1999

Digital Imaging and Preservation Microfilming: The Future of the Hybrid Approach for the Preservation of Brittle Books

Conway, Paul; Chapman, Stephen; Kenney, Anne R.

Washington, DC: Council on Library and Information Resources, 1999
http://hdl.handle.net/2027.42/149488

Digital Imaging and Preservation Microfilm: The Future of the Hybrid Approach for the Preservation of Brittle Books

Stephen Chapman, Harvard University
Paul Conway, Yale University
Anne R. Kenney, Cornell University

I. INTRODUCTION

We are nearing the end of a decade of intensive investigation into the use of digital imaging technology to reformat a range of library and archival materials. This effort has in part been stimulated by the phenomenal growth in network access capability, principally spurred by the advent of the World Wide Web. The effort, in part, also finds its roots in the cooperative microfilming projects the Research Libraries Group (RLG) initiated in the mid-1980s and funded by NEH; in the formation of the Commission on Preservation and Access (CPA) in 1986; and in the 20-year brittle books program that the National Endowment for the Humanities (NEH) launched in 1989 at the request of Congress. These initiatives promoted wide acceptance of a definition of preservation as prolonging the life of information *in* documents, rather than the documents themselves when the documents could not be preserved in their original forms.

Following a perceived consensus in the field, NEH has considered microfilm the preferred preservation choice for embrittled published materials and an acceptable access option, although some view digital imaging as an attractive alternative. A number of the earliest imaging projects supported by the Commission on Preservation and Access focused on digitization for preservation as well as access. Despite predictions that microfilm could be replaced by digital imaging,[1] early users of this technology came to appreciate that simply digitizing material did not guarantee its continued preservation. "Being digital means being ephemeral," Terry Kuny concluded in an article entitled "The Digital Dark Ages?"[2] Concern over digital longevity prompted RLG and CPA to collaborate in producing a highly influential report, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* This report presented the clearest articulation of the problems associated with digital preservation, and galvanized a number of institutions and consortia both within the United States and abroad to consider finding ways to assure the safekeeping and accessibility of digitized knowledge to be among their highest priorities. Despite this attention, to date there is no universally agreed upon technological approach or institutional/consortial capability in place to guarantee continuing access to digitized materials of enduring value. As such, microfilm remains the preferred preservation reformatting strategy even as digital imaging has assumed a prominent role in enhancing access to such materials.

This working paper examines the dual use of microfilm for preservation and digital imaging for enhanced access in the context of the brittle books program. It seeks to build on work that has already been accomplished, principally through projects conducted at Cornell University and Yale University; to propose a hybrid strategy; and to raise questions and suggest means for answering them before such a strategy can be broadly implemented. Support for this paper comes from the three principal advocates

[1] See in particular, Eldred Smith, "Why Microfilm Research Library Collections when Electronic Data Bases could be Used?" *Chronicle of Higher Education* (July 18, 1990): A44.

[2] Terry Kuny, "The Digital Dark Ages? Challenges in the Preservation of Electronic Information," *International Preservation News*, 17 (May 1998): 8-13.

of investigations into the duality of microfilm and digital imagery: the Council on Library and Information Resources, the National Endowment for the Humanities, and the Research Libraries Group, Inc.[3]

ASSUMPTIONS UNDERPINNING THE SCOPE OF THIS PAPER

- Reformatting remains the only viable long-term strategy for dealing with the preservation problems posed by brittle paper. Although there may be strong incentives to retain the original volumes for as long as possible, they should be copied to ensure that knowledge they contain will survive.
- Until digital preservation capabilities can be broadly implemented and shown to be cost-effective, microfilm will remain the primary reformatting strategy for brittle books. Microfilm offers acceptable levels of quality, media longevity, little machine dependency, and the means for producing additional copies with acceptable informational loss. Although digital imaging can be used to enhance access, preservation goals will not be considered met until a microfilm copy or computer output microfilm recording of digital image files has been produced that satisfies national standards for quality and permanence.[4]
- Recommendations presented in this paper will be limited to brittle monographs and serials containing monochrome text and simple graphics. We will further restrict our discussion to microfilm that meets current recommended standards—in other words, film produced from the mid-1980s onward or film to be created today as part of a hybrid effort. We acknowledge that the problems of brittle paper extend beyond these formats, but such problems will be our starting point because we can draw on work already completed to provide definitive recommendations.
- Only strategies that are both quality-oriented and cost-effective will be recommended. As such, this paper will focus on the use of high contrast microfilming and bitonal digital imaging.
- We will present options for both *film-first* and *scan-first* strategies, providing guidance to institutions in determining the best course of action based on their particular collections, capabilities, and needs.

[3] The authors wish to thank in particular the following individuals: Abby Smith and Deanna Marcum of the Council on Library and Information Resources for editorial and financial support of this paper; George Farr and Charles Kolb of the National Endowment for the Humanities for their encouragement to pursue the next steps after the conclusion of these research projects; and Robin Dale and Nancy Elkington of the Research Libraries Group for their willingness to initiate follow up work to the Cornell and Yale studies.

[4] See for instance, Nancy Elkington, editor, *RLG Preservation Microfilming Handbook*, (Mountain View, CA: The Research Libraries Group, Inc., 1992); ANSI/AIIM MS23-1998, *Practice for Operational Procedures/Inspection and Quality Control of First-generation, Silver Microfilm and Documents,* (Silver Spring, MD: Association for Information and Image Management).

II. WHAT IS THE HYBRID APPROACH?

The marriage of microfilm and digital technologies has been a part of the information technology landscape for over fifty years. The visionary computer pioneer, Vannevar Bush, suggested in his oft-cited 1945 article "As We May Think" that much of the world's knowledge could be stored on microfilm in something akin to a mechanical jukebox and retrieved through computerized searching techniques.[5] In 1992, renowned microfilm expert Don Willis drew upon developments in the infant technology of mass digital storage to suggest the possibility that microfilm and digital technologies could be combined to meet the needs of both archival storage and digital access. "By taking advantage of the strengths of film combined in a hierarchical system with the access capabilities provided by digital imaging," Willis concluded, "a preservation system can be designed that will satisfy all known requirements in the most economical manner."[6]

Willis argued that scanning microfilm was already technically possible—and was the least risky preservation option in 1992—but that scanning directly from original source documents and then backing up the digital data on computer output microfilm (COM) was also feasible. Ultimately, he suggested that scanning first would prove to be the most flexible and efficient way to create high-quality digital products while taking care that preservation concerns were met.

Embedded in *A Hybrid Systems Approach to Preservation of Printed Materials* were assumptions Willis made about the quality of microfilm and digital products produced either through the film-first or the scan-first route. The report includes clear but untested arguments about the costs—and cost-effectiveness—of the hybrid systems approach. The real issue, Willis concluded, would be determining the circumstances under which either approach should be pursued. The Commission on Preservation and Access and the National Endowment for the Humanities agreed, and provided support to Cornell and Yale universities over a five-year period to test the assumptions outlined in Willis' important report.

YALE UNIVERSITY'S PROJECT OPEN BOOK

Project Open Book (1991-96) was a multifaceted, multiphase research and development project. Its purpose was to explore the feasibility of large-scale conversion of preservation microfilm to digital imagery by modeling the process in an in-house laboratory. The project unfolded in a sequence of phases designed in part to allow the project to evolve as the digital imaging marketplace changed. In the organizational phase,Yale conducted a formal bid process and selected the Xerox Corporation to serve as its principal partner in the project. During the set-up phase, Yale developed a single integrated conversion workstation that included microfilm scanning hardware and associated conversion and enhancement software, tested and evaluated this workstation, and made the transition to a fully-engineered production system. In the final production-conversion phase, Yale built a workstation

[5] Vannevar Bush, "As We May Think," *Atlantic Monthly* 176 (July 1945): 101-07 [Online]. Available: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm.

[6] Don Willis, *A Hybrid Systems Approach to Preservation of Printed Materials* (Washington, D.C.: Commission on Preservation and Access, 1992), 14 [Online]. Available: http://www.clir.org/pubs/reports/willis/index.html.

conversion network, hired technical staff, converted 2,000 volumes from microfilm (representing 440,000 images), indexed the volumes, stored the results, and tested a prototype Web access tool developed by Xerox.[7]

CORNELL UNIVERSITY'S DIGITAL TO MICROFILM CONVERSION PROJECT

Cornell University Digital to Microfilm Conversion Project (1994-96) was one of a sequence of research and development projects commencing in 1990 that explored the feasibility of adopting digital technology for preservation purposes. The two-and-a-half year demonstration project tested and evaluated the use of high resolution bitonal imaging to produce computer output microfilm (COM) that could meet national preservation standards for quality and permanence. In the course of the project, 1,270 volumes and accompanying targets (representing 450,000 images) were scanned and recorded onto 177 reels of film. All paper scanning was conducted in-house; Cornell contracted the production of COM to Image Graphics, Inc. of Shelton, Connecticut. The project led to an assessment of quality, process, and costs, and to the development of recommendations for the creation and inspection of preservation quality microfilm produced from digital imagery.[8]

Both Cornell and Yale recognized the significance and complementary nature of each other' s projects. The projects had in common:

- Relying on high quality 35mm microfilm as the preservation master
- Creating approximately the same number of high quality digital images from similar collections of nineteenth and twentieth century brittle books
- Developing a high-production, in-house scanning operation
- Regularizing procedures for quality control in scanning

---

[7] Donald J. Waters, *From Microfilm to Digital Imagery* (Washington, D.C.: Commission on Preservation and Access, June 1991); Waters and Shari Weaver, *The Organizational Phase of Project Open Book* (Washington, D.C.: Commission on Preservation and Access, September 1992) [Online]. Available: http://www.clir.org/pubs/reports/openbook/openbook.html; Paul Conway and Shari Weaver, *The Setup Phase of Project Open Book* (Washington, D.C.: Commission on Preservation and Access, June 1994) [Online]. Available: http://www.clir.org/pubs/reports/conway/conway.html; Conway, "Selecting Microfilm for Digital Preservation: A Case Study from Project Open Book." *Library Resources & Technical Services* 40 (January 1996):67-77; Conway, "Yale University Library' s Project Open Book: Preliminary Research Findings," *D-Lib Magazine* (February 1996) [Online]. Available: http://www.dlib.org/magazine.html; Conway, *Conversion of Microfilm to Digital Imagery: A Demonstration Project* (New Haven, CT: Yale University Library, 1996).

[8] Anne R. Kenney, "Digital-to-Microfilm Conversion: An Interim Preservation Solution," *Library Resources & Technical Services* (Oct 1993): 380-402, (January 1994 erratum): 87-95; Kenney and Lynne K. Personius, A Testbed for Advancing the Role of Digital Technologies for Library Preservation and Access (Washington, D.C.: Commission on Preservation and Access, October 1993): 19-26; Kenney and Stephen Chapman, *Digital Imaging for Libraries and Archives* (Ithaca, NY, Cornell University Library, 1996): 179-186; Kenney, *Digital to Microfilm Conversion: A Demonstration Project, 1994-1996. Final Report to the NEH* (Ithaca, NY, Cornell University Library, 1997), [Online]. Available: http://www.library.cornell.edu/preservation/pub.htmhtm (hereafter cited as the COM final report); Kenney, "The Cornell Digital to Microfilm Conversion Project: Final Report to NEH," *RLG DigiNews* 1:2 (August 15, 1997) [Online]. Available: http://www.rlg.org/preserv/diginews/diginews2.html.

- Using the same basic technology for indexing (metadata creation) and file management
- Collecting and comparing data on costs, production, and quality

The Cornell and Yale projects had similar goals but there were some distinctive differences in implementation between the two efforts. Cornell's project may be characterized in the context of prospective conversion of brittle paper: how to exploit available technologies to create microfilm that meets preservation objectives and digital images that meet access objectives in the most cost-effective manner. Yale's project fits into the context of retrospective conversion of extant microfilm: how to exploit available technology to create digital images that meet a full range of access objectives in the most cost-effective manner.

III. ISSUES AFFECTING QUALITY, COST, AND ACCESS

The research projects at Yale and Cornell addressed digital image conversion of text-based materials and the production of archival-quality microfilm. This microfilm is stored as a "permanent" replacement of the brittle book, and also used as a source for image conversion and/or as a backup to digital images if they are lost in a disaster. As the two projects revealed, the relationship of film to digital lies in aligning quality, cost, and access in terms of three underlying concepts. These include: (1) the characteristics of the source material being converted; (2) the capabilities of the technology used to accomplish the digital conversion; and (3) the purposes or uses to which the digital end product will be put.

1. THE CHARACTERISTICS OF THE SOURCE MATERIAL BEING CONVERTED

The first challenge in choosing the path from analog to digital is to understand the relationship between the technology of digital image conversion and the analog resources to be transformed. In a brittle books application, the three most important aspects are:

- the format of the source (including size of object, its structure, and its physical condition)
- visual characteristics (including the centrality of text versus illustration), and
- the level of detail (including the size and style of type faces, and the type of illustrative content).

For the purposes of this study, we assume that brittle books consisting of text (font sizes as small as 1mm in height) and simple line art or halftones (with no color information) can be reproduced successfully using high-contrast microfilm or high-resolution bitonal scanning.

2. THE CAPABILITIES OF SCANNING TECHNOLOGY

Another key to understanding the relationship of analog to digital is to measure the capabilities of the digital imaging hardware/software system against the purposes to which the images will be placed. The

expected uses of the product drive the level of detail that must be captured from the source material. In the course of this working paper, we will differentiate between two different digital products: a digital access master and a digital preservation master. In the case of the former, the overriding functional requirement is to meet a full range of user needs in the electronic environment, now and in the future. In the case of the latter, the digital product must also be of sufficient quality so that it can be used to create COM that meets national standards for quality and permanence. The key distinction between these purposes is the level of detail and tonal representation that must be captured from the source material. Digital files created with the intent of producing analog (eye-readable) versions that meet contemporary archival standards place the highest demands on digital capture technology.

Although the expected uses of the product may drive the choice of technological applications, the converse is not necessarily true. It is important to recognize that standards and best practices developed to support both access and preservation masters should not be driven by the present limitations of digital image capture, display, and output. Matters such as the limited resolution of today' s display screens, the limited bandwidth of wide and local area networks, and the limitations of resolution and tone reproduction in printers should not set the quality thresholds of image system design.

3. THE PURPOSES THE DIGITAL IMAGES MUST SERVE

The third issue at work in the hybrid approach is the relationship between the characteristics of the source documents and the use requirements for the digital images. The most important aspect of this relationship turns on the clear understanding of what needs to be represented in digital form. In the case of brittle volumes, there are two perspectives. The first concerns the appearance of the document at the time it is *converted* (including an accurate portrayal of blemishes, stains, tears, and other evidence of past use or damage). The second concerns the appearance of the document when it was *created*, allowing for the use of digital enhancement techniques to reverse the effects of fading, water damage, image loss, and the like. Reference to the original document when representing it in digital form also relates to questions of the completeness of the digital version (for example, should blank pages in the work be converted) and the extent to which a facsimile copy on paper is a requirement of the digital version. Ultimately, the conversion from microfilm to digital entails some degree of loss; defining the level of acceptable loss will remain a challenge.

The position taken on the issue of representation of the original printed material has many practical consequences for the characteristics of the digital product, particularly when microfilm represents the source material for scanning. These range from the presence or absence of data depicting the physical border of the original document to the accurate representation of the dimensions of the original pages to the acceptability of sophisticated digital enhancement tools to improve the quality of the end result. Additionally, the relationship between purpose and source characteristics may influence the choice of materials in terms of their intellectual content, visual characteristics, and physical attributes.

The relationships among source characteristics, technology capabilities, and the purposes of the end product bear upon the definitions of quality, cost, and access. In the area of quality, for example, an input source with particular characteristics (such as high-contrast, 35mm, black & white microfilm), the limitations or costs of scanning technology at a given point, and the expected uses of the product interact

to set the threshold requirements for image quality. Similarly, the expected purposes of the digital product (for example, preservation replacement) and the characteristics of the source (for example, brittle books) interact with imaging technology capabilities to determine the cost of creating the product with the intended purpose. The same is true for access. The intellectual complexity of the source documents and the specification for the ways in which the image product will be used interact with the hardware and software tools for building metadata files to define access parameters.

IV. RESEARCH ISSUES TO BE ADDRESSED

The Yale and Cornell projects speak to the relationships of quality, cost, and access through their joint exploration of four issues:

1. the characteristics of microfilm as a source for digital conversion;
2. the characteristics of microfilm as an end-product of digital conversion;
3. the choice of a digital conversion path (film-first or scan-first); and
4. the development of metadata elements associated with the digital image product.

RESEARCH ISSUE 1:

THE CHARACTERISTICS OF MICROFILM AS A SOURCE FOR DIGITAL CONVERSION

In this section we will discuss issues associated with quality and cost in scanning from preservation microfilm. The Yale project scanned microfilm that met national standards for quality and permanence. We will discuss this project' s findings as well as consider issues associated with creating new microfilm that may be digitized in the future. The primary question to be addressed is: will modifying existing microfilming guidelines make it cheaper to scan from film and/or make it more possible to generate a higher quality digital product?

The creation of preservation microfilm since the early 1980s has been governed by a well-defined set of international standards that specify the preparation of documents, bibliographic control, the physical composition of the film media, processing techniques, the visual quality of three generations of film, and storage requirements. With the publication of its *Preservation Microfilming Handbook* in 1992, RLG contributed procedural guidelines that expanded upon international standards, helping to assure that preservation microfilm is created consistently, stored properly, and that access to preservation microfilm is improved.

The findings of Yale' s Project Open Book suggest that modest modifications to the Research Libraries Group guidelines may result in preservation microfilm that produces better quality digital image products but that the costs incurred in creating such film will not be recouped through reduced digital conversion costs.[9] If quality is a proportionately greater concern than cost, these modifications may be worth the

[9] An investigation underway at Harvard University is testing this premise. Within the scope of its current NEH project to microfilm collections in the history of science, Harvard is applying several of the recommendations from Project Open Book to regularize the placement of images on film. In addition, project staff are creating electronic files

effort. Ultimately, future developments in digital technology—such as affordable grayscale scanning capabilities—may offer far greater promise to increase quality and reduce cost than any specific modifications in the creation of preservation microfilm.

Recommendations from Project Open Book

Specific recommendations from the Yale project follow, organized in terms of cost reduction and quality improvement strategies.

A. Decreasing the Cost of Converting Microfilm

One of the most important components of Yale' s Project Open Book was a multi-faceted analysis of the costs of microfilm scanning and the factors that influence conversion costs. The study investigated the impact on conversion cost of thirteen characteristics about the books included in the project and eight characteristics of the 35mm microfilm.

As the Cornell project found, book characteristics, such as the presence of halftones, tight gutters, yellowed or faded paper and inks, and similar factors associated with deterioration, damage, or heavy use, can increase the costs of the digital imaging conversion process. There may be very little we can or should do about this, however, when beginning with film because the process of selection for digital conversion cannot and should not drive the preservation imperative. The choice to reformat a brittle book or journal on microfilm should be made because film is the best way to extend the life of the information contained in these items. In the Yale Project, books were preserved on microfilm because of their informational value, not their physical appearance. Yale made no effort to improve the images on problematic books because of the "production-converson" nature of the project.

The following table, excerpted from the final report on Project Open Book, provides the details on the impact of film characteristics on process time. In the Yale model, time equals cost. An "X" in a particular column indicates that a given characteristic has a statistically significant impact on the cost of a given process step. The ten major steps in the conversion process are abbreviated above each column. The steps are: (1) **inspecting** the film before scanning, (2) **benchmarking** the film for scanning quality, (3) **setting up** the scanner software, (4) **scanning** the film in automatic mode using special edge detection software, (5) initial **quality control**, (6) assigning **page** numbers in an associated index, (7) establishing and tagging the **structure** of a volume in a relational database, (8) secondary **quality control**, (9) database **registration** of the completed image file, and (10) **file transfer** activities associated with the management of the conversion process.

---

to "index" the books and journals as one of the preparation steps preceding filming. Early findings suggest that the additional costs of filming are insignificant, but those related to indexing are meaningful and will need to be recouped in scanning. In essence Harvard seeks to prove that the aggregate costs of creating microfilm and digital images can be lowered by making modest improvements in microfilm, and by combining digital metadata creation with microfilm preparation activities. It is hoped that these investments will be fully recovered by eliminating the following activities in scanning: cropping images, enhancing scans of illustrations and/or foldouts, paginating individual images, and indexing digital books and journals. A report detailing project findings, including costs, will be available in the summer of 1999.

TABLE 1.

Impact of Film Characteristics on Process Time										
	Processing Steps									
	1	2	3	4	5	6	7	8	9	10
	IN SP	B E N C H	SE T U P	S C A N	Q C	P A G E	S T R U C	Q C	R E GI S	FI LE
Film Characteristics										
Contrast/density variation (92.0%)										
Skewed pages (76.2)	X			X	X	X				X
Inconsistent gutter (19.7)	X									
Internal splices (6.2)	X	X	X			X		X	X	X
Other film factors (16.5%)										
Reduction ratio	X			X	X	X	X		X	X
Cleanliness (dMin)										
Average density (dMax)		X		X				X		X

The table shows that contrast and density variation, which was present in 92 percent of the books on film in the study, had no measurable impact on the timing of any of the ten process steps. Skewed pages, detectable in 76.2 percent of the film, affected the cost of inspecting film prior to scanning and also had a noticeable impact on scanning, quality control, and the process of assigning page numbers.[10] Evidently, skewed pages generate more data than properly aligned pages, accounting for the increased cost of file transfer. Inconsistent gutter margins that result when a book disbound prior to filming is not aligned or centered consistently by the camera had an impact only on the inspection process. Internal splices had a statistically significant impact on seven of ten processes, yet they rarely occurred. The reduction ratio of the film, however, was a particularly important factor. As the ratio increased above 10:1, the costs of inspection, scanning, quality control, assigning page numbers, indexing the structure of the book, and final acceptance routines all became more expensive. The clarity of the film (dMin) made no difference in the scanning process. Finally, although density variation had no impact, if the average density of a given volume was less than .90, there were noticeable increases in the cost of benchmarking, scanning, quality control, and the size of the image file.

Observing the table vertically rather than horizontally yields additional findings. Few microfilm characteristics had any appreciable impact on the most time-consuming (i.e., costly) image conversion processes. Skewed pages, higher reduction ratios, and greater average density readings combined to increase the cost of the scanning process. Skewed pages, internal splices, and increased reduction ratios combined to increase the cost of assigning page numbers to the digital files.

[10] For purposes of the study, any variation of .1 or greater across the length of the film for a single book was considered to have "contrast and density variation." None of the books identified for image conversion exhibited skew on the film in excess the amount allowable in the RLG guidelines. Noticable skew was determined by inspecting the film on a light table without magnification.

These findings indicate that the characteristics of the film had little or no impact on conversion costs in the Yale project. They suggest that investments to improve the quality of new film may not be recouped through reduced conversion costs. The most cost-effective conversion of existing microfilm will result when selection takes place from a large pool of preservation-quality film created without expectation of digital conversion. Modest changes to the RLG guidelines -- for example, reducing skew, lowering reduction ratios, or reviving the use of blipping (see below) -- should lead to improved quality and more cost-effective film scanning. Whether the additional costs associated with making improvements at the point of microfilming can be offset by lower scanning costs should be examined.

Technology Solutions

The greatest promise for improvements in the cost of the digital conversion process resides in improved technology to reduce dramatically the times associated with scanning and indexing. Those improvements would be to:

1. Utilize appropriate computing and networking capabilities to avoid slow downs in data transfer.

2. Create software-assisted processing tools that routinize low-level tasks (such as setting scanner filter parameters for the entire reel, or automating the process of deleting microfilm targets), and move as much of the file transfer process "off-line" as possible.

3. Develop continuous scan techniques that minimize the need for scanner set-up and that eliminate the present reliance on edge-detection techniques that are prone to costly error, especially when text and illustration are present on the same page.

4. Develop software that semi-automates paginating digital images.

Process Considerations

Beyond the potential contribution by new technology, two additional modifications in the process of microfilm conversion hold promise to reduce conversion costs.

1. Select materials on high-quality preservation microfilm that lend themselves to high-quality digital conversion. Quality requirements can drive cost variables while the opposite equation (cost driving quality) may not always apply.

2. Acknowledge the benefit that a skilled, highly-trained production team can provide. Recognize and measure the learning curves of all parties involved in the conversion process and budget for production with fully trained technicians. This may best be achieved by outsourcing film scanning to reliable service bureaus that understand the needs of cultural institutions.

B. Improving the Quality of the Digital Image Product

The findings of Yale' s Project Open Book suggest two clusters of recommendations concerning the creation of new microfilm that could improve the likelihood of producing better quality digital image products. The first set of recommendations concerns the quality of the individual images. The second set

of recommendations pertains to what we choose to call the "technical rigor" of the film. Cumulatively, the recommendations do not challenge the primacy of international standards governing the creation of preservation microfilm. The recommendations suggest minor enhancements to such standards, particularly in the area of targeting. Similarly, the recommendations largely suggest the need to reduce some of the flexibility that is built into the RLG guidelines for creating preservation microfilm.

Quality of Individual Images

1. Polarity: scanning duplicate negative microfilm (never master negative) yields higher quality images than scanning positive film.[11]

2. Density: the maximum density (Dmax) for medium contrast (Dmax of .90-1.10) to high contrast (Dmax of 1.00-1.30) film results in higher quality images using bitonal scanning than low contrast (Dmax of .80-1.00) negatives. RLG minimum density guidelines (< .10) holds.[12]

3. Reduction ratio: orient material on film to obtain lowest possible ratio.[13]

Technical Rigor of the Microfilm Product

1. Consistent placement: minimize or eliminate "centerline weaving."

2. Skew: minimize or eliminate—no greater than 2 degrees from parallel.

3. Splices: internal splices compound the difficulties of film scanning and suggest that splices inside a given volume be eliminated. This practice would no doubt increase the cost of filming. Additional investigations are needed to determine whether the total cost of creating film and digital images would be less if greater rigor were demanded in the filming stage.

4. Duplicate images: duplicate frames created in the microfilming process to improve the quality of the image on the film have minimal negative impact on the ultimate quality of the digital product. Scanner operators will have to select the most appropriate frame for the retention in digital form and delete any duplicate images as part of the quality control process.

5. Blank frames: no recommendation on best practice on this important issue is possible at this time. The decision to retain or delete digital images of blank pages in the original book or empty frames

[11] The use of negative duplicate film is also recommended by the Working Group of the German Research Council. See: Hartmut Weber and Marianne Dorr, *Digitization as a Method of Preservation? Final Report of a Working Group of the Deutsche Forschungsgemeinschaft*, (Washington, D.C. and Amsterdam, Commission on Preservation and Access and European Commission on Preservation and Access, 1997): 5.

[12] The Working Group of the German Research Council recommended increasing the contrast between the background and the material to be filmed in order to expedite the detachment of the background material from the whole digitized image. Weber and Dorr, 7.

[13] For oversized material, filming one page per frame in the IA position will result in the lowest reduction ratio possible.

on the microfilm hinges on two issues: whether or not a paper facsimile of the original book must be produced seamlessly from the digital preservation master; and the importance of representing the look and feel of the original book in digital form.

6. Reduction ratio: accurate recording of reduction ratio is crucial for reproduction at original size.

7. Dimensions of original: record accurately on bibliographic target, particularly when variable reduction ratios are used as it is necessary to know the original page dimensions in order to compute the exact reduction ratio.

8. Test charts: incorporate RIT Alphanumeric Test Chart and Kodak Gray Scale; seek additional advice from vendors and imaging scientists on the use of Modulation Transfer Function targets.[14]

Technology Solutions

Ultimately, the findings of Project Open Book suggest that future improvements in the quality of digital image products created from microfilm sources depend more upon technology advances than on the characteristics of microfilm. Among possibilities, there are four areas that hold promise for near-term quality enhancements. Close cooperation between the imaging technology community and imaging product developers in libraries, archives, and museums is needed to advance the capabilities and efficiency of the technology of scanning.

1. Automatic calibration of scanners: A significant variable that determines the quality of the digital image converted from a microfilm frame is the human intervention needed to set up the scanning equipment for optimal quality. Set-up is not only time-consuming, but is fundamentally subjective in nature. The scanner operator must continually resolve questions about the settings of any given scanner vis á vis the display on any given screen or the hard copy that emerges from a print device. Software that can automatically optimize for data capture from microfilm would greatly reduce the subjective nature of the scanner set-up process, decrease the time required to scan microfilm, and result in a more consistent image product (assuming, of course, that the microfilm input source has the kind of technical rigor specified above).

2. Continuous scanning and post-scan processing: Another technical limitation in the achievement of consistent high quality image conversion from microfilm is weaknesses in current edge detection software that determines where a frame-image begins and ends. Edge detection software may be easily "confused" by the presence of dense illustrations, shifts in frame size (due to changes in page size or reduction ratio), and similar irregularities common in microfilm of brittle books. One solution is incremental improvements in the "intelligence" of edge detection software. A more radical solution may be to abandon edge detection altogether and produce a continuous image data stream from a roll of microfilm that can then be segmented into individual images through post-scan data processing.

[14] Don Williams, "What is MTF…and Why Should You Care?" *RLGDigiNews*, February 15,1998, Volume 2, No.1. [Online]. Available: http://www.rlg.org/preserv/diginews/diginews21.html#technical.

3. Post-scan image splitting: Depending on the orientation of the book on the film, the production of individual digital images that correspond to individual book pages is more or less complicated. In the so-called "IIA" (cine) orientation of a book, the spine of the book runs parallel to the edge of the film and two book-pages are captured in every frame of film. In "IIB" (comic) orientation, the book is rotated 90 degrees so that the spine is perpendicular to the edge of the film and two book-pages are captured in every frame. In Project Open Book, the vast majority of the books converted were filmed in the "IIA" orientation. The scanner was outfitted with special hardware and software components that resulted in the creation of one digital image for every book-page at a higher resolution than could have been achieved had both pages been captured at once. Microfilm created in the "IIB" orientation requires post-scan processing to split a single image of two book-pages into two discrete digital images. Technological improvements in image-splitting designed to automate and improve the accuracy of the process of creating single book-page images would result in dramatically improved product quality at decreased cost.

4. Blipping: The marking of microfilm to indicate pagination, the beginning and ending of a given book, as well as internal transitions (e.g., chapter breaks) is an old fashioned technology now being given a second-look. One goal of blipping with digital imaging in mind would be to assist in the automation of index-level metadata that now must be created in a time-consuming (and error-prone) manual process. To date, no rigorous testing of modern blipping techniques has been undertaken in the United States.[15] Another goal is to use blipping to note frames that must be rescanned to achieve consistencies in image quality. One example is the frame that contains a complex illustration that would be better captured in grayscale scanning; another is the always difficult foldout, which is larger than the images that immediately precede and follow it. The authors of this report, therefore, make note of the potentials of blipping technology and take no formal stand on its cost advantages and disadvantages.

RESEARCH ISSUE 2:

THE CHARACTERISTICS OF MICROFILM AS AN END-PRODUCT OF DIGITAL CONVERSION

In this section we will discuss issues associated with quality and cost in outputting digital images to COM that can meet preservation standards for quality and longevity. This discussion will begin with a presentation of Cornell' s findings, and conclude with recommendations governing the use of COM. The primary question to be addressed is: How should we specify the creation and inspection of digital image products from brittle books and journals and their subsequent placement on COM?

A. Issues Affecting the Quality of Computer Output Microfilm

The Cornell project showed that computer output microfilm created from 600 dpi 1-bit images scanned from brittle books can meet or exceed national microfilm standards for permanence and image quality.

[15] See Weber and Dorr, p. 6-8 on the use of blipping: "Filming with the use of blips is always necessary for an efficient working method with microfilm scanners."

Permanence

Permanence requirements were satisfied in that the film stock, processing, associated packaging, and storage conditions all met ANSI/AIIM standards. The 35mm film stock used was Kodak Image Link HQ; all reels passed third party inspection for residual thiosulfate concentration; and appropriate reels, fasteners, and boxes were used to store the film. The COM is stored under controlled environmental conditions in RLG' s vault at National Underground Storage in Boyers, PA.

Resolution

Achieving acceptable levels of image quality rested in the two-step process of converting original materials to COM:

- digitization—creating digital image files that adequately capture all the significant informational content of the original source materials, and

- COM recording—utilizing a COM system that is capable of recording faithfully onto film all of the information contained in the digital image files.

The quality of the COM will principally be determined by the quality of the initial scanning. Although there are no national standards governing image quality for digital files, Cornell University Library' s Department of Preservation and Conservation has spent nearly a decade analyzing digital conversion requirements for books published from 1850-1950. This work included scanning over 2.5 million images (in-house and via contract), a systematic review of 105 printers' type sizes commonly used by publishers during this period, and visual inspection of digital facsimiles for Roman and non-Roman scripts. Based on this experience, Cornell has concluded that a scanning resolution of 600 dpi 1-bit is sufficient to capture fully the monochrome text-based information contained in virtually all books published during the period of paper' s greatest brittleness. Illustrated texts—containing line art and halftones, for which photocopy or microfilm are considered adequate for replacement purposes—can also be captured using 600 dpi bitonal scanning with enhancements. For publications containing more complex illustrations that are essential to the meaning of the text or heavily deteriorated volumes, bitonal scanning, even at high resolution may prove to be inadequate—in those circumstances, grayscale or color scanning is recommended. As with other conversion processes, the quality of the resulting image files must be confirmed through a rigorous quality assurance program.[16]

[16] For information on defining digital conversion requirements for text-based materials, see: Kenney, "Digital-to-Microfilm Conversion: An Interim Preservation Solution," and Kenney and Stephen Chapman, *Tutorial: Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality* (Washington, D.C.: Commission on Preservation and Access, 1995). For recommendations on capturing halftones, see: Carl Fleischhauer, "Digital Formats for Content Reproductions," Library of Congress, July 13, 1998 [Online]. Available: http://memory.loc.gov/ammem/formats.html. Subsequent studies are addressing issues associated with more complex book illustrations and the presence of significant color. Cornell, Picture Elements, and the Library of Congress are conducting an investigation into the digital conversion requirements for nineteenth and early twentieth century relief, planographic, and intaglio book illustrations. The report of this project will be available in early 1999.

Having determined that 600 dpi bitonal scanning could produce digital files that faithfully rendered all textual information contained in brittle books, Cornell turned its attention to the quality of the computer output microfilm. The goal was to ensure that there was no loss of resolution or image quality in recording the digital images onto COM. Cornell used the RIT Alphanumeric Test Object, which consists of block characters and numbers represented in two directions, to measure the effective resolution achieved on the COM. Cornell staff also conducted subjective evaluation of the COM rendering of the smallest lower-case "e" contained in a volume, using the ANSI/AIIM Quality Index rating for microfilm inspection. Staff visually inspected the COM on a light box under 75x magnification. In all cases, the images met the "high quality" standard for Quality Index (8.0) in the rendering of the smallest "e." RIT target readings on the COM ranged from line 8 through line 15, which proved identical to those read on-screen during quality control of the digital images.[17]

Polarity, Density, and Placement

Cornell produced a first generation negative film that revealed remarkably consistent density, as well as spacing and placement. RLG standards permit a minimum density of no greater than 0.10. The minimum density values for all reels fell well within specifications, ranging from .02 to .04. Background densities ranged from .90 to 1.06, again within the acceptable range of .90 to 1.10 for medium contrast (appropriate for brittle books with moderately darkened paper). Density variation within titles ranged from .00 to.04, and between titles from .01 to .06, far below the maximum acceptable variation of .20. The images were recorded two images per frame in the IIA (cine) position. Spacing between images and between frames was uniform and consistent, and there was no detectable skew that was attributable to the COM recording.

Reduction Ratio

The Technical Advisory Committee to the Cornell Project approved the use of variable reduction ratios to "fill the frame" for each book.[18] This enabled Image Graphics to use the smallest reduction ratio

---

For information on the conversion of bound volumes via color scanning, see: "Producing Digital Images," *The Electronic Archive of Early American Fiction*, (Charlottesville, VA: University of Virginia Library, July 1998), [Online]. Available: http://www.lib.virginia.edu/speccol/mellon/image.html.

[17] The readings on the RIT target when scanned on the XDOD at settings optimized for its capture represented at least line 15 legibility in all four quadrants. However, when the settings optimized for the brittle books were used, the RIT readings differed considerably, with lower readings seeming to correlate to the capture of low density originals. The quality of the resulting COM was excellent in all cases. This led Cornell staff to suspect that the target was not a sufficiently accurate indicator of resolution when its density varied considerably from that of the original book. Many of these books exhibit low contrast between text and background. The RIT target used in this project was a high contrast target (density of 1.9). Cornell staff subsequently scanned three different versions of the RIT target with high density (1.9), medium density (1.3), and low density (.7) at various settings analogous to ones we would use to capture high, medium, and low contrast books. The best readings were uniformly observed on the low density (.7) RIT target, with the exception of the instance when the "autosegmentation" feature was used, which interpreted portions of the low density RIT target as a halftone and applied descreening and rescreening filters to it.

[18]Image Graphics achieved variable reduction ratios by recording all pixels across the width of an image onto 15mm of the film. There was a 3mm spacing between images in the 2A position, and 3 mm of space reserved between frames. The physical page dimensions of foldouts were recorded on the production note. If foldouts exceeded 11" x 17", they

possible, thus ensuring the highest recording of resolution on film, and to produce an extremely uniform product that potentially would facilitate the scanning back from COM if the original digital files ever became unreadable.[19]

The Committee approved the use of variable reduction ratios, provided that the dimensions of the original documents were recorded on a film target in order to reproduce paper facsimiles at the same physical dimensions as the original volume. Because file size information for each image was recorded in the TIFF header, a target noting the pixel dimensions (e.g., 2,400 x 3,600) and resolution (600 dpi) could be generated automatically from the TIFF header by the program for reel composition. With this information, one could then calculate the original page width by dividing the first pixel dimension by 600, e.g., the original page width for a 2,400 x 3,600 pixel image would be 4 inches (2,400 divided by 600 equals 4), and the length could be similarly calculated. COM recording at fixed reduction ratios is also possible, and is being used by Image Graphics in a contract with the Virginia State Archives.

Use of the Electron Beam Recorder

Cornell did not discern any drop in resolution or degradation in quality from the digital images to the microfilm copy. Given the capabilities of the Image Graphics COM system, the Electron Beam Recorder, to record extremely fine resolution with excellent image acuity, virtually all of the information in the 600 dpi 1-bit images could be represented on the 35mm microfilm at the reduction ratios used (between 5x and 10x). According to IGI product literature, the electron beam provides 10 times better resolution, 10 times faster speed, and 10 times greater dynamic range that traditional cathode ray tube imaging. It appears that other COM recording systems may not be able to match the capabilities of the IGI electron beam recorder in recording 600 dpi images in 2A position on 35 mm film.[20]

Recommendations for the Creation and Inspection of Computer Output Microfilm

Although COM can meet preservation microfilm standards, procedures for production and inspection of the COM will differ from those appropriate to conventional microfilm. Significant changes in film

---

were reduced via preservation photocopy and the photocopy scanned, excepting in cases where significant information would be lost by the reduction process. To maintain information on the actual size of the foldouts, and to calculate the reduction ratio used, the size of the reduced photocopy was also recorded (the pixels representing the smaller dimension of the foldout were always recorded on 32mm of film).

[19] Ron Whitney, Manager of Electronic Production, Primary Source Media, scanned the COM using the Sunrise SRI-50 film scanner. He noted that it was "a pleasure working with the film overall." Its consistent density and image placement resulted in "flawless edge detection and distinction between frames," and made film scanning "a snap." Care must be taken in scanning from film with variable reduction ratios so that original page dimensions can be recreated in printed facsimiles.

[20] The editors of *RLG DigiNews* surveyed COM service providers, and found no other company that could meet IGI's capabilities. See "Technical Review: Outsourcing Film Scanning and Computer Output Microfilm (COM) Recording," *RLG DigiNews* 1:2 (April 15, 1997) [Online]. Available: http://www.rlg.org/preserv/diginews/diginews2.html. This finding was also reached by the German Research Association, which evaluated some COM recording capabilities but not that of the electron beam recorder of Image Graphics. See: Hartmut Weber and Marianne Dorr, *Digitization as a Method of Preservation? Final Report of a Working Group of the Deutsche Forschungsgemeinschaft*, (Washington, D.C. and Amsterdam, Commission on Preservation and Access and European Commission on Preservation and Access, 1997): 19.

creation and quality control are introduced in COM recording. Images are generated digitally, not photographically, and factors affecting image quality, such as resolution and density, are made up-stream—at the point of scanning—and not at the point of filming. This has significant ramifications for final film inspection.

The quality of the resulting COM will in large measure be determined by the quality of the initial scanning, not the film recording. It is imperative, therefore, that digital imaging requirements be established and used to capture fully the significant information contained in the source documents, and that a rigorous scanning quality control process be instituted, with visual inspection occurring both on-screen and via printouts from the digital images.

- In reviewing the findings on image quality and COM inspection from this project, the authors recommend the following guidelines be followed in the creation and inspection of computer output microfilm:

- Permanence requirements: film stock, COM processing, associated packaging, and storage conditions should all meet ANSI/AIIM and RLG standards.

- Resolution and pictorial quality: a minimum resolution of 600 dpi with 1-bit scanning should be used to create digital images for brittle books and journals consisting of monochrome text and line art. Halftones capture will require the use of appropriate enhancement capabilities. The COM recording system should be able to output the 600 dpi 1-bit files onto film in a manner that results in no loss of resolution or (apparent) tonal range. Both the digital images and the COM should undergo technical and visual inspection. On-screen and paper printouts can be used to judge the quality of the digital images, and 100% inspection of the image files is recommended. An RIT Alphanumeric Test Object should be scanned at the same time as the brittle books. Advice from imaging scientists and vendors should be sought on the inclusion of a grayscale target and whether targets should be scanned at the highest possible quality that can be achieved by the scanner and/or scanned at the same exposure setting used for the brittle text.

  The achieved resolution on film should be evaluated by comparing the on-screen readings of the RIT target to the readings taken from the COM. Detail capture should be confirmed by examining the smallest significant lower case letter contained in a document as recorded in the digital image and on the COM. The appearance of halftones and fine line drawings should also be evaluated for detail capture and the introduction of moire and other evidence of aliasing. The COM should be inspected over a light box using a 100x microscope.[21] Once satisfied with the quality of the product, a 10% sampling of COM for resolution verification is recommended if the digital files have been 100% inspected. (Early in the project, Cornell detected erratic "dropouts" of lines of data on film. These were later traced to a faulty raster generator board.) After the hardware was replaced in the COM recorder, the problem disappeared. The COM service provider should not be required to take any resolution readings.

[21] A microscope with magnification between 100x and 200x is recommended in Lisa L. Fox, ed., *Preservation Microfilming. A Guide for Librarians & Archivists* (Chicago, IL, American Library Association, 1996), 213.

- Polarity: the COM should be produced in negative polarity. The master negative COM should be properly housed and stored. In the future event of either a digital disaster or a request for a film copy from another institution, a duplicate negative could be printed from the master using a conventional film duplication process.

- Density: Given that all density readings were highly consistent and fell within acceptable range, we recommend that fewer maximum density readings be required for COM than conventional microfilm. RLG guidelines specify 3 maximum density readings per title or 2 readings for volumes with fewer than 50 pages, and a minimum of 8 readings per reel. We recommend that COM service providers take three Dmax readings per reel and one Dmin reading per reel. The home institution should take one reading per title. Over time, this requirement could be even further reduced. Density variation should be consistent with requirements for creating new microfilm (see previous section).

- Image Placement: images should be recorded in the cine position, either one image/frame (IA) or 2 images/frame (IIA). The images should be centered on the film, with a consistent distance between frames.

- Reduction Ratio**:** Use of variable (and non-standard) reduction ratios is acceptable, provided that information regarding resolution, bit-depth, resulting pixel dimensions, and recording space on film (e.g., 15mm) are included on a technical target. If a standard reduction ratio is used, that ratio must be conveyed on a technical target, according to RLG guidelines.

- Film Size: the exclusive use of 35mm microfilm for preservation purposes should be reexamined. More commercial options for high resolution COM recording (and film scanning) exist with 16mm and 105 mm formats than 35mm film.

- Bibliographic Integrity: there should be 100% inspection for bibliographic integrity conducted either at the time of scanning or after COM recording. If full bibliographic inspection occurs on the digital images and accompanying metadata, a 10% inspection of the COM should also be conducted. Delaying bibliographic inspection until reviewing the COM can eliminate one inspection stage, but may actually increase the time spent in inspection and processing if many errors are detected on the COM.

- Technical Targets: targets containing information on the scanning process used (e.g., resolution, bit depth, use of enhancements, file formats, type and level of compression) should be created, as well as those conveying essential document characteristics, such as physical page dimensions of the original (including all variations from that size, including foldouts, reduced photocopy versions of oversized items), and level of detail and illustration content. Include as a target either the collation form or preferably the actual tables containing the document control information to aid in recreating pagination and indexing if the COM needs to be scanned to recreate digital files.[22] (See below, Research Issue 4, Development of Metadata Elements.)

[22] Appendix I of the COM final report contains copies of the forms and target sequence used in the Cornell Project.

B. Issues Affecting the Cost of Computer Output Microfilm

Cornell undertook a more modest cost study than Yale, collecting data in the following categories: preparation, scanning, file management, tape creation, and COM inspection. These categories roughly correspond to the categories used in the Yale cost study. For comparison purposes, Cornell calculated "Yale-adjusted" salaries and mean times to reflect the difference in the average size of books scanned at Yale (216 pages) and Cornell (341 pages). We provide comparative cost figures in the next section.

Book Characteristics

As the Cornell and Yale projects found, book characteristics such as the presence of halftones, complex illustrations, darkened paper or faded inks, and similar factors associated with deterioration or heavy use can increase the costs of bitonal digital conversion from either the original book or its microfilm version. In the Cornell project, a book containing low contrast pages required additional set-up time to ensure that the threshold setting was not going to lead to feature drop-out or character fill-in. Books that exhibited inconsistent density between pages resulted in higher inspection costs, as the number of pages that had to be rescanned increased. The presence of halftones had the greatest impact on capture costs, and involved a separate form of scanning. The first form of scanning was done in an "auto-mode" in which standard settings were used to capture all pages of the volume. The second form of scanning, "manual mode," involved windowing halftone information on a page, and treating it differently than the surrounding text. The time taken to scan in "manual mode" was considerably longer than in "auto-mode" (running an order of five times longer per page). Fortunately, not all pages of a book contain halftone information, and the per page cost differential spread across the entire book represented an additional $0.02/page. The use of "manual mode" increased the scanning time per Cornell book by 40 minutes (from 86 minutes to 126 minutes); if book length is adjusted to the Yale average of 216 pages, the time increase was only 17 minutes (from 56 minutes to 73 minutes). Nonetheless, Cornell bibliographers decided that only books containing halftones that were considered significant to the meaning of the text would receive "manual mode" treatment, and scanning staff relied on curatorial review of illustrated materials to determine which mode to use.[23]

The need to disbind the book and trim the binder' s margin for scanning on the Xerox flatbed scanners (XDOD) increased preparation times considerably. On average this took nearly 20 minutes per volume, representing an additional $.023 per page cost. On the other hand, if brittle books cannot be disbound for scanning, either the costs of digital capture will be higher or the quality of the resulting images will be lower, given the current state of scanning technology. (A discussion of bound volume scanning is presented in the next section.)

Programming Characteristics

[23] See Appendix I of the COM report for "Guidelines for Autosegmentation/Manual Windowing."

A great deal of time was spent at the beginning of the project to develop systems programming capabilities for handling, rotating, and moving the image files and relevant targets (some image, some text-based, some created on-the-fly). Creating microfilm directory structures and tape generation scripts (to automate the copying of files onto 8mm tape to send to Image Graphics) and log files for quality control also required considerable programming time.[24] Additionally, Cornell, with the support of the Xerox Corporation, developed an "export tool" to convert the XDOD-created RDO files into files that could be directly readable by the UNIX tools used to generate the tapes. Costs associated with development and ramp-up were not recorded, but on average file management and tape creation activities in the production phase of the project increased costs by slightly over a penny an image. However, the programs developed at Cornell may not be transferable to other institutions or to other companies besides Image Graphics. To ensure that systems development costs are kept low, reel programming requirements must be standardized and microfilm reel generation scripts developed that are platform and equipment independent.

Film Characteristics

When Cornell went out to bid for its Digital to Microfilm Conversion Project, only one vendor, Image Graphics, was able to meet its exacting needs. A number of vendors could meet all other requirements, excepting the need to produce film on the 35mm format. Most companies produce COM on 16mm film and 105 mm fiche. Some companies are able to record onto 35mm film, but can not handle the 600 dpi image files or the small reduction ratios. The preservation community should reevaluate the exclusive use of 35mm microfilm for preservation purposes, especially if digital image files are to serve as the access masters.[25]

Other Considerations

Additional cost savings will also certainly be realized if film inspection procedures are streamlined in the manner suggested in the previous section. Recommendations associated with metadata will be discussed below. Finally, it appears that there are cost savings accrued by combining the digitization and COM recording processes into one effort. If digital files are not to be output to COM directly after scanning, some additional steps may be required, thus increasing costs. In the final section of this paper, we will discuss the pros and cons of deferring the production of COM to a later time.

---

[24] See Appendix II of the COM final report for information on reel programming.

[25] See survey of COM recording companies in the April 15, 1997 issue of *RLG DigiNews*. On the other hand, the working group of the German Research Association strongly endorses the use of 35mm microfilm as the starting point for digitization: "Its image size guarantees sufficient quality, even with problematic material, up to a size of 60 x 80 cm." Weber and Dorr, 5. A second German report on digitization acknowledges that good results can be obtained from 16mm film, but predicts that 35 mm film digitization will become more heavily used in the next few years. See *Retrospective Digitization of Library Collections for a Distributed Digital Research Library*, 45.

RESEARCH ISSUE 3:

THE CHOICE OF A DIGITAL CONVERSION PATH (FILM-FIRST OR SCAN-FIRST)

In this section we will examine various paths in the process of creating both digital images for access and microfilm for preservation. The primary question is: What are the circumstances governing the decision to scan-first versus film-first?

Table 2 below suggests possible hybrid workflows. It describes some of the circumstances that may lead to a film-first or scan-first decision. The sequence of steps may be coupled in a single workflow (as in the Cornell project), or they may be separated by several years (as in the Yale project). In some cases, the choice of how to begin will be technical. For instance, if both books and microfilm exist, but the brittle paper has deteriorated to such an advanced state that it can no longer be handled, microfilm is the only viable source for scanning (so the project would begin at the second step of the film-first option). In other cases, the circumstances are resource or policy related: funding is available only to create a single format (whether microfilm or digital images) or institutional policies regarding disposition and handling preclude some reformatting options, such as flatbed scanning.

TABLE 2.

Potential Hybrid Work Flows		
First step	**Second step**	**Circumstances (not a complete list)**
film book	scan film	• desire to handle originals once, disbinding not an option, scan for access
scan book	output images to preservation COM	• desire to handle originals once, book can be disbound, scan for preservation
film book or scan book	scan book or film book	• preservation quality not achieved in scanning, infrastructure provides options to save costs

Caveats and premises about quality, technology, workflow, and cost

From the managerial perspective, the best approach to reformatting brittle material is the one that meets objectives for preservation (film) and access (digital images) at the lowest cost. Until we have full confidence in digital archiving, "permanent" continues to mean analog, so it is appropriate to compare the quality and costs of preservation microfilm to digital COM to determine whether the film-first or scan-first approach yields any advantages. Findings from the Cornell project establish that digital COM can be of equal or superior quality to traditional 35mm preservation microfilm for costs that are slightly under $0.12 per page-image. The Cornell and Yale reports underscore a number of caveats about this and other costs reported in their projects, such as the fact that the $0.12 per image for COM refers only to one generation of film. These costs also presume that bibliographic targets have already been created and are stored with the digital images.[26]

[26] See COM final report, "Quality Finding No. 1" for discussion of film quality, p. 7-9, and p. 30 for costs associated with creating COM, which averaged 11.6 cents per page. [Online]. Available: http://www.library.cornell.edu/preservation/com/comfin.html. See also, Bericht der Arbeitsgruppe Technik zur Vorbereitung des Programms, Deutsche Forschungsgemeinschaft (German Research Council) (DFG), *Retrospektive Digitalisierung von Bibliotheksbeständen für eine Verteilte Digitale Forschungsbibliothek* [*Retrospective*

Based upon these quality and cost findings for film, we may reach two preliminary conclusions about the preservation component of the hybrid approach. Film-first and scan-first offer comparable microfilm quality, but COM production currently appears to be less expensive than microfilm production.

With respect to the the digital images, preliminary conclusions from the Yale and Cornell projects are that scanning from paper and scanning from film offer comparable cost, but the quality of scan-first digital image is superior. The cost comparison tables for the two projects report that for 600 dpi 1-bit images production scanning falls into the range of $0.22 to $0.26 per image for paper scanning, and $0.24 to $0.28 for film scanning.[27] The conclusion about quality is based upon two standards: system resolution and best representation of the original. System resolution is a shorthand way of referring to the phenomenon that today' s microfilm scanners cannot achieve the same legibility on a technical target (such as the RIT Alphanumeric Test Object) as a flatbed scanner at the same dpi and bit depth. The quality standard of "best representation" of the original needs a bit more explanation.

As noted above, the authors of this working paper agreed to distinguish "preservation quality" from "access quality" when describing the digital masters produced in the scan-first and film-first approaches. Referring only to issues of pictorial quality—metadata attributes are characterized in Research Issue 4 below—these quality differences are summarized as follows:

- *digital preservation masters* can serve to create replacements via output to COM for the original brittle book. These files can also be used to recreate a printed counterpart that matches the original page as closely as possible in height x width dimensions, fidelity to detail (including serifs, stroke widths, and smoothness of edges) of text and simple line art, image orientation, and skew. As noted in the Cornell project, the creation of digital preservation masters required a bitonal scanning resolution of 600 dpi (QI of 8, high quality), the disbinding of books for flatbed scanning, and the use of image enhancement algorithms to represent some of the (apparent) tonal range of halftones and other photomechanical processes.

- *digital access masters* can serve as high-quality surrogates for the original brittle book. These files are created to support the widest range of potential uses (short- and long-term), including: on-screen study, OCR processing to generate full-text for searching and mark up, and high-resolution printing. Although these images may be highly functional, objective measurements (such as physical page dimensions, presence of skew) and subjective measures (such as the Quality Index or visual examination of book illustrations) would indicate that these images fall short of the more precise fidelities to the original that are specified for preservation. In addition, lower resolution will increase the risk of feature drop-out due to improper thresholding or information loss in subsequent image processing (e.g., OCRing, compression, derivative creation). If a QI of 5 (medium quality) were used as a benchmark, then a resolution of 385 dpi would be needed, which led the working group of the German Research Association to recommend film scanning resolutions between 350 and 400

Digitization of Library Collections for a Distributed Digital Research Library], 1997. Appendix 4 presents a number of tables that summarize costs and processes associated with book and microfilm scanning. COM costs in the DFG Report refer to 16mm film, the use of a laser COM recorder, and presumably 400 dpi resolution. [Online]. Available: http://www.SUB.Uni-Goettingen.de/GDZ/vdf/entwurf3.htm

[27] COM final report, see "Table 3. Producing Digital Images from Paper vs. Microfilm."

dpi.[28] We recommend that a scanning resolution of 400 dpi be used whenever possible; for oversized items (and reduction ratios over 12 x), a dpi of 300 may be all that is currently affordable. Additional tests to evaluate the quality and utility of 300 vs. 400 dpi image files to serve the full range of functional uses for access are needed.[29]

The following caveats are offered to reiterate the capabilities of scanning technology and their associated requirements for document handling during the period of the Yale and Cornell projects, which, as yet, have not been superseded:

- given the characteristics of "the brittle book" as well as the traditional standards for image quality in preservation microfilming, 600 dpi 1-bit scanning represents the acceptable minimum specification to achieve full information capture without item review of the original volumes
- neither overhead scanners nor digital cameras have demonstrated the capability to achieve quality comparable to 600 dpi 1-bit flatbed scans *in a cost-effective manner*, so the original books must be disbound and pages trimmed in the scan-first approach[30]
- even with high-quality film, microfilm scanners may not achieve the quality of direct-from-paper flatbed scans. In a direct comparison of paper versus film scanning, the Cornell and Yale projects showed that 600 dpi bitonal digital images were superior in quality when created directly from paper rather than from microfilm versions. The most obvious difference in quality was seen in the reproduction of halftones. Current bitonal film scanners do not offer the same enhancement capabilities as flatbed scanners for treating halftone information.[31]

It will be important to revisit the question of managing the hybrid approach as technology and our assumptions about image quality for digital images and microfilm evolve. In the meantime, we have

---

[28] Weber and Dorr, *Digitization as a Method of Preservation: Final Report of a Working Group of the Deutsche Forschungsgemeinschaft, 11.*

[29] A number of film scanning projects have chosen to scan at 400 dpi, including those of the Library of Congress, RLG's Studies in Scarlett (NC State, NYPL), Cornell's SagaNet Project, the Australian Cooperative Digitization Project and the Burney Collection at the British Library. The Early Canadiana OnLine Project is scanning from fiche in the 300-600 dpi range, depending on the reduction ratio. For a review of some film scanning projects, see the August 15, 1997 issue of *RLG DigiNews*, which is devoted to film scanning and COM recording issues.

[30] Digital cameras that meet or exceed the quality of 1-bit flatbed scanning are widely available, but only when used to produce 8-bit or 24-bit images. See, for example, the report on scanning 18th-century rare books, "Producing Digital Images," *The Electronic Archive of Early American Fiction*, University of Virginia Library, July 1998. Available: http://www.lib.virginia.edu/speccol/mellon/image.html. The authors of this Working Paper agree that until face-up scanning is comparable in quality and cost to 1-bit flatbed scanning, the scan-first hybrid approach requires using a high-quality flatbed scanner and disbinding the originals.

[31] See details on the quality comparison in Kenney, *Digital to Microfilm Conversion: A Demonstration Project*, 11-14. This finding was also reached by Yale: "Bitonal scanning is not appropriate for preservation microfilm containing materials with rich tonal qualities, such as photographs, halftones, and dense line art, even if the microfilm containing these types of illustrations is of high quality," See Conway, *Conversion of Microfilm to Digital Imagery*, 10.

created a decision tree that follows from the caveats and assumptions described above. Deciding where to begin the hybrid project requires a consideration of issues associated with the source materials, with assumed capabilities of technology and cost, and with local policies regarding disposition of originals. Each is important, but we have taken as our starting point the question, "What is your quality objective for the digital masters?"

Finally, in considering not only how to begin, but also how to manage a hybrid reformatting project, it is important to distinguish between one-time and two-time approaches. For brittle collections that have never been reformatted, one could create digital images and microfilm (or COM) in a single workflow, or in two different projects separated by time. The impact of workflow on cost needs to be more fully explored.

Hybrid Approach Decision Tree

I. **Goal is to produce digital preservation masters and preservation quality film**

A. When only brittle volumes are available:

Assess brittle volumes. (Contents must be complete.) Will disposition policies permit disbinding?

Yes) Disbind and scan first at 600 dpi 1-bit in a manner to expedite COM production.

No) Assess the bindings, structure (sewing), and inner margins. Without alteration, can volumes be fully opened (180?) with each page flush to the platen on a flatbed scanner?

Yes) Scan first.

No) Consider preparation and disposition costs related to alteration. Can the sewing threads be cut in order to facilitate flatbed scanning of fully open volumes, with each page flush to the platen?

Yes) Cut threads and scan first.

No) Film first. **Note:** if film-first is determined to be the preferred approach, assess book contents. Can 600 dpi be achieved on the source document blown back to its original dimension? And can information loss from "complex illustrations" be accepted?

Yes) Digital preservation masters might be achieved by scanning the film.

No) Assume that digital preservation masters cannot be created in a film-first approach. You must decide whether handling and disposition policy will be changed from "keep intact" to "allow for modification." If so, return to **I.A**. If quality objective for digital images can be changed from preservation masters to access masters, proceed to **II**.

B. When both brittle volumes and microfilm are available:

Assess microfilm. Does the second-generation negative meet relevant standards for preservation quality and permanence (ANSI/AIIM/RLG)?

Yes) Scan the film only if (a) 600 dpi can be achieved on the source document blown back to its original dimension, and (b) information loss from "complex illustrations" is acceptable. If not, presume that digital preservation masters cannot be created cost effectively in a film-first approach. If access masters are acceptable, proceed to **II**.

No) Assess brittle volumes. (Contents must be complete.) Will disposition policies permit volumes to be scanned at 600 dpi 1-bit on a flatbed scanner? (see **I.A**)

Yes) Scan first and produce preservation COM.

No) Film first (i.e., refilm the brittle volumes to create preservation quality microfilm). Digital preservation masters may be created by scanning the film (see above), or the quality objective for the digital images may need to be adjusted from preservation masters to access masters.

Note: under certain circumstances, we presume that digital preservation masters can be created from either the originals or preservation microfilm. If both options are available, conduct a cost-benefit analysis with a representative sample of materials, if necessary, to determine whether the preferred approach is to scan the film or the original volumes.

C. When only microfilm is available:

Assess microfilm. Does the second-generation negative meet relevant standards for preservation quality and permanence (ANSI/AIIM, RLG)?

Yes) Scan the film. Further testing must be conducted to determine whether digital preservation masters can be created in 1-bit microfilm scanning. If you have high-quality microfilm, determine whether the quality produced by the highest resolution offered by the microfilm scanner (e.g., 600 dpi) satisfies the requirement for digital preservation masters as described on p. 23. If so, scan the film at the highest possible 1-bit resolution. If not, consider the more expensive option of grayscale scanning, or conclude that digital preservation masters cannot be created cost effectively and proceed to **II**.

No) Presume that digital preservation masters cannot be created cost effectively and recognize that you have not met preservation requirements for quality or permanence in the film. If digital access masters desired, proceed to **II**.

II. Goal is to produce digital access masters and preservation quality film

A. When only brittle volumes are available:

Is disbinding permitted?

Yes) Disbind and scan first at 600 dpi 1-bit and output to COM.

No) Film first and scan film at a minimum of 400 dpi 1-bit, or if more cost effective, film first and scan the bound volumes with an overhead scanner.

B. When both brittle volumes and microfilm are available:

Assess microfilm. Does the second-generation negative meet relevant standards for preservation quality and permanence (ANSI/AIIM, RLG)?

Yes) Scan the film at a minimum of 400 dpi 1-bit.

No) Either refilm the originals or scan the originals at 600 dpi 1-bit and output to COM . Extant film *may* produce digital access masters but preservation film requirements for quality and longevity not met.

C. When only microfilm is available:

Assess microfilm. Does the second-generation negative meet relevant standards for preservation quality and permanence (ANSI/AIIM, RLG)?

Yes) Scan the film at a minimum of 400 dpi 1-bit.

No) Presume hybrid approach not viable as preservation standards for quality and/or permanence have not been met.

The Hybrid Approach Decision Tree offers a means for assessing some of the circumstances governing whether to scan first or film first. Additional information is needed in the context of the national brittle books program to make definitive recommendations**.** In order to get those answers, the authors suggest that the National Endowment for the Humanities convene a meeting to discuss selection criteria for hybrid reformatting that could form a basis for appropriate policy governing the conduct of such projects. This discussion should address several questions related to the national brittle books program. First, should preservation master quality be a requirement for both the film and the digital masters in hybrid projects? Second, if it can be established conclusively that materials must be disbound or must lie flat to create bitonal digital preservation masters, will funding agencies support the preparation and/or disposition activities necessary to meet this objective? And third, in the scan-first approach, must COM be produced at the time the digital product is created? Although the authors argue that preservation needs are not met until a film version has been created, there may be times in which risk management suggests that an institution postpone the creation of the COM. For instance, we can envision a circumstance in which an institution presents satisfactory evidence that it can responsibly manage its digital image files and also agrees in writing to output the files to COM if circumstances change.

Additional questions should be addressed to microfilm scanning manufacturers, other industry experts, service bureau representatives, and project managers with experience in film scanning projects. Why is film scanning currently more expensive than scanning disbound paper? Given the high throughput of microfilm scanners, shouldn't access digital masters be created at a much lower cost from microfilm than from paper? What specific changes need to occur in scanning technology, microfilm creation, or the procedures associated with metadata creation (including file naming) to meet the goal of reducing film scanning costs?

RESEARCH ISSUE 4:

THE DEVELOPMENT OF METADATA ELEMENTS ASSOCIATED WITH THE DIGITAL IMAGE PRODUCT

In this section, we will examine requirements for metadata to accompany the digital image files in order to create a usable digital object. The question to be answered is, "What are the fundamental metadata elements required by the hybrid approach?"

Context

In his keynote address at the conference *Managing Metadata for the Digital Library* in May 1998, Clifford Lynch observed that it is fallacious to talk about different types of metadata, as if "data" were always clearly recognizable and information about that data was distinct. The meaning of metadata, he observed, is extremely contextual, where the boundaries can become diffuse, if not endless. He suggested that we picture metadata as "a cloud around an information object that diffuses indefinitely."[32]

[32] Clifford Lynch, "Metadata in Context, What We Know and What We Don't Know," keynote address at *Managing Metadata for the Digital Library: Crosswalks or Chaos?*, May 4-5, 1998.

To put metadata into the context of the hybrid approach, we have chosen to view preservation microfilm and the scanned images as "the data," and all collateral information related to these objects as metadata. For the purposes of this discussion, we will use the aggregate term "digital object" to refer to all of the electronic files associated with an original brittle title (in the case of most monographs) or volume (in the case of most journals). A digital object will consist of:

1. digital masters (scanned page-images; each with a unique file name)
2. associated administrative metadata (described below), and
3. associated structural metadata (described below).

When the digital objects have been saved in their appropriately named subdirectories within a digital repository, the workflow is considered to be complete. Thus, the digital repository, or database, creates the *potential* for enhanced access. Delivery of the digital objects (to the screen or a printer) will depend upon other technical capabilities (e.g., use of internet browsers, image viewers) and the contribution of the owning institution. Some institutions will have demanding audiences, well-developed infrastructures, and sophisticated interfaces for their digital collections; others will have more modest capabilities. Our objective is to create digital masters that, in the words of George Farr of the National Endowment for the Humanities, "close no doors."

This research question focuses exclusively on metadata elements related to the digital books and journals because we endorse the practices already in place to ensure physical and bibliographic control for microfilm. These meet the function of ensuring that the film can be identified and distributed easily, and that a given brittle book will not be microfilmed more than once. As noted in the *RLG Preservation Microfilming Handbook*, "Appropriate bibliographic control for titles preserved on microfilm consists of a bibliographic record created according to established national standards and made widely available in the national databases."[33]

To this point, our discussion of digital masters has primarily focused on the relationships among source material, scanning technology, digital image quality, and cost. Metadata elements should be viewed in the same context: the attributes of the source material (complexity of pagination and internal organization), our managerial and functional objectives for the digital object (quality), and the capabilities of technology (to automate or semi-automate metadata) to determine total cost. Depending upon the extent of metadata specified for a book or journal—*and even when excluding OCR or mark-up*—these costs can be significant. Paul Conway has noted that indexing "represents almost 40 percent of the labor invested in Project Open Book."[34]

Purposes

The first purpose of creating metadata to accompany digital images is to promote digital resource management (including preservation), discovery, and use. To fulfill the promise of digital technology to

[33] Elkington, *RLG Preservation Microfilming Handbook*, 1.

[34] Paul Conway, *Conversion of Microfilm to Digital Imagery: A Demonstration Project, Performance Report on the Production Conversion Phase of Project Open Book* (New Haven, CT: Yale University Library), August 1996, p. 15.

enhance access to research materials (especially as compared to the linear organization of microfilm), digital images must facilitate at least two levels of on-line navigation:

- go to a specific page, and
- "open" a digital book or journal at a meaningful section (e.g., title page, table of contents, index).

Online navigation that transcends these two minimum levels by providing hierarchical access to the structural components of a book or journal may also be desirable.

Metadata also satisfy the requirements for physical and bibliographic control and enable the following:

- locate images in the digital repository
- provide easy ways to identify and obtain digital resources and their surrogates, and
- minimize the likelihood of duplicate digital imaging activities.

The adoption of a number of guidelines for metadata creation in the hybrid approach will help control project costs and regularize the functionality of digitized books and journals.

Types of Metadata

For the convenience of summarizing practice and making recommendations, we will classify metadata in two broad categories: administrative and structural. The former refers to the descriptive elements that reside within or outside a digital object to ensure that it will be managed over time; the latter refers to the elements within a digital object that facilitate navigation.[35]

Administrative Metadata

Examples of administrative metadata elements are found in the "Production Notes" that have been produced for every title scanned in Cornell's projects, including the COM project, since 1990. Comparing early production notes to more recent ones, we observe that practice has changed slightly over the years, but a number of elements have been used consistently:

[35] The classifications used here are consistent with those used by the Digital Library Federation. DLF also classes descriptive information as "intellectual metadata," but this third broad category is not addressed in this working paper. Our primary objective is to raise questions about the functionality of digital objects rather than the way they might be described or pointed to in catalog records. See Donald Waters, "I know it's out there but where?" Problems and prospects of discovery and retrieval in digital libraries, presentation at *Managing Metadata for the Digital Library: Crosswalks or Chaos?*, May 4-5, 1998. [Online]. Available: http://www.clir.org/diglib/dlfpresent.htm. Fuller explanations of administrative and structural metadata, particularly as they relate to nineteenth century materials are provided in, *The Making of America II Testbed Project White Paper*, Version 2.0 (September 15, 1998) [Online]. Available: http://sunsite.Berkeley.EDU/moa2/.

Production Notes for the Math Book Collection (1990-92) and the COM project (1994-96) [36] included ten administrative metadata elements:

1. owner/creator (Cornell University Library)
2. note regarding quality (to replace the original)
3. note regarding source material (irreparably deteriorated original)
4. type of scanner used (Xerox software and equipment)
5. scanning resolution (600 dpi)
6. compression (CCITT Group 4)
7. note regarding output from digital images (paper meets ANSI standard for permanence)
8. funder(s) (CPA and Xerox)
9. copyright (Cornell University Library)
10. date (1992).

Production Notes for the Making of America Project titles (1995-96) are slightly modified: the references to type of scanner and paper output have disappeared, and several elements have been revised or added: the statements of quality and source now read "to preserve the informational content of the deteriorated original;" there is an additional note describing thesource ("best available copy has been used"); another for bit depth (bitonally); and another for project name—for a total of eleven elements.

These production notes illustrate that administrative metadata are recorded for the managers of the digital images rather than the users. The creation date of the digital object and the compression scheme (format), for example, are two of the critical elements needed to schedule migration of files. The file format and version (for example, TIFF 5.0) are also important for management and migration; in the Cornell and Yale projects, these metadata are recorded in the file name extensions and the file headers respectively.

Targets

The Cornell project incorporated a technical target in scanning as part of quality control. As is the practice with preservation microfilm, these targets are used to determine whether a scanner performs consistently at its optimal levels. In the Cornell project, 600 dpi images of these targets were scanned with the brittle books, then included with the master images for each volume. Saving these targets as documentation of system quality serves two functions: the targets document the upper limits of quality (detail reproduction in 1-bit systems) of the scanner that was used; and they help programmers ensure that information loss will be minimal (to none) when creating derivatives or migrating master files to new formats.

For digital preservation masters, a bibliographic target is also required. It is saved to facilitate output of the digital object to COM and to satisfy preservation requirements for bibliographic control. There are

[36] To view the pre-MOA Production Notes, bring up "Image 1" of any of the 571 titles in the *Cornell University Library Math Book Collection*. Available: http://moa.cit.cornell.edu/dienst-data/cdl-math-browse.html; the appendix to COM Final Report includes an image of the Digital-to-COM Production note, at: http://www.library.cornell.edu/preservation/com/Appgifs/app32.htm.

both title- and reel-specific targets that must accompany digital objects delivered to a service bureau for COM production. The latter can exist in their own project directory for production purposes, but the former should be maintained with the digital object so they can be easily organized for output to film, and the likelihood of errors can be greatly reduced.[37]

We propose a list of required administrative metadata elements (see Table 3 below) to document the following attributes of a given digital object:

- bibliographic and technical data associated with the conversion from an analog original to digital imagery—what Ann Swartzell of Harvard refers to as a "digital colophon;"
- management data needed to manage and migrate digital files to ensure continuing access to digital access and digital preservation masters; and
- reel programming data needed to organize digital preservation masters on COM.

The broader community of practitioners and industry experts should participate in discussions of regularizing structure (i.e., where to record these metadata elements), syntax, and workflow (i.e., noting how many of these elements we can automatically generate) for administrative metadata in the hybrid approach. The following table provides a starting point for discussion of which metadata elements should be required.

[37] See the following examples in the COM Final Report: "Sample Bibliographic Record Target," at http://www.library.cornell.edu/preservation/com/Appgifs/app28.htm; and "Target and Image Arrangements for Future Reels," at http://www.library.cornell.edu/preservation/com/Appgifs/app28.htm.

TABLE 3.

Proposed Administrative Metadata Elements
1. a technical target that documents the capabilities of the scanner that was used • for bitonal scanning, the RIT Alphanumeric Test Object is recommended
2. for digital preservation masters, bibliographic targets for COM output
3. name of project
4. name of funding agency(ies)
5. unique identifier for the object
6. designation of object as "digital preservation master" or "digital access master" • *must* be recorded in bibliographic record, according to procedures routinely followed to designate ownership and location of microfilm master negatives[38]
7. owning institution
8. copyright statement (including note of any use restrictions)
9. date object was created (i.e., scanning date)
10. scanning resolution, bit depth, file format and version, and compression
11. change history of object: current version (edition) of object, with dates of migration, and notation of which features in #10 were changed

Structural Metadata

The creation of structural metadata is central to the digitization of nineteenth century materials. The authors of this paper agree that the minimum elements associated with digital masters created from the brittle book should be pagination and "feature codes." In other words, for each digital image that has been stored in an image database, there must be a related field that indicates whether it has a page number, and another to identify an associated feature (e.g., blank, none, title page, table of contents, index). In this scheme, most images have a page number, and few images have an associated feature. We defer to the broader community to decide what the features should be and to discuss whether authority control should be used for feature names.

The structural metadata elements of pagination and "features" organize a sequence of images in a way that they can be retrieved and used more flexibly than simple linear access (page-forward, page-back). Today we agree that it is essential to provide the capability to go to a specific page and to move easily from one part of a book to another. As we obtain a greater understanding of user behaviors in and expectations of the electronic environment, the feature list will likely evolve. This is all the more true for journals, where it is necessary to generate a hierarchical structure to facilitate browsing at the top level of a volume or title.

How does one embed this functionality in a series of digital images? We will address this broad question by examining the following sub-topics:

- structure (what to encode and where to record this information)
- syntax (the names for these elements and the authorities we use to control language)
- workflow (the possibilities of creating this metadata at various points in the scanning process).

[38] MARBI recently approved a new MARC 007 field for digital preservation/reformatting. See FAQ response by Diane Hillman in the February 15, 1999 issue of *RLG DigiNews* at http://www.rlg.org/preserv/diginews/.

Although we presume the need for navigation on screen, we also view printing to be among the fundamental access needs. One of the important issues to resolve in generalizing the hybrid approach is whether or not blank pages *must* be included so as to provide the correct representation of rectos and versos when two images are displayed side-by-side on screen, and perhaps more importantly, to be able to recreate the codex if entire books are to be reprinted.

Structure

Reports from Yale and Cornell describe the benefits and limitations of having used software from the Xerox Corporation (XDOD) to structure digital masters. The principal advantage of this software is the ease in associating page numbers with image numbers, and creating internal hierarchies among the image files.[39] The main limitation of the proprietary RDO (Raster Document Object) file format is that it is optimized for sending images to a Xerox printer. Both Yale and Cornell concluded that the RDO is not compatible with the digital library architecture they will use to manage digital collections. Important research and product developments within these projects centered upon the need to transfer the structural metadata from one database (the RDO) to another (the repository).

It is important to recognize that digital masters must not only be created in "widely supported formats" to ensure longevity, but that they must also conform to the database architecture of the digital repository in which they will be stored. Without this compatibility, the delivery of digital books and journals in the networked research environment is highly complex or impossible to accomplish. It is one thing to create masters that can be sent to a printer, quite another to provide a capability for internal navigation in an online mode. We believe that masters created in the hybrid approach must have the potential to be output to the screen, to print, and, in the case of digital preservation masters, to COM.

A number of architectures have been used to structure "page-image" digital books (as opposed to full text), and discussion among a broad community is required to determine whether one model, with common rules for structure and syntax, will emerge.[40] Table 4 lists what we believe to be the minimum set of mandatory structural metadata elements.

TABLE 4.

Proposed Structural Metadata Elements

[39]For examples of the Xerox interface to create structural metadata, see "Appendix 5, Index Samples," in Conway, *Conversion of Microfilm to Digital Imagery*.

[40] These issues have been discussed in some detail at a recent conference regarding SGML and TEI. We fully support the proposal offered by a working group at this conference for the DLF to convene a group to draft a list of common structural and administrative metatdata elements for digital books. See Catheriene Tousignant, "Structural and Administrative Metadata in Page-Image Conversion Projects: Discussion Summary and Recommendations." TEI and XML in Digital Libraries Conference, June 30-July 1, 1998, Washington, D. C. [Online] Available: http://www.hti.umch.edu/misc/ssp/workshops/teigrp3.html. For an earlier model, see W. Turner, Network Working Group, *Request for Comments: 1691: The Document Architecture for the Cornell Digital Library*, August 1994. Available: http://www.netbook.cs.purdue.edu/othrpags/rfcs/rfc1691.txt. The structural metadata associated with the digital files in the Cornell Digital-to-COM project were "liberated" from Xerox' s .rdo format and mapped to the Cornell Digital Library architecture with custom software co-developed by Cornell and Xerox.

1.	correct page number associated with each digital image • except in cases of printer's errors, page number must be transcribed (e.g., Roman or Arabic, upper or lower case) exactly as they are printed
2.	internal navigation/structural points (syntax TBD; see pp.30-31 above), sometimes referred to as features or feature codes, when present in the original. • *for books*, minimum elements: blank, title page, table of contents, index • *for journals*, minimum elements: blank, title page or cover, table of contents, index at the issue level when present; at the volume level when not

Syntax

At first glance, the syntax for pagination appears to be relatively straightforward. Both Yale and Cornell, for example, transcribed page numbers exactly as printed. Handling unpaginated material, such as illustrations and foldouts, front and back matter, was a more complex matter. Along with the issue of blank pages, questions about rules for pagination should be addressed in the broader discussions of the hybrid approach.

There are clear cost implications for specifying how many features must be encoded. If the hybrid approach is to be generalized, this question deserves broader discussion, where a final specification must balance cost and functionality. One comparison serves to illustrate the range of practice. Yale assigned "typically upwards of 25" feature codes to titles in Project Open Book; Cornell chose to tag 5-10 features for comparable monographs.[41]

A number of institutions have developed dictionaries for structural elements for digital projects involving books and journals.[42] Differences in terminology among the institutions is readily apparent, but it is also worth noting that most share the philosophy of generalizing, rather than transcribing, the parts of a book or journal. For example, some institutions use the term "table of contents," while others simply use "contents." What is important is that each is consistent in applying the general terminology across

[41] One of the assertions in *Project Open Book*—a research as well as production project, after all—"as yet untested" is that value increases in some proportion to the amount of structural metadata. Paul Conway speculates " . . . that the cost of creating a high-quality, structured index for a complex digital image file is recouped through more efficient navigation of the file and more accurate and successful retrieval of needed information by the system's users." See, Conway, *Conversion of Microfilm to Digital Imagery*, p.12.

[42] Cornell and Yale have published their lists of terminology. The Elsevier journal specification was adopted, with varying degrees of effort, to structure 19th- and 20th-century journals and books in the following projects: TULIP, CORE, and Making of America (Cornell and the University of Michigan); see, Marthyn Borghuis, Hans Brinckman, et al., *TULIP Final Report* (Elsevier Science, 1996), see Section II 1. Available: http://www.elsevier.nl/homepage/about/resproj/trmenu.htm. The University of California at Berkeley developed a Document Type Definition (DTD) to encode similar materials in their Ebind project; see, University of California at Berkeley, *Digital Page Imaging and SGML: An Introduction to the Electronic Binding DTD (Ebind)*, 1996. Available: http://sunsite.berkeley.edu/Ebind/; for examples of Ebind's structural metadata, view any of the worksheets for the "Ebind-Encoded Documents," at http://sunsite.Berkeley.EDU/Ebind/samples. The National Digital Library Program at the Library of Congress uses a number of structures for their digital collections and they have made much of their documentation available for review; see, for example, "Attribute Use Examples, Structural Metadata Dictionary for LC Repository Digital Objects," July 1998, Available: http://lcweb.loc.gov:8081/ndlint/iwg/examples/att-use-ex-toc.html#top

collections of materials. Whether this terminology should be regularized or not is open to question, but authorities should be designated to ensure consistencies of practice, at least at the project level. Yale used the *Chicago Manual of Style* as the single authority in Project Open Book.

Workflow and Cost Issues

All digital conversion production processes boil down to three categories: manual tasks, tasks that can be fully automated, and those tasks, such as paginating digital files for "irregular" publications, that can be semi-automated. So far as we know, assigning feature codes, or "structuring" digitized books and journals, is entirely a manual process.

In terms of workflow, three models have been followed in hybrid and/or book scanning projects:

- to record the structure of a work on a workform prior to scanning; to record some of the structural elements during preparation, then others during scanning; to record features and/or pagination from the digital images after scanning.

Rather than summarize the pros and cons of each of these approaches, we believe that representatives from the hardware and software industry should be invited to consider the scope of the challenge and help us achieve the goal of being able to gather as much of this information in programmatic fashion. An obvious goal for any hybrid project would be to identify the commercial products that can be used today to automate some of the process of paginating and indexing digital images.

V. CONCLUSIONS AND NEXT STEPS

In order to generalize the results of the two studies undertaken on hybrid conversion and make them available to be put into production by other institutions, there remain several key issues to be resolved. They can, we believe, be decided only with the engagement of others: institutions that have done hybrid projects, imaging service providers, key industry and technology developers, funding agencies, and preservation and cultural institutions. In most cases, we believe that the additional information needed can be obtained by holding a series of meetings with representatives from the above stakeholders. These meetings could be held over the course of the next six months. At their conclusion, the working paper can be finalized and the key findings disseminated broadly to the preservation community both within the United States and around the world.

The issues that need further consideration include:

Decreasing the costs of converting microfilm to digital images by

- introducing modest changes to RLG microfilming guidelines to reduce skewing, lower reduction ratios, and revive the use of blipping
- assessing the potential impact of technologically oriented approaches to cost-reduction on specifications for the creation of preservation microfilm
- improving technology to reduce dramatically the times associated with scanning and indexing

- improving microfilm conversion by modifying two aspects of the processing

Improving the quality of the digital image product by
- making minor enhancements to existing international standards that govern the creation of microfilm, especially in the area of targeting
- advancing the capabilities and efficiency of scanning technology through the automatic calibration of scanners, continuous scanning and post-scan processing, post-scan image splitting, and blipping

Promoting COM as a preservation product by
- adopting common guidelines for image capture
- adopting common quality procedures
- reexamining the recommended film format to include 16mm or microfiche

Furthering development of metadata for digital books and journals by
- stipulating where to record the metadata, syntax, and workflow
- standardizing terminology used by different institutions
- developing models of the internal structures of books and ways of representing those structures in a digital environment
- determining when to record the structure of the work—before or during scanning
- developing commercial products to automate the process of paginating and
- indexing digital images.

Underlying all of these issues is the need to develop ways to track and assess the real costs of conversion projects, with greater and greater numbers of institutions reporting on their allocation of resources to enlighten the community generally about such a critical investment into preservation and access.

We recommend that this paper serve as the starting point for further collaboration to find answers to these questions. One or more meetings with concerned partners could develop a consensus among cultural institutions engaged in hybrid conversion and the groups that support such work, such as vendors, technologists, and funders. Such consensus could inform the funding of preservation and access programs by individual libraries and federal funding agencies.