NATIONAL PARK SERVICE
DEPARTMENT OF THE INTERIOR

Conserve O Gram

September 2000 **Number 19/22**

Managing Digital Projects For Preservation And Access

Managing digital projects is a complex process that involves:

- selection of materials to be digitized
- intellectual rights management, including copyrights, privacy/publicity rights, and cultural/donor sensitivities
- preparation of materials to be digitized including preservation and cataloging/ description/indexing of originals
- creation of photographic or microfilm copy images to be scanned instead of the originals
- digitization through trained use of appropriate equipment to produce high quality files
- quality control checking of files
- metadata production and indexing of files
- storage, refreshing, and ongoing migration of digital files as hardware and software changes and storage media age
- management of network and infrastructure
- ongoing management of all metadata (data on authority, contents, format/structure).

Also see *Conserve O Gram* 19/21, Planning Digital Projects for Preservation and Access, for more information.

Legal Issues: Don't digitize items unless:

- you have copyright notices, or
- you have a special license, or
- the material is in the public domain, or
- you have written permission from rights holders, such as interviewees, and
- you have consulted with appropriate cultural groups and individuals that may be affected by digital publication of the materials

See NPS *Museum Handbook*, Part III, Chapter 1: Evaluating and Documenting Museum Collections Use, and Chapter 2: Legal Issues.

Preservation Concerns: Teach scanning staff handling, scanning, and packing/shipping techniques. Limit exposure to bright lights. Prohibit pressing down on or form-feeding of items. Monitor the scanning laboratory environment and security. If possible, scan from copies, duplicates, or surrogates. For information on reformatting originals before scanning see *Conserve O Grams* 19/10-19/12.

Cataloging/Finding Aids/Indexing/Metadata: Identify, index, and provide metadata of each file produced. Follow metadata standards, such as the Dublin Core (see < http://purl.org/dc >).

Image Quality: Image quality is the sum total of the following:

- scanning resolution and dynamic range
- hardware used (including the monitor and printer)
- software (see below)
- production benchmarks (see below)

- techniques and formats used, including compression and file formats
- operator skill level
- qualities of the originals (such as media, contrast, size, color, process, technique)
- quality of the negative or microfilm copy
- security of the file from tampering

Good image quality for all digitization requires:

- Large file sizes of original scanned data
- High capacity storage devices
- High bandwidth networks
- Memory to display the larger files
- Image compression of large derivative files to maintain them less expensively
- System stability and security from tampering
- Good quality control
- Recopying (refreshing) and migration (copying to new file formats) as hardware and software change and file formats change

Scanning Standards: Use the Quality Index (QI) techniques and formulas described by Kenney and Chapman (1996), which include:

- determining matched image quality levels for scanned images and Association of Information and Image Management (AIIM) microfilm resolution standards
- converting measurements to metric
- calculating the equivalency ratio between digital dots per line in the digital file and widths of text in the document
- adjusting for misregistration in bitonal scanning

Reformatting Originals: Producing and working from high quality copies, such as photos or microfilm, helps preserve originals and provides analog usage copies that will survive. Scanning from copies is also cheaper than scanning from fragile originals. Scanning from copies, however can produce a somewhat lower quality scan than scanning from original documents. Not all document sizes and formats can be filmed in the same size and format of microfilm or photographic negative prior to scanning. Also, if you are scanning originals and outputting to computer output microfilm, different document sizes/formats may require different scanners.

Optical Character Recognition (OCR): Choose OCR for digitizing text when you want rapid access, low cost storage, small files, original page layout maintained, terms from the pages used as the index language, and can tolerate fuzzy (imprecise) searching.

Plan for slow, post-scanning clean-up of raw OCR text. OCR error rate is often 5% + on high quality source material, meaning staff will spend about 8.5 minutes per page manually correcting text at 50 words per minute when scanning from good quality originals.

> According to recent estimates, fully searchable OCR costs 1.5-18 times the cost of scanning and indexing. Total costs of re-keying data is roughly $8.80-$11.40 per page while average OCR costs run about $4.40 per page. Although low cost, OCR doesn't always work for all types and qualities of documents. See Kenney and Chapman (1996) and Puglia (1999).

File Formats: There are thousands of file formats, many for special uses. If scanning both image and text together, the file format you select will need to be a compromise between what is best for each type of document. Use different file formats for your master file (your digital preservation master) and your derivative files (your access and usage copies).

For digital master files, select a non-proprietary (not owned by a vendor) file format, such as TIFF or GIF for ease of migration. Vendors rarely support proprietary formats for long. Many proprietary systems can't use files created on earlier software and hardware.

Digital master file size depends on your established benchmarks. Use lossless compression (such as TIFF in which the decompressed copy looks like the original) rather than lossy compression (such as GIF where data is discarded to achieve the compressed file size) for all master files. Scan your materials as described below:

- Scan text, maps, drawings and graphics < 11"x17" at 300-600 dots per inch (dpi) according to item bench-marks. Halftones need 24-bit color or bitonal scanning.
- Scan documents > 11"x17" at 200 dpi, or segment the file
- Scan photos using uncompressed TIFF. Adjust the dots per inch (dpi) to get 3,000 pixels across the long dimensions at 300 dpi.

For usage files, select a widely used, easy-to-download file format, such as GIF for documents or JPEG for images. Usage files may use lossy compression.

In-house vs Contract Scanning: Before choosing equipment, decide if you want to do the scanning in-house or contract out. Consider such issues as transportation, site security, costs, supervision needs, needs to purchase and maintain calibrated scanners and software, and whether the contractor can match your standards and needs.

Scanner Features: Different models have different capabilities. When considering a scanner, query your peers about potential problems. A good scanner has the following features:

- automatic edge finding
- automatic exposure
- bordering
- brightness adjustment
- centering
- color correction
- contrast adjustment
- cropping
- dropping out of background noise
- ease of use
- highlight adjustment
- manual override for automatic functions
- merging capabilities
- speed

Sheet-feed scanners scan stacks of documents fast, copying both sides. They have limited image enhancement capabilities, document size limitations, and are not suitable for photos, or fragile, oversize, or bound items.

Slide scanners scan transparencies of most sizes/shapes, produce images with a good dynamic range, and provide archival back-up. They only work with transparencies and may not work well if the original is opaque. Slide scanner copy resolution is often poor; some scanners are slow.

You may also want to consider:

- Video cameras that capture digital image files on diskettes.
- Video digitizers that record 3-D objects, sound, and activities/events, similar to a video camera.

Scanning Software: Good scanning software is essential. Scanning software controls the scanner and manages bit depth, compression, enhancements, file formats, image manipulation, resolution, and thresholding. Select your software with care. Buy scanning software that allows color management, customized setting reuse, document and file-naming control capabilities, default settings for image processing, multiple document types, multiple compression

formats, multiple file formats, and page segmentation.

Quality Control: Check the collection description and indexing (100%). Record device settings during scanning and transfer the information to the metadata you maintain and manage. Follow these guidelines:

- Proofread all data entry. Check indexing against thesauri.
- Verify the relationships between text and image.
- Check digital copies against benchmarks.
- Inspect the image/text (100%). Ensure that copies capture the appearance (tone, detail, color, process) of originals.
- Use targets (test charts) to measure resolution. See Kenney and Chapman (1996) and National Archives and Records Administration specifications at < http://www.nara.gov/nara/vision /eap/eapspec.html > .
- Review the copy side-by-side against the original.
  - For Images: Check the color balance and gray scale against the original and targets. Use your system benchmarks. To judge the color quality of a copy, compare it to the original. Don't fix color inaccuracies by adjusting device settings; maintain masters to your benchmarks. Adjust derivatives. Check exposure, fine details, image alignment, and uniformity. Watch for distortion and poor focus or sharpness.
  - For Text: Check page completeness, contrast, legibility, text density, character size, line widths, and letter clarity.
- Coordinate rescanning of unacceptable items.

References

Ester, Michael. *Digital Image Collections: Issues and Practice*. Washington, DC: Council on Library and Information Resources (CLIR), 1996. On the Web at < www.clir.org/ pubs/abstract/pub67.html > .

Hirtle, Peter and Carol DeNatale. *Selecting a Digital Camera*. On the Web at < www.rlg.org/ preserv/dignews/diginews2-6.html > .

Kenney, Anne, and Stephen Chapman. *Digital Imaging for Libraries and Archives*. Ithaca, NY: Cornell University, 1996.

Lawrence, Gregor W., William Kehoe, Oya Rieger, William Walters, and Anne Kenney. *Risk Management of Digital Information: A File Format Investigation*. Washington, DC: CLIR, 2000. On the Web at < www.clir.org/pubs/ > .

Puglia, Steven. *The Costs of Digital Imaging Projects*. 1999. On the Web at < www.rlg.org/ preserv.diginews/diginews3-5.html > and *U.S. NARA Electronic Access Project Scanning and File Format Matrix*. Washington, DC: National Archives and Records Administration, 1997.

Diane Vogt-O'Connor
Senior Archivist
Museum Management Program
National Park Service
Washington, DC 20240

The *Conserve O Gram* series is published as a reference on collections management and curatorial issues. Mention of a product, a manufacturer, or a supplier by name in this publication does not constitute an endorsement of that product or supplier by the National Park Service. Sources named are not all inclusive. It is suggested that readers also seek alternative product and vendor information in order to assess the full range of available supplies and equipment.

The series is distributed to all NPS units and is available to non-NPS institutions and interested individuals by subscription through the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402; FAX (202) 512-2250. For further information and guidance concerning any of the topics or procedures addressed in the series, contact NPS Museum Management Program, 1849 C Street NW (NC 230), Washington, DC 20240; (202) 343-8142.