndon Pkwy.
Herndon, VA 22070

Visionshape
1434 West Taft Ave.
Orange, CA 92665

Wang Laboratories Inc.
One Industrial Ave.
Lowell, MA 01851

Wicks & Wilson Ltd.
555 I West Lambrt Rd.
Brea, CA 92621

X-Ray Scanner Corp.
1687 E. Del Amo Blvd.

Carson, CA 90746

Xerox Engineering Sys
5853 Rue Ferrari
P.O. Box 210061
San Jose, CA 95151

Xerox Imaging Systems
9 Centennial Drive
Peabody, MA 01960

xillix Technologies
Suite 200
2339 Colombia St.
Vancouver, BC V5Y 3Y3

XL Vision
10300 102nd Terrace
Sebastian, FL 32958

XRS Scanner
4030 Spencer Street
Torrance, CA 90503

Contents

Appendix J
Test Charts

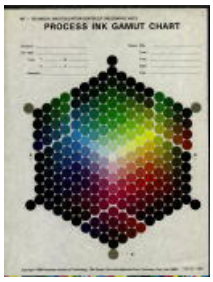
This section shows JPEG-compressed example scans of the test charts that were captured at the start of each day's scanning. These examples are from 12/7/95.

IEEE-167A Facsimile Test Chart

- Screen width grayscale as progressive JPEG, 50 KBytes
- Full size grayscale as JPEG, 666 KBytes

AIIM MS-44 Test Chart No. 2

- Screen width grayscale as progressive JPEG, 39 KBytes
- Full size grayscale as JPEG, 802 KBytes



Rochester Institute of Technology (RIT) Process Ink Gamut (PIG) Chart

- Screen width color as progressive JPEG, 47 KBytes
- Full size color as JPEG, 884 KBytes

Contents

1. One set of the Phase I sample images was produced using a special high-quality printer and these paper copies may be examined in the NDLP office at the Library of Congress. This electronic version presents selected digital files as hyperlinked illustrations; Appendix A provides access to the entire set of sample images. The Phase I investigation also compiled an extensive set of descriptive and technical literature from companies that manufacture scanners and create special imaging software; this body of literature is summarized in the appendixes and may also be examined in the NDLP office.

Return to Text

2. A task force, co-chaired by Donald Waters and John Garrett, on Archiving of Digital Information was formed in 1994 by the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG). It investigated and recommended means to ensure "continued access indefinitely into the future of records stored in digital electronic form." The final report of the 21-member task force, Preserving Digital Information: Final Report and Recommendations, is available on the RLG web site. This report provides an overview of the organizational and technical issues involved in migrating data from near-obsolescent formats or media (or media whose software drivers, operating system environment, or hardware is near-obsolescent) to new formats or media.

Return to Text

3. A sequence of projects at Cornell University has progressively refined an approach to the digitization of printed books using 600 dpi binary images developed from 400 dpi resolution scanners. Disbound books are replaced with laser-printed and bound editions and a preservation microfilm copy created on an electron beam recorder from the same digital source images. For further information, see the works by Anne R. Kenney and collaborators in the References section.

Return to Text

4. The writers of this report believe that digital preservation and reformatting will be accepted by libraries and archives when both questions of image quality and questions of longevity are answered. The organizational and technical issues involved in migrating data away from near-obsolescent media (or media whose software drivers, operating system environment or hardware are near obsolescence) or in refreshing data whose media is experiencing deterioration (as indicated by an increase in monitored error rates) are specifically not addressed in this project.

   A project which should be undertaken by the preservation community is to coax drive manufacturers to make available externally to their units the detailed error detection and correction (EDC) information which already exists within them. This could then be used to monitor trends in soft (recoverable) error rates and provide a basis for migration schedules long before hard (unrecoverable) errors begin to occur. In the networking world, the SNMP (Simple Network Monitoring Protocol) standard permits remote monitoring of the error rates and performance of devices like modems and routers for similar reasons.

Return to Text

5. The size of an uncompressed file can be calculated by multiplication. In the two examples provided, determining the number of "dots" or pixels is identical to determining the area of a rectangle. Multiplying 8.5 inches times 300 dots per inch yields 2550 as the number of pixels in each horizontal row; 11 time 300 yields 3300 for number of rows. The "area" is 2550 times 3300, or 8,415,000 dots or pixels. If 8 bits (1 byte) of information are captured for each pixel, the uncompressed file will have an extent of 8.4 million bytes; if 24 bits are captured, the size triples to 25.2 million bytes. For comparison, an uncompressed binary file (1 bit per pixel) will have one-eighth the extent of the 8-bit file, or about 1 million bytes. The term megabyte is used to refer to 1024 times 1024 bytes (1,048,576), so 8,415,000 is only 8.0 megabytes.

Return to Text

6. For examples of trade press articles, see Bruce Fraser, "Color Under Control," in *Adobe Magazine*, September/October 1995, pp. 41-45, or Andrew Rodney, "Desktop Color Management," in *Photo-Electronic Imaging*, January 1997, pp. 40-42. The IPI report is *Recommendations for the Evaluation of Digital Images Produced from Photographic, Micrographic, and Various Paper Formats.*

Return to Text

7. A group at the library at the University of California, Berkeley, is testing what they call ebind. A paging approach has been applied to a set of serials in the Making of America project organized by Cornell University and the University of Michigan. The Library of Congress has prepared the Walt Whitman notebooks collection in a similar manner.

Return to Text