Digital Imaging and Preservation
Oversize Color Illustrations

Janet Gertz
Director for Preservation, Columbia University Libraries
110 Butler Library, 535 West 114th Street, New York, NY 10027

Abstract

Digital imaging can play an important role in helping to preserve brittle library and archival materials, especially those characterized by color and oversize illustrations, neither of which can be satisfactorily preserved by analog methods. Used together, film intermediaries can provide long-term stable copies while digital versions can provide improved scholarly access. Setting benchmarks for high-quality color images is an important aspect of this approach, and the paper describes the results of a project carried out by Columbia University Libraries to use digital imaging in combination with single-frame color microfiche to preserve oversize brittle maps.

Analog and digital preservation: the hybrid approach

The point from which I come to digital imaging is as the preservation officer of a large academic research library hoping to employ digitization to copy brittle, damaged materials and make them more useful to scholars. Our collections contain millions of printed books, manuscripts, and archival materials. We are especially rich in collections that include graphic materials: architectural drawings, photographs, maps, and print collections in art history and other heavily illustrated disciplines. My job is to preserve the content of all of these media, but as I will discuss later, I have recently been concerned particularly with oversize illustrative materials.

The preservation ideal is to stabilize and repair the damaged item itself. Unfortunately, in the case of brittle paper items which exist in multiples — modern published books and archival records are prime examples — we must often settle instead for creating stable copies which are as permanent as we can make them.[1] These copies must have two characteristics. First, they must be long lasting. Preservation microfilm, for instance, has a life expectancy of 500 years. Second, scholars must be able to use the copies for most if not all the purposes they would have employed the originals. A long-lasting copy that can not be used is not much improvement over a brittle original that crumbles as the page is turned. This holds as well for digital images. The potential for enhancing materials through digitization is almost unlimited, but if what is digitized is illegible, or so badly indexed that readers cannot find what they need, then we have provided neither preservation nor access.

The validity of digitization as a source of long-term preservation is very much an open question. Digital storage media have a short life relative to microfilm, and even relative to acid paper, which can last 50 to 100 years when carefully handled. In contrast, disks and tapes last twenty to thirty years at best. More troubling, software and hardware change with almost frightening speed and leave older iterations behind, their files all too often unreadable. We do not know whether digital files will change under repeated refreshment and migration through generations of software and hardware.[2] Especially when the original books or manuscripts will not survive long after scanning, we cannot afford to rely on a digital copy as our only record of what the item was.

At least for the present, therefore, digital and analog preservation need to go hand-in-hand. For true long-term security, we must assure that we have the most durable analog version we can achieve: the original item itself, properly repaired, or a copy on film housed in permanent archival storage. Complementing the analog version we then create a "digital preservation version" by scanning at the highest resolution needed to produce full legibility, with grayscale or color (as appropriate) carefully matched to the original; and saving it in a lossless format. No enhancement is carried out, since the digital preservation version's role is to present an accurate record of the original for scholarly purposes. Certainly digitization offers tools to enhance images — a manuscript with text obscured by coffee stains can be digitally altered to be more legible, but this changes the facts of what the original really was at the time of scanning. Authenticity and accuracy in representing the original are particularly at issue if the original will be discarded after scanning.

From the digital preservation version we can then derive as many copies as desired, and it is these use copies that can be enhanced and manipulated at will. The point is to open up all the possibilities of use without compromising the authenticity of the digital preservation version; and to maintain the analog version for those who will need to consult it, and of course in case of accidental loss or change to the digital preservation version. Should it ever be needed or desired, we will have the analog version to re-scan.

[1] Permanence is of course a relative concept. See for instance James O'Toole, "On the Idea of Permanence," *American Archivist* 1989, pp.10-25.

[2] Cf. the recent discussion by Jeff Rothenberg, "Ensuring the Longevity of Digital Documents," *Scientific American* 1995, pp.42-47.

This model, the "hybrid approach" as it has been called,[3] assumes that it is in fact possible to make a digital version of the original, that digitization is appropriate for preserving that item, and that the digital copy is good enough to serve the needs that justified selecting the item for preservation in the first place. Much of the current work in digital applications for preservation is aimed at determining whether the types of access which scholars require and desire can be provided by a digital version, and what level of quality is needed to capture at least as much information (preferably more) as the traditional methodologies.

Image quality issues

How do we decide how good is good enough? How can we set definitions for appropriate image quality? Capture of graphic materials is complex, encompassing continuous tone, halftone, or color illustrations, all mixed with black and white text. Not only do books include illustrations, but words and numbers are found within graphic media, for instance on architectural drawings or maps. In most of these cases legibility of the textual elements alone is far from sufficient to define an adequate copy.

Some experts advocate capturing all materials at the highest technically possible resolution and pixel depth (i.e., dynamic range) as the ideal preservation digitization goal, regardless of the nature of the originals, in order to assure that all potential uses will be met and to assure that a second, better scan will not be needed in future. But the higher the quality of capture, the larger the file, and the higher the cost of capture, retrieval and manipulation, and storage media. Funds are always limited, and going beyond the quality reasonably needed consumes money which could otherwise be employed in preserving further items.

Rather than always aiming for the highest technically possible quality, it is essential to determine what the upper limits of quality are that will fully meet foreseeable needs, including different uses of the same image by different disciplines. A historian concentrating on dating a manuscript may well want any stains preserved for the historical evidence they convey. Another scholar may prefer to see what is underneath the stain. A geologist concerned with coding on a map needs legibility and distinguishable colors, while an art historian concerned with the aesthetics of the map needs color faithful to the original. All of these variations imply different requirements for the digitization of the original items.

[3] Don Willis, *A Hybrid Systems Approach to Preservation of Printed Materials*, Washington DC, Commission on Preservation and Access, 1992

The definition of a successful digital imaging project will at minimum require decisions on pixel depth: whether to capture information simply in black or white (binary), grayscale (256 gradations captured using 8 bits per pixel), or color (normally using 24-bits per pixel to give 16 million different potential colors). While a binary scan of 600 dots per inch is clearly superior to one at 300 dpi, a binary scan at 600 dpi is not so simply compared to a 300 dpi grayscale scan. The 8 bits of grayscale convey additional information not carried in the binary version, and color adds even more.[4] Since file size increases immensely for color, it is not feasible to simply scan everything at high resolution and 24-bit color, despite the potential for great accuracy of reproduction when color is properly handled.[5]

Substantial work has already been done on the level of quality required to capture the content of black and white printed text accurately. The Cornell University Preservation Department has been studying published 19-20th century volumes and has established that a 600 dpi binary scan will accurately and legibly capture black print on light background down to 4-point type, the size of the smallest picture captions and footnotes (where the lower case letter e is approximately 1 mm high).[6] The Yale University Preservation Department's Project Open Book investigates the level of quality required to capture black-and-white text from preservation-quality microfilm. Again a 600 dpi binary scan appears to be the highest resolution needed for accurate capture.[7]

Based on this work, Cornell is testing a series of benchmarks to predict what resolution is needed for materials of different sorts based on the size of the smallest letter which must be legible.[8] They have devised formulae for binary and grayscale to determine how many dots per inch are required to produce resolution equivalent to that of microfilm at quality index level 8. Quality index is a measurement of the degree of resolution provided on microfilm, and national standards require that the third generation microfilm (the service copy, the copy people actually read) must have a QI of 8.[9] Cornell's benchmarks should help assure an equivalent level of quality for digitized text, as well as allowing quick

[4] Cf. the discussion in Michael Ester, "Image Quality and Viewer Perception", *Visual Resources*, 1991 pp.51-63

[5] A chart of file sizes for binary, grayscale, and color images is provided in Peter Robinson, *The Digitization of Primary Textual Sources*, Oxford: Oxford University Office for Humanities Communication, 1993, pp.11-12.

[6] Anne Kenney, "Digital-to-Microfilm Conversion: An Interim Preservation Solution," *Library Resources & Technical Services* 1993, pp.380-402; and 1994, pp.87-95, especially pp.88-89.

[7] Paul Conway and Shari Weaver, *The Setup Phase of Project Open Book*, Washington, DC: Commission on Preservation and Access, 1994.

[8] Anne Kenney and Stephen Chapman, *Digital Resolution Requirements for Replacing Text-Based Material: Methods for Benchmarking Image Quality*, Washington, DC: Commission on Preservation and Access, 1995.

[9] *RLG Preservation Microfilming Handbook*, Mountain View, CA: Research Libraries Group, 1992, p.41.

determination of appropriate resolution without time-consuming experimentation on every item to be scanned. It should be possible to group materials and match them to pre-set benchmarks against which the success or failure of the scanning project can be measured.

Digital preservation of oversize illustrations

A great deal of research remains to be done for manuscripts, photographs, and color media, where digital imaging can potentially be even more important for preservation than for printed text. The preservation of brittle materials characterized by text plus color and oversize illustrations, especially art and architecture publications and geology and geography with their maps and charts, presents particular challenges. Scholars need to view the illustration as a whole, to follow information across the breath of the surface, but must also be able to read the finest details at every point. Color and pattern are important for aesthetic reasons and as coding devices on maps and charts. Further, the juxtaposition of the illustrations with the narrative that describes and comments upon them is essential.[10] This means that the technology chosen for preservation must be able to create a long-lasting copy which preserves the information content of both the text and the illustrations in a single, easily-used medium.

Traditional microfilming is highly unsatisfactory as a preservation option because too much visual information is lost when the color is lost. Black-and-white continuous tone illustrations also reproduce badly on microfilm. Add to this the necessity to film oversize illustrations in sections in order to keep them legible, and the result is a major loss of functionality; sectioned illustrations can be genuinely unusable. Even torn and crumbling original illustrations can be preferable to such poor reproductions.

As stated succinctly in the report of the Scholarly Resources in Art History seminar convened by the Commission on Preservation and Access in 1988, "1) scholarship in art history is dependent upon images; 2) the current preservation process of high contrast black and white microfilm is not satisfactory for the reproduction of halftone and continuous tone images; and 3) the preservation process must result in enhanced access to the scholarly resources."[11] Finding new methods to preserve text plus illustration (and illustration plus text) thus ranks very high among preservation priorities. Digital imaging is an obvious candidate.

[10] Cf. *Scholarly Resources in Art History: Issues in Preservation. Report of the Seminar, Spring Hill, Wayzata, Minnesota, September 29-October 1, 1988,* Washington, DC: Commission on Preservation and Access, 1989; and Joint Task Force on Text and Image, *Preserving the Illustrated Text,* Washington, DC: Commission on Preservation and Access, 1992.

[11] Scholarly Resources in Art History, pp.2-3.

The Columbia project

In 1994 the Columbia University Libraries Preservation Division undertook a two-part project funded by a contract from the Commission on Preservation and Access. First we wished to test the hybrid approach for illustrated materials by scanning microfilm and microfiche, and, second, to experiment with a means for integrating the digital files of text and illustrations. The goal is a digital volume in which the scholar can view the text and the illustrations together as was possible with the original paper volume.

Moving through film intermediaries to digital images was our preferred option. We wanted an archival-quality film version for long-term preservation and for those scholars who still face limitations in their ability to access digital files, especially the very large files needed for pictorial materials. They may still prefer to access the volumes in the film version. We also wished to avoid handling the materials twice, once for filming and once for scanning, since they were brittle and fragile. Because most of the cost lies in selecting, cataloging, and preparing the item to be copied, and in physically manipulating the item during capture, we wanted to pay for these operations only once and then transfer the captured information among different media and technologies. Thus, an important question for the project was whether scanning a film intermediary would produce satisfactory results.

We used two types of film intermediaries, single-frame color microfiche and 4"x5" color transparencies.[12] The single-frame microfiche employ the entire field (normally 105x145 mm) to capture a single image at a low reduction ratio. A previous project had proved that single-frame color microfiche could successfully capture and preserve the content of large illustrations in fine detail.[13]

Phase I of the project, now completed, compared direct digital scans of five original maps from the *New York State Museum Bulletin*, scans of the single-frame color microfiche of those maps made in the earlier project, and scans of color transparencies made from the same originals.[14] The Phase I final

[12] The microfiche were produced using Ilfachrome/Cibachrome color film, the transparencies with Ektachrome 100 Plus. The Image Permanence Institute of the Rochester Institute of Technology is currently more than half way through a two-year project sponsored by the State of New York, "Isoperms for Color Photography," which will provide data to establish the life expectancy of color photographic film in optimum storage conditions. Cibachrome/Ilfachrome microfiche produced and stored according to national standards is already known to have at least a 100-200 year life expectancy.

[13] The Commission on Preservation and Access in 1991/92 sponsored a project designed to experiment with preserving twenty-eight brittle numbers of the *New York State Museum Bulletin* by a variety of techniques. One activity of the project was to create single-frame color microfiche. See Susan Klimley, "Notes from the Cutting Edge," *Microform Review* 1993, pp.105-107.

[14] The *Museum Bulletin* contains text, black-and-white and colored plates, and colored oversize maps as large as

report is available from the Commission in paper form and is also available on the Internet along with over 300 map images; the URL is http//www.columbia.edu/imaging/html/largemaps/oversized.html.

In order to evaluate the images, we had to define the desired level of quality. We were not seeking to achieve the best technically possible quality image, but rather, the image good enough to serve the needs of researchers in lieu of the original paper maps. We defined two levels of success, adequate and full. Our definition of adequate quality stated 1) that all of the textual details had to be legible although they might be somewhat fuzzy, including the smallest print, which happened to be 1 mm high contour elevation numbers; and 2) that all color codes had to remain distinct, even if colors shifted in comparison to the printed originals.

The definition of fully successful quality stated that all print, including the 1 mm contour elevations, must attain full clarity; that all color codes must remain distinct; and that there must be no color shift in comparison to the printed originals. In other words, we put primary emphasis on the ability to read the content of the map, including the color codes. Color fidelity to the original was secondary although desirable. Obviously, definitions of successful capture would differ significantly for art materials and other disciplines where color fidelity is essential.

During the project the original maps, the microfiche, and the transparencies were scanned by a number of vendors at a variety of resolutions. Vendors used a combination of digital camera, flatbed scanner, and drum scanner. All scanning was done at 24 bit color. The original maps were scanned along with standard color bars to permit later color balancing and other manipulation.

The digital images created by the vendors met the definition for adequate success and partially met the definition of full success. When the microfiche were scanned at the same effective resolution as the original maps, the images produced were just as legible as those of the scanned originals. We found that scanning at a resolution of 200 dpi and a pixel depth of 24-bit color produced full legibility of 1 mm-high type on the original map. This means that a map 20 inches across requires 20 inches x 200 dots or 4,000 dots across its surface in order to reproduce the 1 mm type at full legibility when scanning with 24-bit color. The microfiche of the same map also requires 4,000 dots across the surface of the map image to capture the same degree of detail; but in this case the microfiche image of the map is perhaps only 4 inches wide, so that we are talking about 1,000 dots per inch on the microfiche.

These findings are in line with Cornell's proposed benchmarks for resolution for black-and-white and grayscale materials.[15] Following the logic of their formulae for those pixel depths, a formula for

30"x40". Turn-of-the-century volumes contain brittle folded plates which have been stored in pockets for decades and cannot be opened flat without breaking; many have torn on the fold lines when opened or have been damaged when refolded and re-inserted in the pocket.

color scanning (24 bit) suggests approximately 200 dpi for quality index level 8, confirming Columbia's evaluation of the images.

While resolution was handled satisfactorily, questions remain about the quality of the color. The project did not achieve color fidelity in scanning from either the originals or the microfiche, although both produced images in which the color codes remained distinct. Color shifts would be expected when scanning the microfiche and transparencies, since change is introduced by every aspect of the photographic process: the specific film stock which may be "cooler" or "warmer" in tone, lighting, developing, printing. Shifts occurring in the images made from the originals again have to do with lighting and other conditions obtaining during capture. Further, certain scanners are inherently biased toward certain color ranges. A scanner designed to capture skin tones well will have trouble capturing blues accurately, for instance.

It is possible to calibrate scanners, monitors, and printers by defining each of the 16 million colors relative to a standard, as long as standard color charts are included in the process.[16] Use of color manipulation software can assist in bringing the colors closer to the original. This is not "enhancement" but rather a way to reduce the amount of interference in the capture of the original color. Such manipulation takes considerable human intervention and time and is therefore costly. In our instructions to our vendors we had not required color manipulation but instead put the emphasis on legibility. For those who are concerned to create a very accurate copy of the original, more investigation is needed to assure that the color has been captured precisely; developing methods for accurate color display and printing is also a priority.

In sum, both color and text were adequately preserved under our definition of success. The information conveyed by the color codes is captured as long as the codes remain distinct, even though the actual hues were altered, and there is a definable resolution at which all the textual information is fully legible. There is no gain in using yet higher resolution because there is nothing further to capture. This contrasts with the situation in scanning works of art or historical artifacts with many subtle color tones, and where important information content may be contained in the very fibers of the paper. Quite different results would be expected for archival materials, reproductions of works of art, and other images with a wider palette and more subtle detail. One example that can be cited is papyrus documents, the subject of ongoing experimentation by a group of institutions. A panel of papyrologists, librarians, and digitization experts have tentatively suggested that 600 dpi plus 24-bit color may be adequate.[17]

[15] Anne Kenney and Stephen Chapman, op. cit. p.10 ff.

[16] Fred Mintzer et al., "A Computer System for Scanning and Cataloging the Art of Andrew Wyeth," *Spectra* 1992, pp.9-15.

[17] Examples of scanned papyrus are available from several sites. Examples may be found at URL

Given that this project dealt only with modern printed maps where the information to be captured is partly textual (place names, labels, numbers), partly linear (roads, borders), and partly codes made up of a relatively limited number of colors and patterns, it is clearly possible to achieve adequate results starting from either the original map or the microfiche. The hybrid preservation approach works for this particular medium. We can create preservation-quality microfiche and transparencies to serve as surrogates or replacements for brittle originals and be confident that we can scan the film intermediaries and capture the intellectual content of the originals even if those originals have been lost.

Capture is one side of the coin, delivery is the other. We can currently capture more information that can readily be transmitted over the Internet or displayed on average monitors. The files of the scanned maps captured at high resolution and 24-bit color can range as large as 20 mb when uncompressed. Clearly they must be scaled back for ease of access; we have mounted on our Web site versions of the file with only 256 colors and resolution only up to about 150 dpi, resulting in files as large as 6 mb. Lossless GIF files at 256 colors and lossy JPEG files at 16 million colors are available for comparison.

Unfortunately, files which are easily transmitted and viewed carry too little detail for the larger maps to be fully useful on-line. The more complex and larger maps can not even be browsed at resolutions which can be transmitted easily; but images at high resolution can only be viewed in small sections at the screen and this can sometimes be disorienting when trying to read a map. Studies have indicated that peripheral vision is very important in map use,[18] but the size of the maps means it is impossible to make the entire map both visible and legible simultaneously. We needed to investigate users' success with different delivery modes in order to find a preferred solution.

Let me turn now to Phase II of the project, which began this spring. Phase II present scholars with images and text in juxtaposition by creating an integrated on-line version. The project is assembling a copy of four volumes of the *Museum Bulletin*, using a combination of microfilm of the text and single-frame color microfiche of the illustrations. We have scanned the microfilm of the text at 600 dpi black-and-white and have scanned the microfiche of the illustrations at approximately 200 dpi, 24-bit color. The project will use indexing and document structure software to integrate the files of pages and illustrations and recreate the original volumes virtually on the Internet. One benefit of working with a government publication, incidentally, is that we can put the images up for public display because copyright is not at issue.

http://www.lib.umich.edu/pap/HomePage.html from the University of Michigan, and at URL http://odyssey.lib.duke.edu/papyrus from Duke University.

[18] See, for instance, the report "Making Maps Easy to Read" by Richard Phillips (URL http://acorn.educ.nottingham.ac.uk//ShellCent/maps/) on a project conducted at University College London at the Royal College of Art, and at the University of Nottingham.

Quite a number of models exist for how to display bitmapped book pages on-line. Our project is specifically concerned with combining the text and the high quality color images. Again we have a specific definition of success. First, the bit-mapped black-and-white text must be legible, including the smallest type. Second, the indexing and document structure must permit quick identification of and access to the files for every page and every illustration. Third, the user interface must be self-explanatory and easy to use so that readers can concentrate on the content of the materials and not on the mechanics of moving from volume to volume and page to page. The result we are aiming for is the full preservation of four sample volumes: long-lasting microfilm and microfiche of all the text and illustrations, along with a truly usable digital version of them where the author's juxtaposition of words and illustrations is also preserved and available for users to read on-line or to create paper printouts.

Conclusion

Many questions remain to be answered. To what extent and for what purposes will scanning to these specifications satisfy the scholarly community's needs? Will the digital version suffice for text retrieval? Will the quality of the color images be satisfactory for classroom use and scholarly research? What role will they play for scholars interested in detailed analysis of the maps — as pointers to the originals which must then be consulted, or to requests for printouts, or will some scholars be able to do much of their work with the digital images alone?

Further guidelines for quality of resolution must be developed, and so must schemes for institutional quality control evaluation to assure that the required levels of quality have actually been achieved. Improving the accuracy of color capture and display is another obvious area for research, as is tiling (where an oversize document is scanned in sections and the sections combined into one large image). Not unimportantly, the real costs of high-quality digital imaging must be determined, so that decisions on what endangered materials to convert and when can be made on a firm financial basis.

Finally, I would like to comment briefly on project management. Administratively, Columbia is using a team approach to digital library issues. Decisions on institutional priorities for which collections should be digitized in response to our clientele's needs, setting policies for commitment of technical, hardware, and financial support, planning for the future to assure that digital files will be maintained, refreshed, and migrated — all of these decisions must be made system-wide. Several standing committees have been established by the Columbia University Libraries to oversee the progress of the digital library and to discuss issues in depth and come up with decisions.

The committee most relevant to the oversize images project is headed by the Deputy University Librarian and it is composed of the Library Systems Officer, staff from Academic Information Systems (AcIS, formerly Academic Computing) responsible for coordinating faculty and library imaging projects and mounting them on the Columbia Web server, the Director for Bibliographic Control, curators of collections in which projects are currently taking place, and the Director for Preservation. Each individual project also has a management team that brings together expertise on the subject matter and scholarly needs, preservation, and the technical side of imaging. The oversize images project has a three-member team from Preservation, Geology, and AcIS. We have found this a very fruitful collaboration in which our pooled expertise has helped each of us learn a great deal, and I strongly recommend it as a model.

The use of digital imaging for preservation purposes offers exciting possibilities. While much further research awaits, it is our hope that this project on the digitization of oversize color images will help to answer some of the questions.