
BUILDING ON THE CORNELL-YALE MODEL: DIGITIZING THE RADICALISM COLLECTION AT MICHIGAN STATE UNIVERSITY

Michael Seadle

Introduction

This article has two purposes. One is to summarize the elements of the Cornell-Yale digitization model in a way that would be useful to other institutions planning to follow their efforts. The people involved with the Cornell and Yale projects have done an exceptional job in documenting their work. They have been the Lewis and Clarks of the digitization process. But because they were the first to explore this unknown territory, not every facet of their experience applies to those who follow the path they blazed. It is necessary to see which parts of their experience are applicable to a large state university like Michigan State University (MSU). The other purpose of this article is to show how MSU's experience has expanded the Cornell-Yale model. In this MSU is following their tradition of documenting features that could be useful and relevant to others.

Seadle is digital services librarian at Michigan State University. He also worked at Cornell University Library when the CLASS project was getting underway. < seadle @pilot.msu.edu > .

The digitizing projects at Cornell and Yale have provided libraries with a legacy on how to produce scanned images efficiently.¹ Although the papers that emerged from their joint experience do not use the word model, that is clearly what their pioneering work represents. They made a conscious attempt to define procedures and standards, and have left a detailed record of their achievements. Together they transformed the digitization process from experimental to production status. This article examines a project at Michigan State University (MSU) which built on the Cornell-Yale model while changing three key factors:

- 1) **Different hardware**—MSU was one of the first users of the Minolta PS-3000 top-down scanner, which offers significant advantages to a library that wants to scan books efficiently without damage to their bindings. Cornell and Yale both used flatbed scanners.
- 2) **Course-based selection**—MSU links the selection of items for digitization directly to the undergraduate teaching program. Cornell and Yale linked selection to long-term scholarly use.
- 3) **A freshman-oriented access-system**—MSU uses a simple Web-based access system designed to serve the needs of freshmen with little or no library research training. Cornell and Yale focus on the access needs of scholars.

The goal of this article is to add to our understanding of the digitization process by showing how each of these changes affected its implementation of the Cornell-Yale model. A five-part framework by Paul Conway of Yale provides a convenient and consistent structure for the analysis. The five parts are²

- **Choice**—The selection process including issues like copyright;
- **Quality**—The resolution (dpi) and depth (in bits) of the initial image as well as its display derivatives;
- **Integrity**—The question of what constitutes an accurate representation as well as an image's provenance;
- **Longevity**—Long-term storage and refreshment issues; and
- **Access**—The access system and related options, including the choice between online and paper final products.

Before beginning the analysis, I will provide overviews of the Cornell-Yale model and the MSU project to give readers some familiarity with the context. I will also briefly describe an alternative approach, the widely used SGML model.

The Cornell-Yale Model

The core of the Cornell-Yale model comes from two projects done in collaboration with the Commission on Preservation and Access and Xerox Corporation. The first of these was Cornell's CLASS (College Library Access and Storage System) project, which began in January 1990, after a period of substantial planning. Cornell used a Xerox scanner that was still under development. It had a nominal capability of 600 dpi (dots per inch) resolution, which it achieved with a 400 by 400 dpi aperture that sampled at 400 dpi with 8-bit grayscale and then resampled at 600 dpi using a multi-line averaging algorithm.³ Images were saved in Aldus/Microsoft TIFF Version 5.0 using CCITT Group 4 compression. The end-product was a printed facsimile from a Xerox Docutech printer. An X-Window-based print-request-server was to be the primary online access method. It is important to remember that this planning predated the expansion of the Web.

Cornell's initial project scanned 950 books, which resulted in about 285,000 digital files. The volumes had to be disbound and the edges trimmed parallel to

the text. No sheet feeder was used initially because of concerns about the fragile condition of the paper. In production mode, trained technicians could do about five pages per minute.⁴

A number of important conclusions emerged from the CLASS project: 1) digital imaging provides a cost-effective alternative to photocopying; 2) it offers a cost-effective alternative to microfilm preservation as long as the concept of technology refreshing is institutionalized; 3) it has the potential to enhance access; 4) it can facilitate internal access and provide links to the library catalog through document control structures; and 5) its infrastructure "supports other applications in the electronic dissemination of information."⁵

Yale's Project Open Book consciously built on Cornell's early experiences. It began in 1991 with a report that became its master plan for using microfilm as the source medium for scanning rather than the paper originals. The scanner was a Mekel M-400 microfilm scanner controlled by Ameritech TurboScan software. The scanner's maximum capacity was 7,042 pixels per inch (ppi) at 16 millimeters or 3,690 ppi at 35 millimeters. Although these numbers look high, the reduction ratio of microfilm must be taken into account in order to get the 600 dpi equivalent from an original page.⁶ These images also were saved in TIFF using CCITT Group 4 compression. The end-product was designed as a network-based system, which allowed for browsing the page images at a relatively low resolution (200 dpi) and for printing them at a high resolution (600 dpi).⁷

Project Open Book aimed at converting 10,000 volumes that had previously been microfilmed. All of the works had been selected as part of the National Endowment for the Humanities-funded Brittle Books Program, and often the originals no longer existed.⁸

Among the major lessons of Project Open Book were 1) that there is a significant practice effect where scanning efficiency improves as staff learn how to master the process; 2) that the technology for digitizing continues to mature and improve as prices fall; and 3) that scanning from the original, when possible, "will almost always produce better quality results than scanning from a microfilm intermediary."⁹ The project was also significant as a test of the selection theories and internal management requirements for a project of that scale.¹⁰

Two key characteristics of the Cornell-Yale model are a concern with reproducing a photographically exact image of the original and the use of paper facsimiles as a user-friendly medium. In a sense, the model takes the well-established tradition of preservation microfilming, discards the user-hostile hardware, and significantly enhances access through the use of modern computer technology. Large-scale production capability is an important goal for the model. Changing the way that

people use written works by adding new search and analysis tools is not.

Since the Cornell and Yale projects, others have built on this model. One of the most notable efforts has been the Mellon Foundation-funded JSTOR project. JSTOR has scanned the back issues of over 40 academic journals. Photographically exact 100 dpi images are available for browsing on the Web, and 600 dpi images are used for printing to a user's locally attached device. Both are typical Cornell-Yale model features. Also typical is the scale of the project. The number of journals is less significant than the fact that JSTOR includes complete runs from the first issue up to five years prior to the present (by agreement with the publishers). It offers a complete solution to the problems of missing issues, razored articles, storage space, and other typical journal-preservation issues. JSTOR has also extended the Cornell-Yale model in a couple of ways. It has added the ability to search invisible and slightly (0.05 percent) imperfect OCR'd versions of each page.¹¹ And it has built on the Cornell-Yale economic model by attempting to become a self-supporting organization by charging for its services.

The SGML Model

The Cornell-Yale model is not the only one available to libraries interested in digitization. A second widely used approach might be called the SGML (Standard Generalized Markup Language) model, such as is used at the Humanities Text Initiative of the University of Michigan and at the Electronic Text Center of the University of Virginia. The SGML model does not use photographically exact images as its chief product. Instead it aims at getting essentially correct text (≥ 99.995 percent) through OCRing, transcription, and various editing, spell-checking, and multiple-comparison processes. This text is then tagged using SGML, which allows the specification of key structural units as well as dates, pages, and other significant features, depending on the definitions used. Those interested in an overview of this model should see Timothy W. Cole and Michelle M. Kazmer, "SGML as a Component of the Digital Library," in *Library Hi Tech* 13:4 (1995): 75-90.

This model is particularly suitable for longer works that are likely to be analyzed in detail. It adds significantly to the value of the original because of the text's searchability, but loses that value again when printed. This model also focuses on the intellectual content and is prepared to sacrifice the original context that a photographic process saves. The SGML approach is neither better nor worse than the Cornell-Yale model. Indeed the ideal for the future may be to combine them. For now, given current technology and current costs,

the Cornell-Yale model may serve larger-scale projects where careful textual analysis is not important, and the SGML model may do best for focused projects where the added value of computerized searching is needed.

MSU Project Overview

The MSU project began in 1997 with a debate about which model to follow and some experimentation with different display techniques. A UMAX D016L scanner was initially used to digitize some items from the Radicalism Collection;¹² then Adobe software was used to OCR, to edit, and finally to build a PDF file. Although PDF (Page Definition File) is a proprietary format, it is widely used by United States government agencies. Both the image and print quality were excellent, but three concerns arose: 1) that some students might balk at downloading and installing the (free) Adobe Acrobat software required to view and print; 2) that the OCR'd text contained errors and did not give a completely true representation of the original; and 3) that the cost per item was high because of the OCRing/editing process. MSU decided instead to follow the Cornell-Yale model because of a concern for presenting a photographically exact and error-free image, so that processing errors would not misrepresent the care (or relative sloppiness) of the radical groups, and because of a need to produce a significant number of images quickly and inexpensively to meet evolving classroom needs.

Student needs, not preservation or scholarship, were the driving force behind the MSU project. Over the years students from MSU's basic freshman writing classes have come to Special Collections to use the Radicalism Collection for their research papers. From the student viewpoint, going to Special Collections was inconvenient because of its limited hours (9:00 a.m. to 5:00 p.m. Monday through Friday and 9:00 a.m. to noon on Saturday). These were not prime freshman paper-writing times, and the weekday hours were prime class time. For students with heavy schedules, the restricted hours represented a legitimate problem. When scanning technology became available, Peter Berg, head of Special Collections, conceived of a project that would make the most heavily used materials available in digital form, so that students could use them from their dorm rooms or from library workstations at any hour, day or night, weekday or weekend.

Most of the scanning for the project has been done on a top-down Minolta PS-3000. The physical structure of this machine makes it relatively easy to get alignment correct. For most materials, the pages need only be turned, scanned, and turned again. The Minolta also can create two separate images from the two-page spread of an open book, a useful time-saving feature.

SIDEBAR 1: MSU DIGITIZING PROCEDURES

Introduction

Digitization procedures have evolved in the roughly ten months that MSU has owned the Minolta scanner. The initial procedures used JBIG compression, which, though efficient, was unreadable by other software such as Hijaak Pro and Adobe Photoshop. The initial procedures also allowed two-page spreads for books with small formats to save paper when printing copies. But the two-page spreads ultimately seemed confusing when used with the page-turning scripts on the Web. Although these mistakes cost time (and money), they were part of the learning process in dealing with a new technology and new machine.

The current digitizing procedures may be summarized as follows:

Scanner Set-up

- 400 dpi (the maximum available at present);
- CCITT Group 4 compression for TIFF files (compatible with Hijaak, and the same compression algorithm that Cornell used);
- bit density for black and white images only (which works well for text).

Materials Preparation

- Check that item fits within the 11" by 17" scanned. Items whose pages are greater than 11" by 17" in fact cannot be scanned because the covers push the page over the outer limit.
- Check that the thickest point of the item when opened does not rise above the back-plate mirror.
- Remove any tags or other items except the covers that stick out beyond the page-block. These will confuse the scanner and make it think the item is larger than it really is.

Scanning

- Items should be up against the ridge in front of the scanner, not over it, and not away from it. This will ensure that items are lined up straight and will position items so that the software can use the mirror to make correct size calculations.
- If pages must be held down, fingers should be placed at the sides, not the bottom, and should not be too close to the text. This will allow the software to remove the finger image properly, and not to mistake the fingers for an extension of the page.
- Either the left-page-only, the right-page-only, or the two-pages-at-once option should be used, except when a single page occupies the whole scanned. This will provide a single page image. Either the foot pedals, the scanner buttons, or the clickable controls on the PC may be used.
- Batch processing rather than single-image processing should be used for any items that are more than two pages. This will ensure that the page images remain in correct order.

Post-Scanning

- After each page is scanned, it should be inspected on the PC screen for fuzziness, misalignment, or other flaws. Small letters especially should be checked. The images should be deleted and rescanned immediately if necessary. If rescanning does not help, a supervisor should be consulted.
- If the Minolta has included too much margin, which it does frequently, the image should be cropped using the mouse pointer and expand tool.
- The image can then be saved into the batch.

The Minolta is much faster than the UMAX, but it does have several limitations. The present version of the machine cannot do color, will not scan at densities greater than 400 dpi, and is "incapable of outputting true grayscale."¹³ Fortunately a board from Picture Elements will soon be available to fix the grayscale limitation and allow 600 dpi images, which should meet most preservation requirements. The Minolta also cannot scan books that are over about three inches thick (because of the way it uses a mirror to correct for page curvature) or books that are over 17" by 11" wide (because of the size of its scanning bed). And, it works

only with proprietary "Epic 3000" software, which has minor bugs such as still requiring the old DOS eight-character format for names for its MS-Windows 95 version. The machine's advantages are that it can scan fragile books that cannot go face-down on an ordinary flatbed without serious stress to the pages and binding, and that the software's batch processing mode automatically numbers images and keeps them together.

A description of MSU's scanning procedures is available in sidebar 1. Once a substantial number of images were ready, MSU began to build an interface oriented to the needs of undergraduates. A description

of MSU's display procedures is in sidebar 2. MSU also put paper copies of the images on course reserve (see figure 1) and plans to survey students about which improvements they most want.

Issues of Choice

Choosing the right materials is critical to any digitizing project. As Paul Conway explains, "Selection is the central intellectual challenge of any program that has a goal of creating a corpus of useful and meaningful digital research materials."¹⁴ A project that digitizes a jumble of intellectually unrelated materials wastes resources, even when individual items are gems. The gems get lost, and the whole corpus is much less than the sum of its parts. Because the Cornell-Yale model emphasizes large collections, choosing what to include can affect a project for years or even decades to come.

Cornell initially chose a set of mathematics texts which were rare, but not so rare that processing damage to the original artifact would eliminate the world's last copy. It chose the mathematics field also because researchers were likely to have workstations powerful enough to view the images—a more important factor in 1990 than today. Yale chose microfilm that had been part of the Brittle Books project partly because the material had already been validated at a national level and partly because of the scale of Yale's involvement in the project (roughly ten percent of the whole).¹⁵ But it also presented some problems. The "no duplication" requirement of the Brittle Books program meant that Yale did not necessarily own a complete set of film for logically coherent collections. Also the films themselves sometimes contained multiple works which did not go logically with the main item that was wanted. Yale included them anyway "in the interests of efficiency and productivity," and perhaps also with the thought that they fit into a larger goal of eventually digitizing all the Brittle Books microfilm.¹⁶

The ideal selection plan in the Cornell-Yale model has nationally significant goals, but does not necessarily rely on any closely coordinated plan of cooperation. A task force that included both Ross Atkinson of Cornell and Gay Walker of Yale gave as one of its recommendations in 1986 that "[t]he most effective and practicable method for cooperative preservation selection is to base such cooperation upon discrete subject collections in individual institutions."¹⁷ This is one of the fundamental ideas behind the "great collections" approach to preservation, which assumes that major research institutions have discrete groups of materials that they feel are important and worth preserving. It relies on the selection decisions of past generations of selectors for coherence. Selection item by item would probably be impossible for any truly

SIDEBAR 2: DISPLAY PROCEDURES: MSU DISPLAY PREPARATION PROCEDURES

Display preparation for images follows the model used by JSTOR in creating GIF files with a lower dpi that will not take long to load on the Web and will still be readable. Black and white images are converted to grayscale to enhance clarity.

The current display preparation procedures may be summarized as follows:

- **Resequencing**—This is necessary because the Minolta's Epic software uses random hexadecimal numbers for file names, and builds a separate control file to keep them in order. A locally written program reads the control file that Epic produces and renames the files as 001.TIF, 002.TIF, and so on.
- **Resizing**—Each image is resized using Quarterdeck's Hijaak Pro. Staff first test an image from within the set (not the first page, which is often an unrepresentative title page), to create the smallest file that is clearly readable online. Good judgment is critical at this step. All files are then converted. The original TIF files of course remain unaltered.
- **Uploading**—The files are uploaded to a Unix-based server using file transfer protocols. The call number of the item identifies the folder for the pages. Nothing else goes in that folder.
- **Building HTML access**—Staff build HTML files with the title, author, image number, forward and backward controls, and links to the subject and alphabetical indexes. Each new title also must be added to the subject and alphabetical lists. The work is done with a combination of Perl programs and manual coding.

large-scale project.

A related and compatible approach tries to identify a core collection. This was used at Cornell's Mann Library, the main library for the university's College of Agriculture and Life Sciences. Samuel Demas, Mann's head of collection development, wrote, "Selection for preservation in certain central components of a discipline's total literature is most effectively conducted by systematically identifying the core, or most significant, historical literature."¹⁸ One key difference between this and the "Great Collections" approach is the amount of contemporary effort that must go into selecting individual items. The project at Mann rested on a grant-funded "monumental bibliographic effort,"



Digital Images from the American Radicalism Collection

Introduction

The Special Collections Division of the Michigan State University Libraries has begun to scan documents from its American Radicalism Collection and to make them available electronically. You can access the digital images in two ways:

Use the alphabetical list if you are looking for a particular work whose title you already know.

Use the subject list if you want works on a particular topic.

We have attempted to make the files as small as possible for easy loading without compromising legibility. You should be able to print the image files page by page.

If you prefer to read them in paper or to copy them at the Copy Center, photocopies are available on Reserve in the Main Library (1st floor, east wing). Look them up in the red Special Reserve notebook under the course name "Special Collections 101."

Other Collections

Comic Art Collection

Fables Collection

Michigan State University Student Activism

How to Print

If you want to print an image, first click on the image, then click on the Print button of your browser. If you do not click on the image first, your browser will not know which frames window you want printed and will do nothing. When printing wide pages, you may wish to change your printer setup from portrait to landscape mode.

Window Size

If you want to see more of the page image on your screen, you can put your cursor on the line that separates the image from the header, hold the mouse button down, and drag the line up. The header frame has a scroll-bar, so you can always scroll down to get to the page numbers for moving from one page to another.

Copyright

According to our understanding of copyright law, all of the works in this digital collection are in the public domain, and therefore have no copyright protection. If you have reason to think otherwise, **please let us know immediately.**

Please send comments to Special Collections

Page Created: 22 Sept 97 MSS

Last Updated: 22 Jan 98 MSS

Figure 1: The introductory page for the Radicalism Collection on the Web <<http://www.lib.msu.edu/spc/digital/radicalism>>.

which Wallace C. Olson directed over a five-year period.¹⁹ It is not clear that a similar project could have succeeded in even twice the time in a less scientific discipline like history where common ground for agreeing on a core literature could be hard to find. But it did succeed in agriculture, and, like the Great Collections approach, set up criteria for digitization that were national in scope and significance.

This summary of the selection goals at Cornell and Yale suggests that the model's key selection characteristics should include two factors. The first is coherence, particularly coherence at several levels. The selections not only should make local sense, but should also fit into a national plan that has significant goals for future scholarship. In terms of actual digitization work, these national plans have been more theoretical than real, since other libraries have not yet stepped in to do their parts, but that may be only a matter of time. The second characteristic is preservation. The whole thrust and justification of the digitizing efforts at the two universities has been the preservation of deteriorating books, and the projects have both been administered from the libraries' preservation departments. These characteristics stand in contrast to the SGML model, which emphasizes 1) the importance of individual works (since it puts more effort into fewer of them) and 2) the potential for added value from digitization (rather than the need for preservation).

MSU's project fits the Cornell-Yale model's characteristics. Coherence ranks high in importance. All the materials selected so far have come from the American Radicalism Collection, and particular topics have been pursued within it, such as the Black Panthers, the Sacco and Vanzetti case, the Ku Klux Klan, Students for a Democratic Society, and the Scottsboro Boys case. Preservation is a factor, too. Many of these materials were printed on the cheapest possible paper, which will soon begin to show signs of aging and acid damage. Bleedthrough is a problem on the leaflet literature from the vertical files, and poor-quality duplicating inks are beginning to fade. High demand from students has made the preservation problem worse. Although preservation photocopies have been made whenever possible without damage to the originals, photocopies are not seen as a good long-term solution. MSU's project might also be viewed as fitting into the Great Collections approach. Its radicalism materials represent a significant national resource, which historians of the future will want to access.

MSU's project did not, however, follow the Cornell-Yale model when it linked its selection process to the curricular needs of the freshman writing classes. This has some significant implications. On the positive side, it means that the digitized items will get immediate use. It creates a demand for digitization from a

prime tuition-paying population, which gets significant attention from the provost, president, and regents. This makes selection a fairly concrete, user-driven decision, rather than one that must guess at the future needs of future scholars and face the skeptical barbs of those whose concepts disagree. This is not to say that selection rests with random freshmen requests. Judgments are made on the basis of years of experience with the kind of research problems the freshmen bring to Special Collections, and on interaction with the faculty about themes they want to pursue in their classes. The faculty role is particularly important. Cooperation with the project gives them the opportunity to introduce issues about the problems with and advantages of writing papers using primary sources, and it gives Special Collections heightened visibility as a service provider for the university. The disadvantage of the curriculum tie-in is that it has potential for weakening the coherence of the digitized collection by leaving out parts that do not fit current curricular plans. That problem is negligible at this point, with the project just getting started, but may show up later as requests for new Radicalism Collection materials slack. At that point MSU will have to decide whether to spend time (and therefore money) on digitizing items with no immediate demand to maintain its commitment to coherence, or to shift focus to other areas where demand is stronger. It is impossible to say now how difficult that decision will be. The end result may be no worse than the gaps in Project Open Book from the Brittle Books no-duplication rule.

Issues of Choice—Copyright

Intellectual property considerations are important to every digitization project. The Cornell-Yale model avoided copyright violations by choosing materials that fell outside of any possible copyright entanglements, mainly pre-twentieth-century items. In this the Cornell-Yale model presents no unique features. An alternative would be to build on the experience of the JSTOR project and make contractual agreements with publishers to digitize more recent materials. JSTOR has shown that this is feasible with the right publishers and with due consideration for their economic interests.

For MSU, both approaches posed problems. Most of the heavily used parts of the Radicalism Collection post-date 1923, the latest date for absolute assurance that no U.S. copyright restrictions pertain. Many of the radical groups no longer exist, and searching for subsequent copyright holders (if any exist) could be a long and frustrating process. And much of the project makes sense only if MSU can digitize materials up to 1978, the date when the revised 1976 copyright law went into effect. This date is important because the

SIDEBAR 3: METADATA

MSU uses a Microsoft Access database to keep metadata about the images. The following fields are included:

- **Title**—for the main title of the item.
- **Author**—when there is an identifiable author.
- **Call number**—the call number at Michigan State University.
- **Disk location**—the location of the TIF file if it is on a hard drive on one of the work machines.
- **Backup location**—the name of the Zip disk (or tape) with the backup copy.
- **TIF dpi**—the dpi (dots per inch) for the digital master copy in TIF format.
- **GIF dpi**—the dpi (dots per inch) for the display copy in GIF format.
- **Scanned by**—name of the staff member who did the scanning.
- **Scanner**—name of the scanning device used.
- **TIF Density**—black/white, grayscale, or color for the TIF master.
- **GIF Density**—black/white, grayscale, or color for the GIF display copy.
- **Date Scanned**—the date when the TIF master was created.
- **Copyright Check by**—name of the person who checked for copyright clearance.
- **Number of Images**—the number of TIF files, which could be more than the number of pages.
- **Compression**—Normally CCITT Group 4.
- **Notes**—For a record of any exceptions in processing.

1976 law did not affect materials already in the public domain under the old 1909 law, and, according to copyright law expert Robert Oakley, a work entered the public domain under the 1909 law immediately if it was published without any copyright notice.²⁰ Since a significant part of the Radicalism Collection fits this criterion, it appeared to be safe to digitize.

Nonetheless the decision to digitize materials in this gray area from 1922 to 1978 imposes a burden not present in the Cornell-Yale model. A responsible staff member has had to check each item carefully for any evidence of a copyright notice, regardless of form. Also there has had to be some reason to believe that the item meets the legal definition of a “published” work, since unpublished manuscripts did not fall into the public domain under the 1909 law, and generally came under the protection of state law.²¹ To add to its legal safety margin, MSU has emphasized materials with no clear authorship. It has also attempted to document its copyright checking as part of the metadata for each item, so that it can show good faith in its intention to follow the law, should any doubts arise about particular items (see sidebar 3). Other institutions should not follow suit in choosing materials from the 1922 to 1978 era without taking equal trouble to inform themselves about copyright law and to check materials, even when they have materials that appear to be equally safe. Further, they should realize that the extra work involved clearly adds to the total cost of any digitizing project.

Issues of Quality

One common misconception is that the Cornell-Yale model set an inflexible quality standard for digital scanning at 600 dpi for all items. Quality in the Cornell-Yale model in fact calls for a much more complicated concept of “full information capture.” As Stephen Chapman and Anne Kenney explain, “The objective is not to scan at the highest resolution and bit depth possible, but to match the conversion process to the informational content of the original—*no more, no less* [author’s emphasis].”²² For printed materials, some factors include

- size of the document,
- size of details in the document,
- text characteristics,
- medium and support (such as pencil on paper),
- illustrations,
- tones and/or color, and
- dynamic range, density, and contrast.

Following the model faithfully requires careful measurement and analysis. Cornell workshops offer practice in doing this, and the accompanying handbook has formulas that help calculate the best dpi for a particular image. Going through the whole analytical process title by title can become time-consuming and expensive. Nor is it really necessary, except in a few difficult cases. Most people quickly gain enough experience to substitute a few rules of thumb, or just use the convenient

table that Kenney and Chapman have prepared, which gives the dpi requirements for particular heights at specified levels of quality.²³ Those who chronically overcapture pay a price in file size and storage costs, but many prefer this to the risk of having to rescan because of inadequate detail in the future. This is one reason why many sites default to 600 dpi. It feels safe.

Resolution in dpi is only one quality measure. Depth is another of nearly equal importance. Depth refers to the number of bits of information that the image file contains about any single dot or pixel. One-bit depth means a black and white image because only two possibilities exist in binary computer code for a single bit. Eight-bit depth allows grayscale with 2^8 or 256 shades. Twenty-four-bit depth is enough for millions of colors. Both Cornell and Yale have emphasized 1-bit images, which are appropriate for books where the paper is supposed to be white and the letters black. But for depth as for resolution, the principle of the Cornell-Yale model is full information capture. Items which in fact have a range of grays or colors should be captured at a depth that will reproduce them faithfully.

For Yale, scanning was not just a matter of setting the machine to a particular resolution, say 600 dpi. The resolution was in fact a "software-controlled mathematical artifact" that depended on the film's density and reduction ratios.²⁴ Library materials are generally microfilmed at ratios from 8:1 to 14:1, and to make matters worse the RLG handbook allows variable reduction ratios. The RLG handbook also set a range for density, which is a measure of the opacity of the film. Of course, the condition of the original—yellowing, print size, bleed-through—matters just as much in working from microfilm.

The Cornell-Yale model requires that all of these factors receive serious attention to achieve the goal of no more and no less than full information capture. With the SGML model there is some temptation to treat the digital image as an intermediate where quality matters mainly in terms of OCR accuracy. But this has not led those following the SGML model to produce poorer images—quite the opposite. Virginia, for example, routinely scans at 24-bit depth because the OCR results are significantly more accurate with the extra depth even though it is more than needed for full information capture.²⁵

Although the MSU project has subscribed to the quality principles of the Cornell-Yale model, it has had to work within hardware limitations in carrying them out. The Minolta scanner in its present form can do only 400 dpi resolution and is incapable of true grayscale. For many of the Radicalism Collection materials these limits did not matter. The radical groups tended to favor large type, so that 400 dpi sufficed to capture

the smallest element. The pages also tended to be simple black on white. But limiting the choice of items to those where full information capture could be guaranteed would have meant an arbitrary exclusion, which would undermine the Cornell-Yale principle that a collection have coherence.

There were options. More scanning could have been done on the UMAX, which was capable of full 24-bit color and resolutions well above 600 dpi. The UMAX, however, is a slow machine, and, because it is a flatbed scanner, the items would require more careful handling than on the overhead Minolta. Both factors would increase the time and therefore the cost per item significantly. Another option might have been to wait until the new 600 dpi-capable Picture Elements board arrived. Both options, however, would make the project less able to respond quickly to faculty requests and evolving curriculum needs. Since the lower-quality images were still higher resolution than could be displayed on the Web, a decision was made to accept the need to rescan some items later rather than delay the benefits the project would bring to students using these materials.

This was a radical quality decision in terms of the Cornell-Yale model, and was driven largely by the different purpose. The MSU project was not primarily for long-term preservation where a few months' delay was trivial. Its main goal was to serve immediate curriculum needs, which sometimes meant having to add particular titles in a week or less. A secondary preservation goal was served by reducing the handling of the original, much as preservation photocopying would do. In effect, MSU chose to do information capture, which was as full as its primary goals required, and in that sense fit the spirit, if not the letter, of the Cornell-Yale model.

Issues of Integrity

Ross Atkinson talks about the importance of the "photographic mentality"—that is, "representation of the text as is...."²⁶ Paul Conway describes the intellectual integrity of an item as a concern for the "authenticity, or truthfulness, of the informational content of an item—maintained through careful and complete reformatting or sensitive treatment...."²⁷ The emphasis on accuracy, completeness, representing the text "as is" is one of the distinguishing hallmarks of the Cornell-Yale model.

There are two major consequences. The first is that images are presented with all the flaws of the original. If the original type font was hard to read, the digital image will have type font that is hard to read. If the original has misspellings, the digital image will have them as well, in exactly the same places. The first

set of books that Cornell digitized had handwritten notes by well-known mathematicians in the margins, and these now-important defacements were reproduced as scrawled, however illegible. Absolute accuracy took priority over all of the minor improvements that a conventional publisher might make in reissuing a work. This follows logically from microfilm preservation standards, where no temptations for improvements are possible. It is also a logical outgrowth of the concern that every change strips away information that may be valuable to future scholars.

The second consequence of the Cornell-Yale stand on integrity is that it does not incorporate ASCII conversion and SGML markup to make the text searchable, though it does not oppose them either when offered as an added service. JSTOR, for example, is quite consistent with the model in providing photographically accurate digital images of every page of its journals, while allowing users to run background searches of a good (but not perfectly accurate) OCR'd copy. This highlights a fundamental difference with the SGML model, for which the improved, marked-up, and searchable version represents the primary product, and the photographically accurate digital image is a crucial, carefully preserved, but generally invisible intermediate. It may well be that improving OCR technology will soon close the errors-per-character gap between the two models. Even when that happens, however, the fundamental split between displaying an image of original or building an enhanced version will still remain.

MSU followed the Cornell-Yale model fully in presenting photographically exact digital representations of original items. This was not an easy or uncontroversial decision. Many of the most interesting items in the Radicalism Collection looked like the sort of papers that a tidy librarian would toss into the recycling bin without thinking twice. They lacked the handsome fonts and careful layout of a professionally published book. Any person with minimal word processing skills and a decent inkjet printer can produce better-looking pages today. The ugliness of the items is in fact an important part of their contents—a part that an ASCII version with SGML markup would mask.

A second reason for MSU to follow the Cornell-Yale model was the cost of providing the integrity levels acceptable to the SGML model. Good OCR software is not cheap, but far more expensive is the time involved in checking the result for errors. Low-wage undergraduate workers with proper training and the right attitude could in theory do the editing, but finding and training the right persons takes time. Spot checking continues to be necessary, and turnover among undergraduate workers tends to be high. Even with a reliable, well-trained worker, half a dozen pages

could be scanned for every one that gets OCR'd and edited. Adding the extra time and training of SGML markup meant that MSU had to choose between providing a few sample items or a broad range of what students actually need and request. The Cornell-Yale model was the only economically viable option where an acceptable level of integrity was possible.

Issues of Longevity

Longevity is a serious issue because of the impermanence of digital files. Margaret Hedstrom points out three key problems.²⁸ The first is that the recording media are vulnerable. Magnetic media such as disks and tapes are particularly susceptible. A strong magnetic field that comes too near can wipe out data, and people are often unaware what objects create magnetic fields. Also the reading process can be hard on the media. There is physical contact with the machinery, and a read-head that is improperly calibrated can cause unseen damage.

The second problem is hardware and software obsolescence. In the roughly 50 years that computers have existed, both hardware and software have mutated rapidly to offer new capabilities. The 5.25" diskettes which were state-of-the-art personal computer storage media 15 years ago are almost useless today because few machines have drives to run them. The same is true for older software formats. Few contemporary word processors will read 1983 Wordstar files. There is some basis for arguing that this rate of change is common for new technologies, and might well decrease for computers in another 50 years or so, but no one concerned with the longevity of today's information should count on such predictions.

The third problem is the lack of standards. Without nationally accepted digitizing standards it is impossible to plan for the future in the way that, for example, those using microfilm can do. A quick push for standards is tempting, but inappropriate standards can be more damaging than none at all. Patricia Battin has, for example, discussed the problems that Cornell faced in trying to meet preservation standards written in the microfilm era—standards which prevented the National Endowment for the Humanities from funding the CLASS project.²⁹

The Cornell-Yale model addresses these longevity issues forthrightly. It does not pretend that digitizing guarantees permanence. Instead the model emphasizes the importance of a commitment to the long-term and potentially expensive process of technology refreshing.³⁰ This means that files are copied periodically onto new media, and that formats and access tools get reviewed regularly and updated as systems change. Since this is not a set of tasks for which most library staff have

the necessary training or resources, Cornell University Library delegated the work to Cornell Information Technologies, the campus computing center, which routinely deals with these problems. Who does the work matters less than ensuring that it gets done in a correct and timely matter. The model's longevity goal is to have the digitally perfect great-great-grand-copies of images be useful and available long after the best microfilm has crumbled into dust.

For MSU longevity has not been a prime concern. The long-term plan has been to follow the Cornell-Yale model and institute a technology-refreshing procedure. The short-term situation is that the low-resolution image copies are stored on the library server and on back-up tapes, and the high-resolution images are stored on the hard drives of the scanning machines with back-ups on Zip disks. The latter is particularly a concern. There is general agreement that Zip disks are not a desirable long-term storage medium, both because they are proprietary systems and because they have not been tested for durability by any disinterested third party. Cornell-style CD-ROM storage has been discussed. Another possibility is tape. Tape's advantages and weaknesses as a storage medium have the virtue of being well known. It is relatively inexpensive and most formats are non-proprietary.

Before any long-term storage can be implemented, several issues remain to be worked out. One is how to arrange a logical storage sequence, especially for tape. Should storage, for example, be in processing sequence, or in shelf (call number) order, or in OCLC number order, or in some other sequence dictated by curriculum or Web or other needs? Another issue is where to keep track of the sequence information. It could be in the metadata database. It might also belong in a computer center tape management system, especially if they have a hand in both the storage and the technology refreshing. A third issue is where to store the CDs or tapes. Tapes in particular require environmental controls that the library itself cannot easily provide. MSU has resisted the temptation to jump to a solution that gives the appearance of compliance with the Cornell-Yale model. Waiting too long could be a problem, but hasty and incomplete solutions could be worse.

Issues of Access—Online

The Cornell-Yale model does not view access in a vacuum. In the words of Paul Conway, "Preservation in the digital world puts to rest any lingering notion that preservation and access are separable activities."³¹ This means that no project should be purely for access, since any form of digital access has also created a copy, and every copy is (like a preservation photocopy) a way

of saving the original. It also means that no project should focus exclusively on creating digital images to be preserved unseen in vaults, because digital files are inherently network accessible with only modest effort. In other words, if an institution takes the trouble to digitize items, it should realize that it has undertaken two commitments: one to create a preservation quality image, and the other to make that image available.

This double commitment leads to a technical problem. The Internet is not capable of moving large 600 dpi files at a speed acceptable to most users. An even bigger problem is the inability of most existing monitors to handle resolutions over 72 dpi. What happens is that the monitors enlarge the image so that an equivalent number of pixels are used over a larger surface. The result is that letters become huge and documents hard to read because constant scrolling is necessary. A preservation-quality image is simply not acceptable for online access today without modification. The Cornell-Yale model accepts this fact, and calls for derivative display images with a significantly lower resolution. The principle that Cornell developed was "readability," with an assumption that the online files would be used primarily "for rapid browsing."³² That assumption is becoming less and less true as monitors improve and as generations of students who grew up reading computer screens enter the universities, but the readability principle still holds. A readable image for a standard 23 centimeter volume whose original paper pages had good contrast and no especially fine print is in the neighborhood of 100 dpi. That is the resolution that JSTOR prefers. Yale in 1992 proposed 200 dpi for its images from microfilm.³³ The ideal display resolution depends on circumstances, not on a universally applicable number of dpi.

Access in the SGML model has a very different set of considerations. It escapes the need to produce lower-resolution derivatives, because the OCR'd or transcribed ASCII images present no serious size problems. Instead it faces problems with ensuring accuracy and with either providing an expensive SGML browser or converting the SGML tag-set into HTML for standard Web-browser display. The cost of creating a Web-accessible version from the scanned image is also significantly higher, both because of the OCR or transcription process, and because a quality SGML markup job takes time and well-trained people. The result is a representation that is not identical to the original, but has the added value of being machine-searchable in sophisticated ways, such as looking for information only in first lines or only in footnotes. It is on this issue of access that the two models most diverge.

MSU follows the Cornell-Yale model for online access. The goal is to produce not the best looking

INSTALLING OFFICER: "Sister Marshal, you will read aloud the names of the newly elected officers to the Klanswomen now assembled."

(After concluding this the Marshal asks all to rise for prayer which shall be given by the Marshal.)

Prayer to be as follows:

"Almighty God, we commit to Thee these women who have been elected to fill offices of this Klan and ask that Thou wouldst fill them with wisdom and grace, that their every effort may be in tune with Thy will and look toward the good of this Great Order. Give them, we beseech Thee, the dignity and devotion that should accompany their new responsibility and also teach them to be impartial in every ruling they may be called upon to make. Give them the courage to set the proper example for all Klanswomen, that by their lives they may emulate that which is according to Thy wishes. Oh, God, we ask these things for the good of our Order and for the Glory of Thy great Name, Amen."

Figure 2: A sample image from "Installation Ceremonies" by the Women of the Ku Klux Klan, page 6.

display, but the best balance between size (in bytes), width (on the screen), and readability (see figures 2 and 3). The assumption is that students will be accessing images through the phone lines with relatively slow modems, and that they will balk at using them if 1) the files are so large that the download seems painfully slow or 2) they must scroll back and forth constantly to reach each line of print. The scrolling problem is particularly important because many Radicalism Collection items are on 8.5" by 11" paper that is wider than a standard book.

Quarterdeck's Hijaak Pro utility is used to do batch conversions of TIFF files to GIF format. The original instructions left images in black and white and just adjusted the resolution downward to fit the screen. Recently staff have discovered how to make Hijaak trade resolution for depth. The result is grayscale images of roughly 100 dpi, which are sharper than the black and white images and take less screen space. Doing this involves using options-settings of 400 dpi, grayscale, interlacing, and a three-inch frame (the result is significantly larger). Staff are instructed to make test conversions of one or two files from the middle of a work to be sure the results are readable. All images for a title are then converted to the same resolution.

This testing process is expensive. For a pamphlet with half a dozen pages it can almost double the number of conversions, and occasionally the readability judgment of staff in their early twenties does not match that of bifocal-wearing supervisors.

MSU is considering switching to *tif2gif*, a software tool which converts TIFF files on the fly to screen-size grayscale images.³⁴ The Making of America project has used this software successfully. One advantage is a saving on storage space for the GIF images, since they exist only temporarily. Another is the ability to convert the image at 25 percent, 50 percent, 75 percent, or 100 percent of the original size, so that users can look at the whole layout of a large page and can adjust the image to a size that is comfortable for their eyes. Binaries for SunOS, Solaris, and AIX, as well as source code are available on the *tif2gif* Web site.

Issues of Access—Organization

One effect of digitizing a document is to dismember it—intellectually if not physically. Each page image becomes a separate file with no fixed relationship to the others, and new virtual bindings are necessary to put the images in logical sequence. As Paul Conway points out, even a powerful index that links a list of terms and concepts with particular pages does not substitute for understanding the "structure, relationship, and intellectual contents" of a whole document.³⁵ In the Cornell-Yale model the tool that recreates an item's organization is a set of separate online documents, which describes the structure of the work.

Cornell and Yale use somewhat different approaches. Cornell takes a linear approach in presenting page numbers sequentially with bookmark-style tags that show where forewords, chapters, indexes, and other sections begin. Yale builds an archival type of "finding-aid" that emphasizes the structural elements of the work, such as the front matter, the contents, and the body. Both seek to reproduce the organization of the original in a way that is both faithful and still convenient for online users. Yale now uses SGML to build these finding aids, and this use of SGML may well grow to be more important for those following the Cornell-Yale model.

MSU currently builds organizational structures that follow Cornell's linear practice for pamphlets, leaflets, and other very short materials. For book-length works it uses the Yale finding aid model with its distinctions between front matter, body, and back matter. Since the MSU project is aimed at freshmen, an effort is made to match the organization to their needs and understanding. No single standard is used. Each work is judged separately depending on its inter-

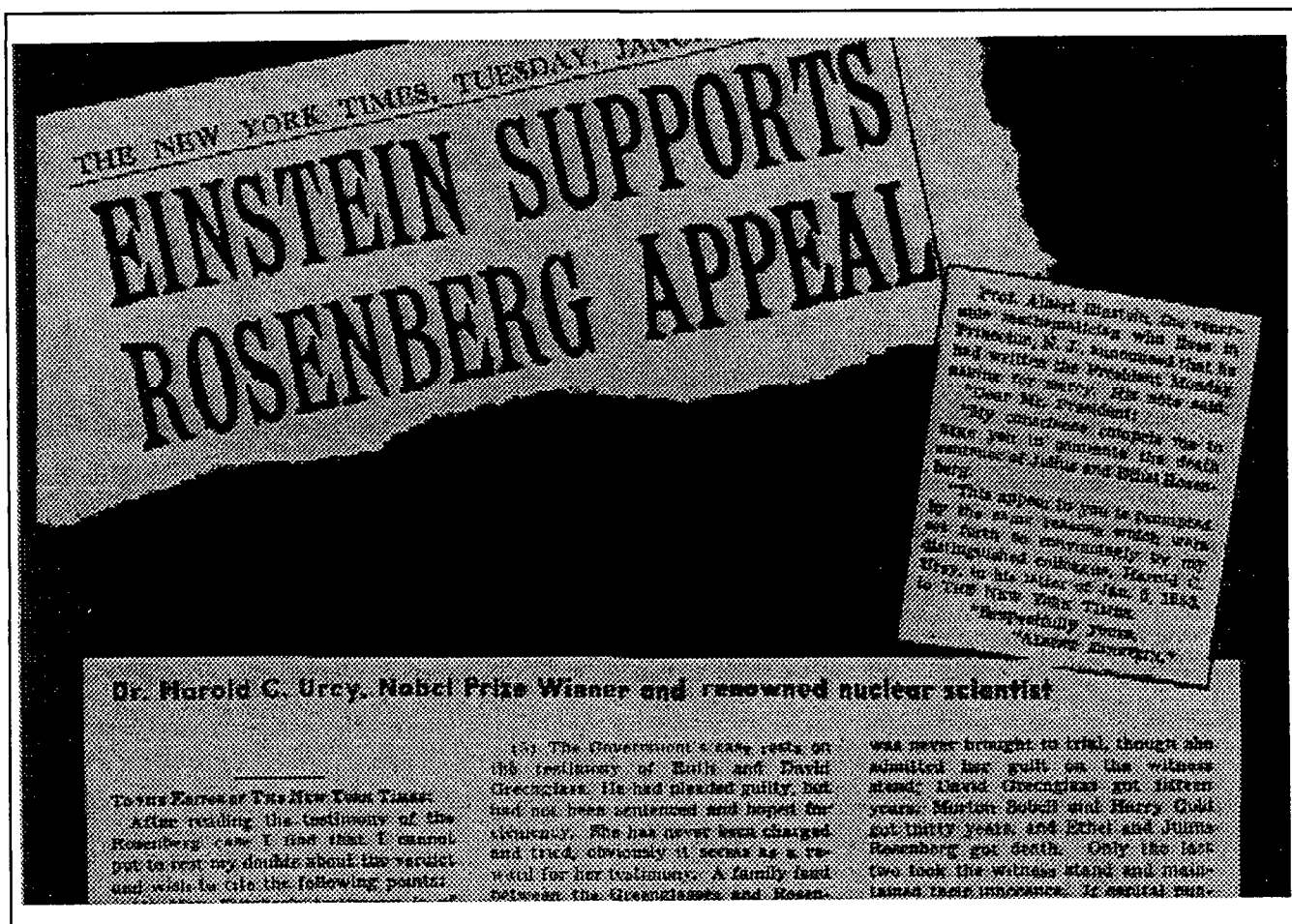


Figure 3: A sample image from "Broadsides" by the National Committee to Secure Justice for the Rosenbergs, page 10.

nal structure or lack thereof. It is important to remember that the Radicalism Collection has a wider range of types of materials than the collections that either Cornell or Yale has digitized. Full-length books of several hundred pages are only a small part of the total. As experience with this variety of materials grows, a set of principles for building appropriate organizational structures should emerge and be codified.

The organization of the digital collection as a whole is as important to the success of MSU's project as the organization of each work. MSU's current OPAC does not allow hyperlinked URLs that point directly at each item, nor would this really be adequate for the undergraduates, who need advice about which titles would go together well to make good papers. At present, two services are provided. One organizes the digitized items alphabetically by title. This is for students who first searched the OPAC and are looking for particular items. The other organizes the titles by course-paper topic (see figure 4). Titles that are appropriate for more than one topic are listed twice. It is, in effect, an online bibliography designed specifically for students in certain classes. The expectation

is that this topical guide will become the chief form of access, and that the list of topics will continue to grow as more and more faculty make requests.

Issues of Access—Printing

Another important part of access for the Cornell-Yale model is the ability to produce printed copies. When the Cornell project began in 1990, few faculty or students had machines powerful enough to display even low-resolution images decently. The project managers expected that users who wanted digitized works mostly would prefer printed copies, which could be done easily and quickly using the Xerox DocuTech printer. In today's world many people continue to prefer printed copies for anything long or difficult to read or for simultaneous browsing with several other documents. But the model specifies no preferred method for providing this print capability. Online on-demand printing is practical for relatively short journal articles from a service like JSTOR. For book-length works, there is no desktop substitute for a fast, high-

Rosenberg Case [back to top](#)

Title	Author	Call Number
<u>Appeal For Clemency, An</u>	Pritt, D.N. (Denis Nowell), 1887-	KF224.R67 P7
[<u>Broadsides</u>]	National Committee to Secure Justice in the Rosenberg Case	KF244.R67 N3
<u>Mercy for the Rosenbergs</u>	Abraham Cronbach	KF224.R67 C7 1952
<u>"Never losing faith ... " : for Julius and Ethel Rosenberg.</u>	various	HX86.N4 1953
<u>Rosenberg Case, The</u>	Pritt, D.N. (Denis Nowell), 1887-	KF224.R6 P7 1953
<u>To Secure Justice in the Rosenberg Case</u>	William Reuben	KF224.R6 R46 1951
<u>Suppressed Facts in the Rosenberg Case, The</u>	Irwin Edelman	HX R6 E3 1953

Sacco-Vanzetti Case [back to top](#)

Title	Author	Call Number
<u>Background of the Plymouth Trial</u>	Bartolomeo Vanzetti	KF224.S2 V3 1926
<u>Case of Sacco and Vanzetti in Cartoons from The Daily Worker, The</u>	Fred Ellis	NC1429.E4 C3 1927
<u>Conspiracy Against Sacco and Vanzetti</u>	various	KF224.S2 S32

Figure 4: Part of the Subject List

quality printer like the DocuTech, which collates and provides a perfect binding as part of its output.

Printing has been a major access concern at MSU, and has been addressed in two ways. The first is to recognize the image-printing capabilities of the Web browsers. This has some problems. Although a 100 dpi image will usually fit on an 8.5" wide page, the print may suffer from strange artifacts, such as decapitated lines whose bottom half appears on the following page. But the result is at least generally useable, if not handsome. Printing convenience is a reason for keeping the images as small (in inches) as possible. With higher resolutions, the lines can become so broad that the right side is truncated and lost completely. Switching to landscape printing helps, but it is one more step users must take, and even landscape has its limits. Since each image is in a separate HTML document, using the browser print capabilities also means that students must print page by page. For anything more than a few pages, this quickly becomes impractical.

The second print option is to put high-quality printed copies of everything that has been digitized onto course reserve. This parallels Cornell's practice of replacing digitized books with new paper copies. An explanation of where to find these copies is on the introductory page of the Radicalism Collection Web

site, and the reserve reading desk is conveniently near the main library's copy center, so that students who want to make photocopies can do so easily. This is a real marriage of traditional and digital methods. Not only does it offer students a choice, but the online copies can function as an extended finding aid to let students decide which articles they want to copy, or to allow them to locate last-minute references without trudging across campus.

Other print options have been considered, but all required users to install additional modules on their computers. This could disenfranchise less technically oriented students, or force them to come to the library to use machines with the print options already set up. Better solutions are being sought to allow students to do quality printing simply and easily from home. MSU is well aware that the SGML model does not have this problem, since the text is already in ASCII and will make a clean, professional-looking print. But the printing issue has not become so important as to suggest a change of models.

Issues of Access—Usage Studies

Usage studies are not a canonical part of the Cornell-Yale model. Those institutions have focused

SIDEBAR 4: SURVEY

Radicalism Collection Digital Images Feedback Form

Introduction: The University Library has provided Web-based digital images of works in Special Collections' Radicalism Collection so that you will have easier access to them. We would like your feedback to make it better.

Background: Please circle the best answer or fill in the blank.

Sex: Male / Female

Age: _____

Years of computer experience: _____ Have you easy access to a computer? Yes / No

How convenient is it for you to get to the Main Library between 9:00 a.m. and 5:00 p.m.?

very convenient 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 very inconvenient

Non-Users Only: If you made *no* use of the digital images, please circle the main reason:

- 1) Poor access to the Web
- 2) Do not like reading on the Web.
- 4) What I wanted was not on the Web.
- 5) Other (please explain on the back).

Users Only: Please mark or circle the best answers in this section.

Did you use the reserve copies in addition to the digital images? Yes / No

How easy was it to find the digital images you wanted?

very easy 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 very hard

How were the images to read?

easy to read 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 hard to read

If you printed copies from the Web, how easy was that to do?

very easy 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 very hard

If you printed copies from the Web, how were the copies to read?

easy to read 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 hard to read

What would you most like to see improved? (Please rank 1 through 4)

- | | |
|---|------------------------------------|
| ___ Number of items available on the Web. | ___ Ease of reading on the screen. |
| ___ Ease of printing. | ___ Quality of printed copies. |

Please add any other comments you have on the back.

on items for scholarly use, and have no captive and easily accessible population from which to get feedback. MSU does, however, have a captive classroom population available and plans to build regular usage studies into its digitizing program. The goal is to learn more about how its target population uses the images and what kinds of added-value features they most want.

To prepare for this, MSU looked at what others had done. There appear to have been relatively few user-studies for digital libraries, but one possible model comes from Columbia University, which did a study of the role of online books for academic library users.³⁶ The author, Mary Summerfield, interviewed 23 faculty and five graduate students, and asked them about background information, computer access, hours of comput-

er use, ranking of expertise, how many books and journals they read, and what percentage of the whole they actually looked at. The study focuses on what people would like in the abstract, rather than at the strengths and weaknesses of a particular digital collection. It also focuses heavily on senior researchers (full professors are 46 percent of the group) and on political scientists (39 percent). Such a sample is certainly not representative of undergraduates at an institution like MSU. Nonetheless it offered a starting point for thinking about what the survey should accomplish.

MSU's proposed initial survey (see sidebar 4) has the practical goal of learning whether undergraduate users require potentially costly improvements in the ease and quality of printing. Closely related to printing is the issue of how comfortable the students feel reading the images online. For those who never used the online images, the survey tries to find out why. Some basic background information is also asked, including how convenient students feel it is to get to the main library during Special Collections' hours, and how conveniently they can access a computer. The results will help inform future developments. Succeeding surveys may explore how students use the finding aids, and what kinds of new materials they would like to see in digital form.

These surveys also are potentially important to a state-supported institution like MSU, which must always be ready to justify the cost of its digitization program. The surveys not only show a (genuine) concern for the undergraduate user population, but allow MSU to evolve the service in ways that will build a strong and enthusiastic support. That sort of support matters. David Seaman of the Etext Center at the University of Virginia writes that in his experience a "stable, local user community, once created, becomes a great asset... [I]t prevents us from forgetting that we are a local service with a defined mission..."³⁷

Conclusion

This article has two purposes. One is to summarize the elements of the Cornell-Yale digitization model in a way that would be useful to other institutions planning to follow their efforts. The people involved with the Cornell and Yale projects have done an exceptional job in documenting their work. They have been the Lewis and Clarks of the digitization process. But because they were the first to explore this unknown territory, not every facet of their experience applies to those who follow the path they blazed. It is necessary to see which parts of their experience are applicable to a large state university like MSU. The elements that have been laid out here are not complete. For example, the economic issues have not been discussed at all, and deserve

separate and extensive treatment. It is also important to realize that other models exist, particularly the SGML model, which has different goals and uses some different technologies. MSU did not choose the Cornell-Yale model because it was inherently better—but because that model fit the needs, resources, and timetable of the Radicalism Collection project.

The other purpose of this article is to show how MSU's experience has expanded the Cornell-Yale model. In this MSU is following their tradition of documenting features that could be useful and relevant to others. Three conclusions can be drawn:

- 1) The Minolta PS3000 is currently a useful scanner for making preservation-quality images of items with large fonts and no grayscale images because of its speed and the ease with which materials can be positioned, and it should become even more useful once it has 600 dpi and true grayscale capabilities.
- 2) Tying selection to the undergraduate curriculum does not violate the Cornell-Yale model's principles of coherence and scope.
- 3) Designing an access system for undergraduate use can be consistent with and even add value to the electronic finding aids and printed copies of the Cornell-Yale model.

Although many institutions are likely to show interest in the overhead scanner, MSU's link with the undergraduate curriculum has broader implications. For many public universities, curriculum support is the primary mission. They cannot justify spending money on the resources needed for digitization if it only serves long-term preservation needs for scholarly use. MSU's experience offers a justification more compatible with their teaching mission. It is an encouragement for a broad range of institutions to try digitization for themselves.

NOTES

1. Many papers are available, among them Anne R. Kenney and Lynne K. Personius, "The Cornell/Xerox/Commission on Preservation and Access Joint Study in Digital Preservation" (Washington, DC: Commission on Preservation and Access, 1992), <palimpsest.stanford.edu/cpa/reports/joint/products.html>.
2. Paul Conway, *Preservation in the Digital World* (Washington, DC: Commission on Preservation and Access, March 1996), 16.

3. Kenney and Personius, <palimpsest.stanford.edu/cpa/reports/joint/products.html> .
4. Kenney and Personius, <palimpsest.stanford.edu/cpa/reports/joint/products.html> .
5. Kenney and Personius, <palimpsest.stanford.edu/cpa/reports/joint/products.html> .
6. Paul Conway and Shari Weaver, "The Setup Phase of Project Open Book: A Report to the Commission on Preservation and Access on the Status of an Effort to Convert Microfilm to Digital Imagery" (Washington, DC: Commission on Preservation and Access, June 1994), <palimpsest.stanford.edu/cpa/reports/conway.html> .
7. Donald Waters and Shari Weaver, "The Organization Phase of Project Open Book: A Report to the Commission on Preservation and Access" (Washington, DC: Commission on Preservation and Access, September 1992), <palimpsest.stanford.edu/cpa/reports/openbook.html> .
8. Paul Conway, "Selecting Microfilm for Digital Preservation: A Case Study from Project Open Book," *Library Resources and Technical Services* 40 (January 1996): 68-69.
9. 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, 1996), 18-19.
10. Paul Conway, "Yale University Library's Project Open Book: Preliminary Research Findings," *D-Lib Magazine* (February 1996), <<http://www.dlib.org/dlib/february96/Yale/02conway.html>> .
11. "The JSTOR Production Process," (New York: JSTOR, 1996), <<http://www.jstor.org/about/production.html>> .
12. MSU's American Radicalism Collection contains materials on Timothy Leary, the Black Panthers, the Christian Right, Neo-Nazis, the Ku Klux Klan, the Communist Party of the U.S.A., the Students for a Democratic Society, and the Vietnam War era. For more information about it, see <http://www.lib.msu.edu/coll/main/spec_col/radicalism/aboutrad.htm> .
13. Lou Sharpe, "Preservation Quality Scanning of Bound Volumes: Integration of the Picture Elements ISE Board with the Minolta PS-3000 Book Scanner," *RLG DigiNews* 1:1 (15 April 1997), <<http://www.rlg.org/preserv/diginews/>> .
14. Conway, "Selecting Microfilm," 67.
15. Conway, "Selecting Microfilm," 69, 71.
16. Conway, "Selecting Microfilm," 72.
17. Margaret S. Child, "Selection for Preservation," in *Advances in Preservation and Access*, vol. 1 (Westport, CT: Meckler, 1992), 150.
18. Samuel Demas, "Setting Preservation Priorities at Mann Library: A Disciplinary Approach," *Library Hi Tech* 12:3 (1994): 83.
19. Demas, 85.
20. Robert Oakley, "Copyright and Preservation: A Serious Problem in Need of a Thoughtful Solution," (Washington, DC: Commission on Preservation and Access, September 1990), <<http://www-cpa.stanford.edu/cpa/reports/oakley/index.html#toc>> , section A. "Is the Work Protected."
21. Oakley, section A.3 "Unpublished works."
22. Stephen Chapman and Anne R. Kenney, "Digital Conversion of Research Library Materials," *D-Lib Magazine* (October 1996), section on "Full Information Capture," <<http://www.dlib.org/dlib/october96/cornell/10chapman.html>> .
23. Anne R. Kenney and Stephen Chapman, "Bitonal Scanning Means for Benchmarking Resolution Requirements," in *Digital Imaging for Libraries and Archives* (Ithaca, NY: Department of Preservation and Conservation, Cornell University Library, 1996), 9.
24. Paul Conway and Shari Weaver, "The Setup Phase of Project Open Book," (Washington, DC: Commission on Preservation and Access, 1994), <<http://palimpsest.stanford.edu/cpa/reports/conway.html>> .
25. David Seaman, "Special Collections Digital Image Creation" (University of Virginia, June 1996), <<http://etext.lib.virginia.edu/helpsheets/specscan.html>> .
26. Thomas F.R. Clareson and Miriam Kahn, "ALCTS Preconference Reports—Electronic Technologies: New Options for Preservation," *ALCTS Newsletter* 4:6/7 (1993): 78.
27. Conway, *Preservation in the Digital World*, 9.
28. Margaret Hedstrom, "Digital Preservation: A Time Bomb for Digital Libraries" (University of Kentucky, [1996?]), <<http://www.uky.edu/~kiernan/DL/hedstrom.html>> .
29. Patricia Battin, "Image Standards and Implications for Preservation," a talk presented at the Workshop on Electronic Texts, sponsored by the Library of Congress, 9-10 June 1992, <<http://palimpsest.stanford.edu/byauth/battin/imagestd.html>> .

30. Kenney and Personius, <<http://www-cpa.stanford.edu/cpa/reports/joint/>> .
31. Conway, *Preservation in the Digital World*, 6.
32. Kenney and Personius, <<http://www-cpa.stanford.edu/cpa/reports/joint/>> .
33. Donald Waters and Shari Weaver, "The Organizational Phase of Project Open Books" (Washington, DC: Commission on Preservation and Access, 1992), <<http://palimpsest.stanford.edu/cpa/reports/openbook.html>> .
34. "TIF2GIF" (Ann Arbor, MI: University of Michigan, [n.d.]), <<http://kalex.engin.umich.edu/tif2gif/>> .
35. Conway, *Preservation in the Digital World*, 15.
36. Mary Summerville, "Columbia Online Books Evaluation Project: Online Books: What Roles Will They Fill for Users of the Academic Library?" (Columbia University, 1995), <<http://www.columbia.edu/cu/libraries/digital/texts/paper/>> , last revision 23 May 1995.
37. David Seaman, "The User Community as Responsibility and Resource: Building a Sustainable Digital Library," *D-Lib Magazine* (July/August 1997), <<http://www.dlib.org/dlib/july97/07seaman.html>> .

For an annotated bibliography of sources on digitization, please see Michael Seadle, "Digitization of the Masses," *Reference Services Review* 25:3-4 (Fall-Winter 1997): 119-130.

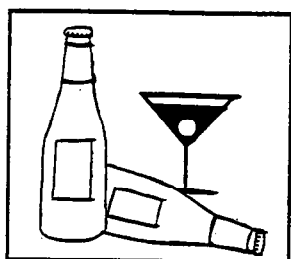
Available in June 1998

Alcoholism

The Health and Social Consequences of Alcohol Use

An Annotated Bibliography and Analytical Introduction

by Cecilia M. Schmitz and Richard A. Gray



SCIENCE AND SOCIAL RESPONSIBILITY SERIES NO. 3

ISBN 0-87650-355-5, 6 x 9, 462 PAGES, \$40.00

THE PIERIAN PRESS, BOX 1808, ANN ARBOR, MICHIGAN 48106

☎ (800) 678-2435 ☎ Fax (734) 434-6409