I. Introduction

A. Objectives

Preservation Resources offers this proposal to Columbia University Libraries (CUL) to provide scanning and indexing services for the New York State Coordinated Project.

CUL is proposing a project to integrate digital scans of black and white text, color and continuous tone illustrations, index the files to the illustration, page and chapter levels. The project includes plans to scan four volumes of the New York State Bulletin. Each volume contains one or more articles/chapters and multiple illustrations. The illustrations include oversized color and continuous tone, color plates, black line drawings. CUL will provide digital scans of the color and continuous tone illustrations.

Preservation Resources will provide a 600 dpi bi-level scan of the text of four volumes on existing 35mm reels, will integrate bi-level scan with color scans, will index all the digital files and deliver the images in TIFF format group 3 or 4. We have estimated 300 pages per volume for a total of 1200 microfilm frames and 2400 images.

The project will be conducted in three broad, overlapping phases. The first phase will entail staging and inspection of film. The second phase will entail scanning of film, editing of scanned images, and integration of digital files. The third phase will entail indexing. Quality assurance will be part of every phase and is addressed as part of each workflow described in this proposal.

The project deliverables will include scanned and indexed images written to 4mm DAT magnetic tape. The files will be indexed to provide hierarchical access to the scanned pages and illustrations for each volume as well as to the chapter/article level within each volume.

Preservation Resources will complete the project in approximately six weeks, from July 15, 1995 to September 1, 1995.

B. Specifications and Standards

The preservation reformatting production procedures and systems in place at Preservation Resources are based on adherence to the specifications, guidelines, and standards from the American National Standards Institute (ANSI), the Library of Congress (L.C.), and the Research Library Group (RLG). Although guidelines for digital image scanning have yet to be broadly adopted by the library community, standards for scan quality have been published by ANSI/AIIM and will be followed by Preservation Resources. A copy of specifications followed by Preservation Resources is attached (Attachment 1).

C. Equipment/Software

The Mekel Engineering produced M400XL ITools™ Microfilm Scanning System will be used to transfer 35mm microfilm into digital image files. The M400XL will scan the microfilm continuously using a high resolution 5000 pixel CCD (charge couple device) linear array camera. The Mekel's TurboScan™ is designed to automatically detect the beginning and trailing edge of each new image. Equipment specifications appear in Attachment 2.

Commercially available software designed specifically for film-to-image file conversion will be used. Available from and supported by Mekel Engineering, (a division of Amitech Corporation), this package is an integrated hardware/software system, and is considered a proven technology. We will be using the most recent release of some of their new applications, which are integrated in with the scanning process.

II. Scanning Microfilm and Quality Assurance

A. Microfilm Quality

Preservation Resources recommends that scanning be done from second generation negatives where possible. Image content can be captured and a positive digital image produced from either positive or negative roll films, however, testing conducted by Amitech using the Mekel 400XL scanner with the latest versions of optimization software and evaluated by Preservation Resources indicates that scanning from negative film produces far less "speckling" or background noise caused by dust specks or fine imperfections in the image. These findings are corroborated by Yale University Library during the setup phase of their "Project Open Book."

Production capacity and quality of bi-level scanned images are limited by two primary factors: (1). the condition and quality of the microfilm, and (2). the condition and quality of the original documents.

Some conditions caused by filming methods, or the nature or condition of the material being filmed require a halt in the automated scanning operation to reset scanning parameters or to manually reposition the film. The frequency of occurrence of these conditions have a corresponding negative impact on production. Given the content of the microfilm, Preservation Resources anticipates that some level of "special handling" will be required. Our experience with preservation microfilm has shown that unlike document imaging in a business context, it is not realistic to assume universally standardized scanning settings when digitizing microfilm of retrospective research materials.

B. Special Handling

Based upon previous experience, Preservation Resources has found that scanning of microfilmed research materials requires some level of special handling. Based on a sampling, it is estimated that 8% of all frames will require special efforts to optimize and/or edit. This 8% special handling assumption is reflected in the price proposal.

Conditions requiring special handling are as follows: un-spliced retake frames; abrupt density changes; reduction ratio changes; skew severe enough to possibly affect leading edge detection, text obscured causing optimization changes and orientation changes. Serious cases of such conditions make optimization using automatic or semi-automatic (using preset special case optimization settings) scanning parameters impossible and require manual, single-frame optimization and image editing necessary. Special handling will also be required when foxing; bleed-thorough; stains and discoloration; ink, handwriting and smudges; and acidification fading occur in original documentation.

Two post-scan editing procedures will be undertaken to ensure quality control. In the Manual Image editing process, the 8% of the images requiring special handling will be corrected as much as is feasible.

III. Scanning WorkPlan

A. Workflow

A summary of the scanning workflow is as follows.

(1.) Reel tracking and Pre-Scan Evaluation:
Inspect microfilm. Determine whether manual optimization settings will be necessary for scanning. If so, test for 'family' settings. If not, send for automatic optimization and determine breaks for reduction changes, density changes, orientation etc. Prepare worksheet.

(2.) Scan, Edit & Quality Assurance:
Based on worksheet, program batches and set optimization as needed. Perform visual inspection of images as they are scanned. As indices are scanned technician will print out hard copy and note image file name for each page of index. ITools software will perform "intelligent" editing for deskew, rotation, filtering and cropping asynchronous to, but at the time of the scan. Targets, duplicate frames etc. will be noted by file name for later deletion. "Special Handling" image files will be noted for manual image scan/editing.

(a.) Run TurboQA™ software on directories of scanned images to check for file errors and user-defined file parameters to detect and list images which are likely to be incorrectly written or badly scanned. This software runs unattended and requires only batch setup. Flagged files can be accessed directly from the TurboQA™ software for disposition - either rescan or appropriate manual scan/editing and indexing.

(b.) Customized manual rescan and image editing (special handling) for images which resist automatic and preset specialized optimization settings. Localized thresholding may be necessary where adaptive thresholding proves insufficient for images such as those containing unwanted areas such as stains, faded edges etc. which are darker than the text in other parts of the page. We estimate 8% of all images will require special editing.

(c.) Batch editing such as splitting and renaming 2-up images into separate image files, and/or rotating a selected group of images is done from worksheet or notes made on worksheet by scanner technician. The worksheet prepared during the preview stage should follow the images scanned from each reel of film to aid in workflow and to maintain bibliographic integrity of sets of image files. At this stage, unwanted images as noted by the scanning technician will be deleted.

(3.) Indexing:
Index the image files by creating a spreadsheet-based, tabular format consisting of rows (records) for each image file and columns (fields) for each level of indexing.

(a.) Locate the appropriate image file for each of the 89 illustrations. The references are to individual page numbers within the volumes, and will correspond to one or more image files.

(4.) Write image and indexing files to DAT tape.

B. Reel Tracking and Pre-Scan Evaluation

In order to establish scanning parameters each reel of microfilm will be inspected prior to scanning. A quality assurance and tracking worksheet will be created for each reel with reel number and basic bibliographic identification data as available from the reel label and filmed targets. This worksheet will provide detailed film and document condition information and will be a tool to track each step in the process, including the scanning, editing, and indexing of each reel of film.

Reels will be placed on manual film rewinds and examined over a light-box with a loupe with 8x magnification. Film will be evaluated for anomalies which will require adjusting the settings during scanning and allow for batch processing.

The following characteristics will be noted: reduction ratio of each reel and any changes, image orientation of each reel (1B, 2B, 1A, 2A) and any changes including the presence of foldouts, presence of film targets. All anomalies will be noted on the inspection worksheet. Each reel will be cleaned to eliminate any dust, fingerprints, or similar problems which would introduce "noise" in the scanning process.

Based on preliminary inspection data, reels will be assigned one or more batch numbers with manual optimization settings noted as necessary. "Batches" or continuous runs of scanning will follow the bibliographic content of the reels, i.e. a new volume will be assigned a new batch number.

C. Scanning

Upon completion of inspection, batched reels will be assigned a scan date with programming batches and setting optimizations. Reels will be scanned at a single Mekel work station.

The actual spatial resolution of the scan is dependent on the original document size, reduction ratio, and orientation on the film (either comic or cine mode). Given the reduction ratio at which the original document was filmed, the software will calculate the necessary spatial resolution to achieve 300 dpi (dots per inch), relative to the original size of the document, as determined by the reduction ratio, per pass along each reel.

The storage and compression processes occur asynchronously (under the control of separate processing boards thus allowing simultaneous processing) to the scan and will therefore not require any additional time.

After scanning, the first line of quality control commences: as part of the normal function of the scanner system, each frame will undergo an image enhancement, via Mekel's Automatic Optimizer™. This process will automatically determine and set enhancement and thresholding parameters to enhance faded, low-contrast materials into cleaner, sharper, and more legible digital images. Next, the ITools™ software package will perform "intelligent" editing for deskew, rotation, scaling, filtering, and cropping asynchronous to, but at the time of, the scan. This concurrent processing of scanned image files will automatically reduce file sizes up to 20%.

After scanning, each image will be opened using image editing software for 100% Quality Assurance inspection and to correct for image orientation, cropping, color, balance, contrast etc., or to determine if re-scanning is necessary to assure appropriate aesthetic and technical quality. At the same time, each file will be duplicated and sub sampled to a lower resolution file suitable for on-screen viewing. Each file, both full-scale and lower resolution versions, will be named using a scheme suitable to differentiate volumes and maintain the linear page to page content of the original material for identification and access.

1. Post-Scan Evaluation

Image quality problems not compensated for before or during scanning will be corrected in the post-scan process. Although each frame will be automatically cropped, scaled and rotated as necessary according to varying pre-set specifications programmed into the scanner, it is estimated that 8% or 281 images (includes 89 illustrations) will require some degree of "special handling" to optimize the images. Optimizing will involve selecting the scanning controls to automatically adjust the shading of images ("thresholding"), the thickening or thinning of lines and letters, despeckling the frame, etc. For example, it is assumed that there will be a need to crop and rotate images containing charts and tables, etc. in order that they be oriented so as to be made readable on the computer screen by the user.

After all steps in the scanning process have been completed, the scanned images will be evaluated to ensure adequate optimization. The Mekel's TurboQC™ Image Quality Control Software will be used to automate the process. It is run on directories of scanned images to check for image file integrity (such as file header or compression errors) and will also employ user-defined file parameters to detect and list images which are likely to have been badly scanned. A found set of suspect files are put into a list which can be accessed directly from the TurboQC™ software for disposition -- either for rescanning or appropriate editing and indexing.

In addition to quality control, batch-editing will also be used for adjustments not made by ITools™ at the time of the scan. For instance, splitting all two page images, rotating, cropping and deskewing them based on determinations noted on the worksheet during the preview/evaluation stage or problems noted during the scanning or quality control stages. Automated processing functions such as these will be accomplished using TurboImageEdit™, a high-speed batch image processing software program which will open each file in turn, perform the necessary editing operations and rename the split images to maintain the linearity necessary to facilitate the indexing process.

IV. Indexing

A. Indexing/Bibliographic Data

An index to the scanned images will be created which may be imported into on-line database access software. Data to be entered into the index will be assembled through each stage of the project workflow, beginning at the Pre-scan Inspection stage.

During evaluation of each reel, the technician will note the volume/chapter/article on the worksheet for the reel/volume. The appropriate directory names, and file name protocol will be assigned at this stage.

During scanning, the technician will enter file names on the worksheet to indicate initial pages from each section of the document being scanned, and will also indicate files which contain illustrations to facilitate the later search for these images.

After the batch editing, directories will be created to hold sequential sets of files and to comprise the hierarchical directory structure.

The entry of data into fields in the index will be accomplished after these preliminary steps have been completed. Sequential file names will be entered using a utility macro written for this purpose. Entries will be made in the appropriate field for the record of the image at which new indexing levels are reached, and repeating field entries between these "milestone" points will be entered using another utility macro.

Individual indexes will be created for each volume in the project and may be merged into a single index for access when creating the on-line interface.

B. Index Structure

The index will be created in a spreadsheet-based tabular format consisting of rows (records) for each image file and columns (fields) for each level of indexing. Field definitions will include:

- Volume
- Title, Year
- Chapter/article
- Subject level page number
- Illustration
- File name

In order to facilitate the creation of the on-line access interface, two predefined conventions will be followed in the creation of the index : (1) vertically sequential file naming and, (2) vertically and laterally consistent directory structure. Additionally, the index will provide identification of individual files for lateral access and cross-referencing to the illustrations.

(1) File names will conform to DOS/WINDOWS style of 8 alpha-numeric characters in the file name followed by a "." and a three character extension. The first three characters (two numbers and a letter) of the file name will number in a linear or vertical fashion the document in the logical bibliographic sequence according to the Volume/Chapter/Article and Title/Year indexing levels. The next four numbers in the file name will sequentially number the images within the document. The final character will be reserved for "a", "b" designations used when splitting two page scanned frames into separate image files. Thus, within each document, all image files will appear in a directory in the correct bibliographic order.

(2) The directory structure in which the image files will be stored will be hierarchical and will mirror the index itself, with each level of indexing being represented by directories, down to the lowest level of indexing. The deepest directories in the hierarchical structure will contain the actual image files.

The illustration references will be entered in the appropriate field in the record for the image file cited. The directory structure and file naming conventions described above will facilitate the location of the image files as these entries are made by going through the guide sequentially, and locating the appropriate record for the referenced image.

V. Data Transfer and Compression

Preservation Resources will deliver indexed images on 4mm DAT tapes, each storing 8 gigabytes of information (compression: CCITT, Group 3).