MAGAZINE

Yale University Library's Project Open Book

Preliminary Research Findings

Paul Conway
Head, Preservation Department
Yale University Library
Paul.Conway@Yale.Edu

D-Lib Magazine, February 1996

ISSN 1082-9873

Abstract

This is a summary report on the production-conversion phase of Project Open Book, a program at Yale University Library exploring the feasibility and costs of converting preservation microfilm to digital image files. The report describes four clusters of issues addressed in the production conversion phase--selection, quality, administration, and cost--and outlines the next steps for the project.

Introduction

Project Open Book is a multi-faceted, multi-phase research and development program at Yale University Library. Its purpose is 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 has three overall goals. First, create a 10,000 volume digital image library and, in doing so, evaluate issues of selection, quality, and cost. Second, enhance intellectual access to the image files by creating structured indexes. Third, enhance physical access to digital materials by providing distributed access over the Yale University campus network. The library's Preservation Department is home to Project Open Book. I have been the principal investigator since 1993.

This article is a preliminary and somewhat provisional report on the recently completed production-conversion phase. This phase resulted in the conversion of 2,000 volumes in a twelve-month production cycle. There were four important outcomes of this phase.

- a practical test of the selection theories developed in earlier phases of the project

- an assessment of image quality obtainable from baseline microfilm conversion hardware running in a production mode
- an assessment of the internal management requirements for a large scale digital conversion program
- a detailed cost model and a test of the model using actual production time data

In the following sections, I will describe issues associated with selection for digital conversion, image quality, staffing considerations, and conversion costs and summarize the research findings. A full report on the production conversion phase is in progress and will be published initially on the HomePage of Yale's Preservation Department and later this year on paper.

[Return to top of page]

Background

Since 1991, Project Open Book has unfolded in a sequence of phases, designed in part to allow the project to evolve as the digital imaging marketplace changed.[1] In the first phase--the organizational phase--Yale conducted a formal bid process and selected the Xerox Corporation to serve as its principal partner in the project.[2] During the second phase-- the setup phase--Yale acquired a single integrated conversion workstation, including microfilm scanning hardware and associated conversion and enhancement software; tested and evaluated this workstation; and made the transition to a fully engineered production system.[3] The Commission on Preservation and Access provided partial funding for the setup phase. In the third phase--the production conversion phase--Yale built a multi-workstation conversion system, hired technical staff, converted 2,000 volumes from microfilm, indexed the volumes, stored the results, and tested a prototype Web access tool developed by Xerox. Generous support for the production conversion phase was provided by The National Endowment for the Humanities.

During the production year--September 1994 to August 1995--we held constant several important aspects of the project to increase the reliability of our findings on quality and cost. There was no staff turnover in the project's management or production teams. The physical setup of the conversion laboratory (Fig. 1) and the equipment configuration of the project (hardware and software) did not change. We fine-tuned the file management process (Fig. 2) early in the project and did not modify it appreciably. Mid-way through the project, we established an Ethernet sub-net (Fig. 3) to improve file transfer speed but did not swap out any of the hardware or software components of the network. Figure 4 is a key to the room layout, the local area network, and the file management process that is integral to the efficiency of the image conversion process.

Conversion of microfilm frames to digital images is a complex process in many ways similar to the process of creating microfilm of brittle books or manuscripts. The workflow of Project Open Book has been described previously in a report on the Setup Phase . The most important innovation of the project workflow was the worksheet that staff completed for each converted volume. A full worksheet (Fig. 5) contains a wealth of important information: bibliographic details on the book, film and book characteristics discovered during pre-scan inspection or in the midst of the conversion or indexing steps, file specifications, and microfilm scanner filter settings. Most importantly, project staff recorded the time it took, in minutes, to complete each of ten processing steps for all 2,000 volumes converted in the project. Corroborative data from daily work logs validated the accuracy of the volume processing data. Staff used the worksheets to manage workflow. I used the data to project the costs of the conversion process and to identify some of the factors influencing those costs.

[Return to top of page]

Selection

The point of departure for Project Open Book is the ongoing program administered by the National Endowment for the Humanities (NEH) and known as the Brittle Books Program. Its goal is to preserve on microfilm three million crumbling books selected by participants in the program from their high-quality research collections. The overall selection strategy of the Brittle Books Program calls for participating libraries to identify large, significant, subject-oriented humanities research collections rich with publications from the 19th and early 20th centuries.

Yale University Library has been an active participant in this program since its inception in 1983. Over the past ten years, Yale preserved on microfilm roughly 15,000 volumes from the American History Collection (1983-93), 23,000 volumes from the European History Collection (1988-93), and 19,000 volumes from the History of Economics and Political Science Collection (1992-95). A major preservation survey a decade ago identified each of these collections as a top preservation priority.

Selection for the production-conversion phase of Project Open Book started with this large body of preservation film. The goal emphasized content cohesion over the technical limitations of the digital imaging system. Project staff hoped to identify significant clusters of film on subjects of interest to Yale's faculty and students. By connecting selection with expected use, a known population of scholars could help evaluate the end product and its usefulness for scholarship.

In the interests of project efficiency and productivity, the production plan called for converting all of the titles on a given reel of film. An explicit decision was made not to "de-select" a particular title from a chosen subject cluster simply because image orientation, film contrast or density, and other technical matters stretched the capabilities of the project's scanning equipment. Reasoning for this decision factored in curiosity to discover the frequency and nature of "problem" books, and a need to measure the impact on the conversion process of these books.

We encountered a number of challenges to selecting by content from a single library's collection of microfilm.

- *Yale's microfilm collection resembles Swiss cheese*. Given the national mandate to avoid duplication, Yale did not film any volume in its collection that already had been preserved on film at another institution or that fell out of the date scope of the project.
- *Borrowing film for conversion is a bother.* Using "Other People's Film" takes all of the intellectual energy required to reformat the volume in the first place: searching for the existing film, matching records and then content, and worrying about quality and completeness. More significantly, libraries are very reluctant to loan their negative film, which happens to be the most appropriate polarity for digital conversion.
- *As Elvis would say: "I'm All Shook Up.*" Volumes on many different topics can and do appear on any given roll in the materials selected for Project Open Book. Like most libraries, Yale leaves "reel programming," which is the process of grouping volumes with similar physical characteristics, to the vendors who create the film. Even when programming is handled in-house, meaningful arrangement by topic is usually not a goal and the result is intellectual chaos from reel to reel.
- *Slash and burn preservation*. For some key collections in a single library, most of the brittle books are now gone. Although the content is preserved, our ability to undertake quality benchmarking or to verify the accuracy of the reduction ratio of the filmed book is severely hampered if the original volume is in a landfill.

Through luck and sheer determination, project staff were able to identify four clusters of titles in the "Old Yale" classification system that contained a critical mass of microfilmed titles from the original collection AND that were of interest to Yale faculty and students. These four clusters are:

- Civil War History (1,000 vol.)
- Native American History (200 vol.)
- History of Spain Before the Civil War (400 vol.)
- History of Communism, Socialism and Fascism (400 vol.)

Yale library bibliography staff played a key role in reviewing this work, assessing several years worth of course offerings at the graduate and undergraduate level, contacting faculty in three disciplines by phone and letter, and obtaining commitments from them to use the project's digital image files for research and teaching. Without this effort, the selection process would have been largely an "academic" exercise.

The Brittle Books Program is creating the first "virtual library" in the world that also happens to be a vital source for digital conversion. This library is largely underground and exists as an "entity" only in national bibliographic databases. Selecting from this national collection of microfilm to help create part of the new national digital library then becomes, by nature of the conditions of its creation, a responsibility that transcends the policies of any of the libraries that have participated in the Brittle Books Program to date.

[Return to top of page]

Quality

Today's digital conversion technology can produce spectacular, high-quality results with the use of very specialized hardware and software and enough time to find just the right image of each and every original document. The quality goal of Project Open Book was to find the maximum quality obtainable with the most widely available equipment running with minimum operator intervention. "Production quality" scanning entails accepting compromises in any number of areas. This section summarizes three aspects of digital conversion of microfilm that have an impact on quality: resolution, depth, and technical rigor of the film.

Resolution: Based upon extensive experiments during the setup phase of the project, the production-conversion phase scanned microfilm at an "effective resolution" of 600 dots-per-inch. If the filmed item were enlarged to its original size then 36,000 pixel points of data would be recorded for each square inch of surface. Figure 6 is a sample of text produced from one frame of microfilm at this resolution. One firm conclusion from the project is that high-resolution scanning (at least 400 dpi) is essential for "preservation quality" digital conversion of text.

Depth: The production-conversion phase utilized binary rather than gray-scale scanning technology. The data produced from a binary scan can be highly compressed using standardized algorithms and, therefore, takes less time to transmit and less money to store. The film input source was the second generation negative, referred to as the "print master" in the microfilm trade. For a variety of technical reasons, negative film produces a better quality digital image than positive film. Another firm conclusion from the project is that high-resolution binary scanning produces high quality digital images of **text**. This quality meets the criteria outlined by Anne Kenney and Stephen Chapman in their recently-published tutorial on image quality.[4]

Technical rigor: In order to scan 35 millimeter microfilm at 600 dpi and create a single digital image for each discrete page of a filmed book, Amitech Corporation redesigned the Mekel 400XL and added special software. The resulting equipment can obtain maximum quality only with books oriented on the film in a particular way (Fig. 7). Rigorously filmed books produce a better quality end product. Some of the technical problems that contribute most significantly to poor image quality are skewed frames, weaving center lines, irregular spacing between frames, significant density variation across a frame, and internal film splices. A third firm conclusion from the project is that high-quality digital images can be generated from film created within the standards and guidelines for archival-quality preservation microfilm. The incremental cost required to produce rigorous film will pay off in higher quality digital images produced at a lower cost.

[Return to top of page]

Project Staffing

We administered the production-conversion phase of Project Open Book with the explicit goal of learning about large-scale image conversion projects in libraries. We used the project to find out how to contract with a service bureau for conversion services and then manage one or more focused projects with confidence about both the cost-effectiveness of the process and the quality of the end result. Of the three major aspects of project administration that we explored in detail--vendor/partner relations, in-house technical support requirements, and project personnel recruitment and training--the most broadly applicable lessons derive from our experiences creating the three-member production team.

At the beginning of the project, Yale employed no staff with experience in digital image conversion. From the findings of previous phases of the project, we knew that meeting the production goals would require a team of three technicians, each one with a slightly different set of skills. The Library's long-standing commitment to nurture the skills of its existing staff -- combined with conditions imposed by a unionized clerical and technical staff -- argued strongly for the creation of a new "job family" within the Yale's personnel structure. Based upon the specifications of the project director, the library's human resources staff prepared three "generic" job descriptions with increasing levels of skill and responsibility (Technical Assistant I, II, and III) and then derived specific descriptions of the duties for Project Open Book. We sought broad consensus on task definition from University departments (for example, the computing center, the medical school complex, and the Yale University Press) that in the future may need to recruit digital imaging technical support staff. The jobs were posted for two weeks in the University's employment system. Recruitment took place from the pool of applicants.

The three staff hired for the project (Fig. 8) exhibited a wide diversity in background experience and technical training. Joe Cinquino, Jr., a recent college graduate, operated the project's digital scanning equipment. As is the case with kids who have grown up with interactive video games and computers in the classroom, Joe is a savvy user of graphical user interfaces and is comfortable earning his livelihood through a keyboard and a mouse. JoAnn Teadtke, on the other hand, came to the project with limited computer experience. Previous library jobs trained her to catalog books and process thousands of brittle volumes for preservation microfilming. JoAnn served as the principal indexer for the project. Bob Halloran led the project team and was responsible for workflow scheduling, quality control coordination, and file management activities. Like JoAnn, Bob had limited experience with computer technology and no previous experience with digital imaging. As chief technician in the library's in-house microfilm laboratory, however, he brought to the project invaluable knowledge of the characteristics of microfilm and techniques for inspecting film prior to conversion.

The entire personnel recruitment process added eight full months to the project. Balanced against the frustrations inherent in such a delay is the positive benefit of a permanent personnel structure in the University community. Our decision to recruit staff with skills analogous to the ones we needed for the project, and then to accept responsibility for computer training, turned out to be a wise one. In today's technology environment, it is far simpler to enhance the computer skills of a novice user and provide training on specific applications than it is to give a skilled computer operator a feel for the visual, tactile, and content characteristics of the library material selected for digital conversion. As part of the next section will show, personnel training issues are a significant cost factor in digital imaging project. Yet the "learning curve" can be predicted and the economic benefits of supporting a well-trained technical staff can be measured and then factored into the overall cost of the project.

[Return to top of page]

Cost

The most important research findings of the production-conversion phase of Project Open Book concern the cost of converting preservation microfilm frames to digital images. No systematic study of microfilm conversion exists in print. In general, published studies on conversion from paper do not adequately consider:

- the cost of equipment purchase, lease, maintenance, and replacement;
- the sub-processes that comprise the overall digital conversion process;
- the factors that contribute to conversion costs; and,
- the possibilities for improvement in processing efficiencies from staff training and practice.

We began by developing a model of the microfilm conversion process (Fig. 9) to guide the collection and analysis of data on equipment and labor costs. In its simplest form, the model says an analog source is converted to digital data and this data is made available. We assumed, however, that microfilm is a complicated conversion source. Characteristics of the filmed book (e.g., size, text clarity, type of illustration, and physical condition) interact with characteristics of the film itself (e.g., reduction ratio, minimum and maximum density, and the technical precision with which the film is created). The conversion process itself is multi-faceted. The total labor cost of the process is the sum of the costs of a number of sub- processes, including inspection, scanning, indexing, and acceptance for storage of the image file on an appropriate medium. The components of the equipment configuration vary in cost and have distinctive technical limitations that affect throughput speed. Finally, access is not a simple end result of the conversion process but rather is a complex process of displaying or outputting image and index data on a variety of platforms connected to a communication network.

The following paragraphs summarize the complex research findings in four areas: annual equipment costs; process time and cost; the impact of book and film characteristics on process time; and the role of staff training on process costs. A complete discussion of the findings is in the works and will be available before the tulips bloom in Connecticut.

Equipment costs: Figure 10 displays the actual costs of hardware, software, integration and support and optical storage media supplied by Xerox Corporation and then converts these costs to a range of per-book and per-image costs. Per-book costs are based upon estimates of the throughput capacity of the system during a single shift. The lowest capacity is 3,194 volumes per year; the highest capacity is 4,166 volumes per year. Per-image costs are calculated by dividing the per-book cost by 216, which is the average number of digital images per book in the project. The equipment replacement model assumes a five-year amortization and an additional 50 percent replacement surcharge for increased functionality. The bottom line is equipment costs of $31.32 per book or 14.5 cents per image.

Process time and cost: Figure 11 has details about the distribution of processing times for the first 600 volumes in the project. Process costs for each of the ten steps are derived from the time data by multiplying the process mean (average) by $0.2563, which is the "benefitted minute" wage rate of the project staff in 1995 dollars. The table displays the actual minimum and maximum values for a given process step, in part to illustrate how median and mean values vary. A few large-value outliers were discarded during certain data analysis procedures. The bottom line is that a typical book in the project took just over 92 minutes to process, at a cost of $23.71 or 10.9 cents per image.

Book and film characteristics: Figure 12 is a complex presentation of the statistical impact of various book and film characteristics on the distribution of time for each of the ten processing steps highlighted in Figure 11. The row marked "In-process model" is the proportion of overall variation in time explained by characteristics identified during the conversion process. The row marked "Pre-scan model" (in bold) is variation explained by factors identified prior to scanning. Characteristics that contribute significantly to the model are marked with an "X". Those in bold pertain to the "Pre-scan model."

In essence, the table shows that film characteristics, for example reduction ratio, density, and "technical rigor," have relatively little impact on conversion costs even if they can make or break digital image quality. The good news in this conclusion is that we can obtain or exceed quality conversion from "poor film: with only a marginal increase in overall conversion costs of "good film." The findings also suggest that significant investment in improving the quality of new film will probably not pay off in terms of reduced conversion costs.

Book characteristics like tight gutters, yellowed or faded paper and inks, and similar factors associated with deterioration, damage, or heavy use tend to increase the costs of most of the processing steps. There is very little we can or should do about this fact, however, because the preservation imperative should not control digital image selection processes. The findings will allow us to predict the incremental increases in cost required to digitize "difficult books" in comparison to "easy books."

Practice effect: Digital image conversion is not an automatic process. The talents of the people who operate the equipment, master and then simplify the process, and add value to the end product can have a tremendous impact on overall costs. Figure 13 shows just how dramatic the impact of training and practice is on processing costs. The table compares the average processing times (and costs) of a 600-volume sample with the costs of the process for the fist and last 50 volumes in the sample. The important thing to know about these figures is that they control, statistically, for all of the film and book characteristics identified in Figure 12, as well as the varying sizes of volumes converted. What's left is improvements in staff efficiency, including simplifying the process itself. In the figure, the time to complete four scanning steps drops from 37.8 minutes to 21.1 minutes--a 44 percent productivity improvement. Similarly, the efficiency of the two indexing steps improved by 50 percent from the beginning to the end of the 600 volume sample.

[Return to top of page]

Next Steps

Two immediate next steps are needed. Each of the 2,000 volumes of imaged books should be available for consultation. In practical terms, this means creating bibliographic records in Yale's on-line catalog. Yale University Library is well-along in planning to provide patrons access to its catalog on the World Wide Web. Links in the displayed bibliographic record will provide direct access to digital image files converted in Project Open Book.

Another important step is the design and implementation of a comprehensive evaluation of the quality of the image product, the usability of the system for research and teaching, and the usefulness of the content from the perspective of faculty and students. There are many complex issues associated with the design of such an evaluation, not the least of which are the need to control for variations in interface design and the narrow content focus of the project. Fortunately for everyone concerned about building digital libraries that are sensitive to the information needs of prospective users, evaluation of digital libraries is a hot topic. A recent conference at the University of Illinois set the methodological agenda. An upcoming research seminar at UCLA is expected to produce a meaningful research program. Conference sessions and workshops at the First ACM International Conference on Digital Libraries (DL'96) will also focus on evaluation.

By mimicking a large-scale production environment in a relatively small-scale laboratory, the production-conversion phase of Project Open Book has done much to meet the project's overall goals. Because the project staff managed directly nearly all technical aspects of the project, we were able to limit the unknowns that plague an assessment when contract vendors must be counted on to supply information on costs, production workflow, and image quality. Cost and process models developed for the project are generalizable to other production-level digital imaging projects. The comprehensive findings shed important new light on the opportunities for libraries and archives to convert their extensive research collections from microfilm frames to digital images.

[Return to top of page]

References

1. Waters, Donald J. *From Microfilm to Digital Imagery*. Washington, D.C.: Commission on Preservation and Access, June 1991.[Return to footnote 1]

2. Waters, Donald J., and Shari Weaver. *The Organizational Phase of Project Open Book*. Washington, D.C.: Commission on Preservation and Access, September 1992.[Return to footnote 2]

3. Conway, Paul and Shari Weaver. *The Setup Phase of Project Open Book*. Washington, D.C.: Commission on Preservation and Access, June 1994.[Return to footnote 3]

4. Kenney, Anne R. and Stephen Chapman. *Digital Resolution Requirements for Replacing Text- Based Material: Methods for Benchmarking Image Quality*. Washington, D.C.: Commission on Preservation and

Access, 1995.[Return to footnote 4]

[Return to top of page]

Home Magazine Comments
NEXT ▶

hdl://cnri.dlib/february96-conway

Typo in handle corrected, Editor, June 25, 1998