# Overview of digital image handling practices

# University of Wisconsin Digital Collections Center

## Principles

Generally, the University of Wisconsin Digital Collections Center (UWDCC) treats digitization as a tool to serve access needs, with the assumption that the original material will continue to be retained into the foreseeable future. At the same time, we recognize the continuing value of quality digital reproductions for all materials, and are therefore pursuing further strategies for initial capture as well as refreshing and migrating our digital files that will provide for preservation of the digital content produced from our material.

## Capture

The precise settings used for a particular project may vary depending on the nature of the material or any requirements specified by the content provider, but in general the following guidelines are used:

For text-only or text with line art: 600 dpi, bitonal capture
Photographs or half-tone-illustrated material: 300-400 dpi, 4- or 8-bit capture
Pages with significant color information: 300-400 dpi, 24-bit color

35 mm slides: 2000 dpi, 24-bit color (results in approx. 2900 x 2000 pixel image)

## Processing

A limited amount of work may be done to the master images in order to improve usability and reduce storage needs. In all cases the emphasis is on reproduction of the intellectual content of the original item. Once the master image has been optimized, several sizes of derivatives are created from each master for use online. In addition, images of textual material are processed using optical character recognition (OCR) in order to provide basic text searching. The following processing steps are generally used:

**Pre-scan:** settings affecting brightness and contrast (i.e. the mapping from the scanner's bit-depth to the desired bit-depth) are adjusted. This may be done by sampling or on a per-image basis.

**Master image processing:** scanned images will generally be cropped, and will be deskewed if necessary. Material is typically cropped to within a 1/4 to 1/2 inch margin around the content of the page; the edges of the page are usually cropped off. Visual materials (photographs, illustrated material) may also be adjusted to fine-tune the brightness and contrast. In rare cases, efforts to correct colors which have faded or shifted will be undertaken. Image processing is done in Photoshop (currently using version 7.0) often taking advantage of Photoshop's batch automation capabilities.

**Derivative processing for user displays:** For multimedia projects (generally photographic images), three derivatives are provided, varying according to the number of pixels on the longest dimension: either 720, 300, or 150. In some cases, we

have provided a fourth size: 1500 pixels on the longest dimension. This processing is done is batch mode using a combination of AppleScript and Photoshop actions.

For electronic facsimile projects (generally text-based images), four derivatives are provided which may be produced by either of two models. Under one model, three of the derivatives vary according to base resolution (200, 120, and 80 dpi) with the fourth created to have a 600 pixel maximum dimension. As with the multimedia images, this model uses AppleScript and Photoshop for processing. Under the second model, derivatives are created based on the following pixel widths: 352, 472, 568, and 712. These derivatives are created using ImageMagic software.

**OCR** text is produced using Prime Recognition software with three licensed engines. The software is capable of handling grayscale and color as well as bitonal images. The OCR results are not proof-read or manually corrected, so some care is taken to sample and optimize settings for large batches of material.

## Storage

Master images are created and saved as TIFF images. Bitonal masters are compressed using CCITT Group IV compression; grayscale and color masters are uncompressed.

Derivative images produced using Photoshop are in JPEG format at various levels of compression; those produced using ImageMagic are reduced-palette GIF images (3-bit/8 levels of gray).

Bitonal masters and all derivatives are ultimately stored on server hard drives, which are regularly backed up to tape. Grayscale and color masters are archived to tape with triple redundancy.

## Metadata

Metadata for multimedia images is primarily descriptive in nature, and based on the Dublin Core set of elements. Metadata is entered into an Access or Filemaker Pro database; then converted and stored in a standardized SGML form. SiteSearch software is the engine which fundamentally drives the searching capabilities of the multimedia collections, although the actual interface is the result of extensive local customization.

For electronic facsimile projects, the metadata is primarily structural with some basic bibliographic elements added. The data is initially organized in a set of Excel spreadsheets, which are then exported and processed into TEI (Text Encoding Initiative)-compliant SGML files. Open Text software, in conjunction with an array of PERL scripts initially developed at Michigan, forms the basis of the user interface, which, again, has been heavily customized for local use.

At this time, administrative metadata attached to the images themselves is limited to that included in the TIFF image header by default, although we are currently developing guidelines for collection and handling of a more extensive set of administrative information.

January 2004