Federal Agencies
Digital Guidelines Initiative

September 2016

Technical Guidelines for Digitizing Cultural Heritage Materials

Creation of Raster Image Files

Document Information

Title	*Editor*
Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files	Thomas Rieger
Document Type	Technical Guidelines
Publication Date	September 2016

Source Documents

Title	*Editors*
Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files *http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf*	Don Williams and Michael Stelmach
Document Type	Technical Guidelines
Publication Date	August 2010
Title	**Authors**
Technical Guidelines for Digitizing Archival Records for Electronic Access: Creation of Production Master Files – Raster Images *http://www.archives.gov/preservation/technical/guidelines.pdf*	Steven Puglia, Jeffrey Reed, and Erin Rhodes U.S. National Archives and Records Administration
Document Type	Technical Guidelines
Publication Date	June 2004

Imaging Workflow

Overview

Digitization workflow planning follows the steps outlined below. It is a balancing act – each part needs to be designed to work with the other, and function at the same capacity. If designed and implemented properly, digitization can be a smooth and efficient process. If any one element of the process is not designed and/or implemented correctly, the whole system can become inefficient and wasteful.

Follow these steps in designing a digitization workflow correct for your project.

Define the Requirement

Planning begins with an evaluation of the requirement. These questions need to be answered:

- What is the scope of the requirement?
- What is the timeframe required for completion?
- What types of materials are involved?
- What is the expected result?

Assess Organizational Capabilities

Once the project is defined, review the capabilities of the organization to accomplish the goal defined above. Questions to answer:

- What staff resources are available?
- What are the digitization skills of the available staff?
- What physical space is available to devote to the effort?
- What digitization equipment is available?
- What new digitization equipment is needed?
- What level of funding/resources is available?

Insource vs. Outsource

If the capabilities of the organization are determined to not be equal to the requirement, the issue needs to be addressed, or the decision to outsource the project needs to be made.

The answer to this question may determine the success or failure of the program. A project with inadequate resources may produce results with little or no long term value, and could even result in damage or loss of the collections. A well planned and executed effort will save time, effort, and resources.

Resist the temptation to dive into a digitization project without proper resources. If the organization does not have the resources, it may be best to outsource the project. Also, avoid the trap of assuming doing the work in-house will cost less. Insourcing may cost more than outsourcing.

Project Management

Critically important to the success of any project is project management. Set target production goals and establish a system to measure progress. Establish a quality monitoring system based on the DICE system to verify the results and inform if adjustments are needed. Keep records of both quality and quantity.

Workflow Plan

Once the project is defined and decisions concerning project management and insourcing vs. outsourcing have been made, a workflow plan needs to be developed.

The workflow plan takes into account the flow from ingest through archive, which includes the following steps:

- Selection of materials
- Condition evaluation
- Cataloging
- Metadata creation
- Production scheduling
- Digitization prep
- Digitization
- Post processing
- Quality review
- Archiving
- Publishing

Large scale project workflow

Selection of Materials

Project workflow plans are unique to each situation, but all follow the same logic. Starting with the selection of the materials, decisions need to be made about the types of materials that will be digitized, and in what sequence. Grouping like materials together can create efficiencies in acquisition process, but may create organizational challenges. Switching back and forth among digitization setups is time consuming and disruptive. Production planning is necessary to balance the competing elements to achieve the most efficient production.

Condition Evaluation

All materials to be digitized should be evaluated for their condition, and appropriate steps taken if problems are discovered. Materials may have mold issues, which present health risks, or may suffer any number of other issues that should be evaluated before digitization.

Metadata

Metadata should generally be acquired before digitization begins. Technical metadata should be generated during digitization.

Production Scheduling

Selection, condition evaluation and metadata creation should be designed to fit efficiently with later processing actions. The process should be designed to have a buffer of project work ready for digitization to permit flexibility in scheduling. If the project involves multiple scanners and material types, buffers for each imaging task should be maintained. Production scheduling is the essence of an efficient project.

Digitization Prep

Materials are assembled in the workspace and made ready to digitize. Information about file naming, shoot order, project specifications, etc., should be clearly organized. Art conservation treatment should be completed, and special handling needs addressed in advance.

Digitization

The actual digitization process will vary dramatically based on the type of materials being digitized and the scale of the effort. FADGI does not provide recommendations on the specific design of digitization projects.

Post Processing

This process is highly specific to the project goals, but in general involves adjustment of the files to a common standard image specification and project specific requirements. It may or may not involve the generation of derivative files and the creation of file folder structures and metadata.

Quality Review

Separate from the post process step is quality review. This process may involve automated inspection for conformance to technical specifications, followed by physical inspection, where an inspector reviews a representative sample of the project to verify conformance to the project specifications. The size of the sample is determined in the project planning stage. Automated tools such as JHOVE can assist in verification of the files.

Archiving

Once the project, or segment of the project if appropriate, has passed quality review, it is passed on to be archived. Additional testing for file integrity may be done at this stage. Refer to the section on storage options for additional information.

Publishing

Master files are not suitable for presentation, except in unusual circumstances. For publishing or presentation, derivative files are created in an appropriate format for the application. These can be created as a structured process in post processing, or dynamically as needed. Care must be taken to assure that color transformations are done correctly as appropriate for each application.

The following section outlines – in broad strokes – the workflows that have been perfected in several Federal agencies. They are presented for reference and as a guide to designing workflows.

Example Workflow #1

The following is an example of a project that has a distinct goal and end point. There is a collection of paper materials that needs to be digitized in its entirety.

- Selection of material
    - All objects are of various sizes and materials but the overall project is so vast that the determination is to scan everything in the same manner on the same device.
        - All objects are scanned with the same resolution
    - Materials are in the same location and stored in the same manner. Therefore, organization is uncomplicated as long as it is tracked through the use of databases amongst staff.
    - Some material throughout the collection has already been scanned. Items that have been digitized previously are reviewed, and it is determined by staff whether or not to rescan based on the quality of digitized file.
- Condition evaluation
    - Objects are pulled and reviewed by conservation staff to determine whether or not digitization is safe for each object.
    - Some objects need to be rehoused or unfolded in order for digitization to occur safely.
    - Special handling may be required for oversized objects (i.e., need at least two people to handle the item) or for extremely fragile items like onion skin maps.

    - Many objects are housed in Mylar and some of them are too fragile to remove from the housing. This determines how the items need to be scanned in order to avoid unnecessary reflections and distortions.
- Cataloging
    - The number of materials in the collection is unknown at the beginning of a project. The digitization process will determine the actual number of items, and limit the amount of handling required.
    - Many objects have not been cataloged to date.
        - Recording as much information about the object into the database for future cataloging is critical.
- Metadata creation
    - Objects that are ready for digitization are reviewed and named with a specific digital filename. The filename is recorded into a database with the corresponding call number, unique data about the item, and information about the digitization including: what date it was scanned, who scanned it, and what equipment scanned it.
    - For objects that have been previously cataloged, the LCCN is also recorded.
    - Filenames contain both letters and numbers and will signify both the abbreviated collection name and a unique number. Copy number and/or bibliography numbers have also been appropriated within the digital filename.
    - At the time of the scan, a signifying character is appended to each filename indicating that the scan is an "unprocessed" scan.
    - Because many items are oversized and cannot be scanned in one image, a grid pattern is determined and carried out with letter or character identifiers appended to the unprocessed filename.
    - Technical metadata is created and saved from the time of the scan.
- Production scheduling
    - Staff resources are limited and the digitization process is largely handled by one person.
    - After a group of objects are pulled, one person names it, digitizes it, and performs post-processing as needed.
    - Most items are fragile and although an extra hand would be helpful during the handling of the item, the process is mainly individual.
    - Other digitization efforts are recorded into the same databases, and having one person responsible for the majority of the digitization of the specific object limits confusion and mistakes.
- Digitization prep
    - Filenames are created and recorded into both the database and onto the storage housing of the object.
    - Because items are scanned at the same resolution, grouping of objects may occur due to type of material.
        - Example – there are multiple onion skin objects that need a backing while scanned.
        - Example – different collections are being scanned in the same session and are therefore grouped according to filename.
- Digitization
    - Determine the appropriate spatial resolution for the project.
    - Determine the appropriate ICC color space (sRGB, Adobe 1998, etc.).
    - Small targets are placed within each unprocessed scan to monitor and track the color accuracy of each image. This target will be cropped out of the final image in order to ensure no confusion for the end-user. These targets measure the digitizing environment of the actual material and not the equipment.
    - Once a day, a target used to measure the equipment is also scanned. This is used by digitization staff to track and monitor the scanning equipment's performance through time. Results from testing are used to create custom color profiles, to fine-tune sharpening and resolution settings, and to monitor any variances in quality.
    - Items are generally scanned in the numerical order that they are named to streamline the process.
    - Many of the items are housed in Mylar and need to be positioned accordingly on the scanner bed to avoid glare and unwanted reflections.

- Processes of digitization are grouped together (i.e., all naming and inputting into database at once, and then all scanning) for efficiency.
- Once object has been scanned, the housing folder will be labeled with a "scanned" label to alert all future staff and patrons of an already digitized object.
- Post-processing includes cropping and de-skewing as well as minimal tonal adjustment, done within the lightness channel in L*a*b* mode.
- Minimal sharpening is done within the scanner software at the time of scanning and not applied in post-processing.
- The ability to read information accurately and in the way that the original is represented is the main concern here.
- Since many objects are so large and cannot be scanned in one image, multi-part scans also need to be combined into a presentation file.
  - Merging files is unacceptable because of the inevitable loss of information.
  - Files are generally "placed" next to their counterparts allowing blank space to occur between each part and therefore alerting the user that the files have been combined.

- Quality review
  - All images are reviewed as thumbnails in Adobe Bridge.
  - Selected images are reviewed 1:1 in Photoshop, and metadata verified.
  - Review is done on color calibrated monitors.
- Archiving
  - Archiving digitized material is done both on a personal staff-member's level as well as on the Library level.
  - Much of the scanning equipment does not have access to the local intranet and shared drives
    - Images will be scanned to a staff-member's portable hard drive.
    - Images are also copied to the staff-member's computer.
    - As both an extra back-up copy and also an access copy, the images are copied to the library's processing space.
  - Because little processing has been done to the images, the processed version of the scanned image is an archival copy.
    - The un-processed version is usually saved until the processed image is reviewed and published.
- Publishing
  - Different versions and file types of the same image are created in order to be accessible to all patrons and users.
    - These derivatives include thumbnail JPEG, TIFF, and GIF.
    - Only the processed copy (with minor adjustments and combined when necessary) is published.

Example Workflow #2

The following is an example of digitizing "on-demand." When there is a group of objects that needs to be digitized (either for an internal or external request), scanning requirements and goals are specially created. There is no distinct end point and the project always varies in terms of materials and goals.

- Selection of material
  - Objects range in size and material and different resolutions, lighting set-ups, and/or equipment are needed to capture both detail and color accurately.
    - Example – engravings need to be digitized at higher resolutions in order to avoid moiré.
    - Example – negatives and other transmissive materials will need backlighting either with a light box on a copy stand or a particular setting on a flatbed scanner.
    - Example – stereographs cannot be scanned on a flatbed scanner because of the curve and must be scanned on a copy stand set-up. They may also require special lighting set-ups to avoid distracting shadows and reflections.
  - Similar materials are sorted into groups to be scanned at the same time with the appropriate device.
  - Objects are reviewed individually to determine appropriate resolution and equipment set-up.

- Condition evaluation
    - Objects have been requested for digitizing and have therefore been reviewed by reference or curatorial staff beforehand.
    - A “shooting list” of needed materials is created, alerting the digitization specialist of specific details regarding each object (i.e., resolution, cropping).
- Cataloging
    - Most objects are cataloged ahead of time.
- Metadata creation
    - The “shooting list” is referred to again for specific digital filenames (digital ID’s).
    - All digital ID’s are unique to the division and may contain specific information about the type of file, the collection that the file is in, or even the bit-depth of the file.
    - Technical metadata is created and saved from the time of the scan.
- Production scheduling
    - Some materials warrant a helper to assist with handling items.
        - This allows the scanner to input filenames and focus on image quality.
    - Many other materials are scanned and post-processed by the same individual.
    - Other processes such as file naming, quality review, derivative making and publishing can be done for multiple files and groups of objects by one individual.
- Digitization prep
    - Items from shooting list are pulled and grouped into “like objects.”
        - Example – all prints that need to be scanned at 300ppi are grouped together and all engravings that need to be scanned at 600ppi are grouped together.
        - Example – photographic prints that are highly reflective and need special lighting set-ups to avoid glare are grouped together.
- Digitization
    - Once every scan session (or when a resolution is changed), a target used to measure the equipment is scanned. This is used by digitization staff to track and monitor the scanning equipment’s performance. Results from testing are used to create custom color profiles and to fine-tune sharpening and resolution settings.
    - If digitization allows for downtime, operator can complete other tasks simultaneously, such as metadata input or post processing.
    - Post-processing includes cropping and de-skewing as well as minimal tonal adjustment.
    - Color accuracy and likeness to the original, in its current state, is the main concern in scanning here.
    - Minimal sharpening is done within the scanner software at the time of scanning and not applied in post-processing.
    - If a material “helper” is available, the scanner can focus on operating the software and appropriately changing quality settings.
    - Some objects are too large to be captured in one scan
        - These files can be merged with a process like Photomerge in Photoshop.
- Quality review
    - Materials that have been scanned in various sessions are reviewed by other staff members on color calibrated monitors.
        - Aspects such as excess dust and hair, skew, improper cropping and poor tonal quality are all reviewed.
        - The goal is only to show the material in its true form and not to improve upon the material. (i.e. removing permanent dust and hair)
- Archiving
    - Archiving digitized material is done both on a personal staff-member’s level as well as on the Library level.
    - Much of the scanning equipment does not have access to the local intranet and shared drives
        - Images will be scanned to a staff-member’s portable hard drive.
        - Images are also copied to the staff-member’s computer.
        - As both an extra back-up copy and also an access copy, the images are copied to the Library’s processing space.
    - Because little processing has been done to the images, the processed version of the scanned image is an archival copy.

        - The un-processed version is usually saved until the processed image is reviewed and published.
- Publishing
    - Different versions and file types of the same image are created in order to be accessible to all patrons and users.
        - These derivatives include thumbnail JPEG, TIFF, and GIF.
        - Only the processed copy (with minor adjustments) is published.
        - In some cases, merged files are presented as the final copy along with its various uncombined parts, making file sizes more manageable when needed.

Small scale project workflow

Small scale projects follow the same steps defined above, but are adjusted to fit the realities of a smaller project. Generally, that means there may be only one person tasked with performing all of the steps, on one computer and with limited resources. Production efficiency that comes with the use of expensive digitization equipment is traded for the affordability of lesser, or slower equipment, and the realization that some digitization tasks cannot be done. In minimalist form, this might be a flatbed scanner and a single operator working with free software on a single computer. This entire system can be assembled for a few hundred dollars, even less with creative acquisition strategies.

Stepping up to a digital camera moves the cost of a workable system to a few thousand dollars, but opens up many more possibilities for digitization of various materials, including dimensional objects.

Small scale projects using basic tools should easily achieve FADGI level 2, and may achieve higher FADGI levels. Scanners and cameras vary dramatically in their capabilities. Careful selection and implementation, using DICE testing as a guide, can reward you with a high quality capability at very low cost.

Sample Image Processing Workflow

The following provides a general approach to image processing that should help minimize potential image quality defects due to various digital image processing limitations and errors. Depending on the scanner/digital camera, scan/capture software, scanner/digital camera calibration, and image processing software used for post-scan adjustment and/or correction, not all steps may be required and the sequence may need to be modified.

Fewer steps may be used in a high-volume scanning environment to enhance productivity, although this may result in less accurate tone and color reproduction. Scan a target, adjust controls based on the scan of the target, and then use the same settings for all scans - this approach should work reasonably well for reflection scanning, but will be much harder to do when scanning copy negatives, copy transparencies, original negatives, and original slides/transparencies.

Consider working in high-bit mode (48-bit RGB or 16-bit grayscale) for as much of the workflow as possible if the scanner/digital camera and software is high-bit capable, and the computer used has enough memory and speed to work with the larger files. Conversion to 24-bit RGB or 8-bit grayscale should be done at the end of the sequence.

Scanning

- Adjust size, scaling, and spatial resolution.
- Color correction and tone adjustment
- Follow established aimpoint guidance – remember there are always exceptions and you may need to deviate from the recommended aimpoints, or to adjust image based on a visual assessment and operator judgment.
- Recommended – use precision controls in conjunction with color management to achieve the most accurate capture in terms of tone and color reproduction.
- Alternative – if only global controls are available, adjust overall color balance and compress tonal scale to minimize clipping.
    - No sharpening or minimal sharpening (unsharp mask, applied to luminosity preferred).
    - Color profile conversion (might not be possible at this point, depends on scanner and software).
- Convert from scanner space to Adobe RGB 1998 for color images or Gray Gamma 2.2 for grayscale images.
- Generally, for color image profile conversion, use relative colorimetric rendering intent for near-neutral images (like most text documents), and perceptual rendering intent for photographic and other wide-gamut, high-saturation images.

Check the accuracy of the scan. You may need to adjust scanner calibration and control settings through trial-and-error testing to achieve best results.

Post-Scan Adjustment/Correction

Color profile assignment or conversion (if not done during scanning)

- Either assign the desired color space or convert the image from scanner space. Use the approach that provides best color and tone accuracy.
    - Adobe RGB 1998 for color images or Gray Gamma 2.2 for grayscale images.
- Generally for color image profile conversion, use relative colorimetric rendering intent for near-neutral images (like most text documents), and perceptual rendering intent for photographic and other wide-gamut, high-saturation images.

Color correction

- Follow aimpoint guidance - remember that there are always exceptions and you may need to deviate from the recommended aimpoints, or adjust the image based on a visual assessment and operator judgment.
- Use precision controls (levels recommended, curves alternative) to place and neutralize the black-point, place and neutralize the white-point, and to neutralize the mid-point. When color correcting photographic images, levels and curves may both be used.
- Alternative – if only global controls are available, adjust the overall color balance.

Tone adjustment

- Use precision controls (levels recommended, curves alternative) to adjust all three aimpoints in an iterative process. There are always exceptions and you may need to deviate from the recommended aimpoints, or adjust the image based on a visual assessment and operator judgment.
- Alternative – if only global controls are available, adjust contrast and brightness.

Crop and/or de-skew

Check image dimensions and resize

Sharpen

For color files, apply unsharp mask to luminosity information only. Version CS of Photoshop has the ability to apply unsharp mask to luminosity in high-bit mode. Sharpening should be done prior to the final conversion to 8-bits per channel.

Save file

The actual image processing workflow for small scale projects will depend on the originals being digitized, the equipment and software being used, the desired image parameters, and the desired productivity.

Directory Structure

Regardless of file name, files will likely be organized in some kind of file directory system that will link to metadata stored elsewhere in a database. Master files might be stored separately from derivative files, or directories may have their own organization independent of the image files, such as folders arranged by date or collection identifier, or they may replicate the physical or logical organization of the originals being scanned.

The files themselves can also be organized solely by directory structure and folders rather than embedding meaning in the file name. This approach generally works well for multi-page items. Images are uniquely identified and aggregated at the level of the logical object (i.e., a book, a chapter, an issue, etc.), which requires that the folders or directories be named descriptively. The file names of the individual images themselves are unique only within each directory, but not across directories. For example, book 0001 contains image files 001.tif, 002.tif, 003.tif, etc. Book 0002 contains image files 001.tif, 002.tif, and 003.tif. The danger with this approach is that if individual images are separated from their parent directory, they will be indistinguishable from images in a different directory.

Versioning

For various reasons, a single scanned object may have multiple but differing versions associated with it (for example, the same image prepped for different output intents, versions with additional edits, layers, or alpha channels that are worth saving, versions scanned on different scanners, scanned from different original media, scanned at different times by different scanner operators, etc.). Ideally, the description and intent of different versions should be reflected in the metadata; but if the naming convention is consistent, distinguishing versions in the file name will allow for quick identification of a particular image. Like derivative files, this usually implies the application of a qualifier to part of the file name. The reason to use qualifiers rather than entirely new names is to keep all versions associated with a logical object under the same identifier. An approach to naming versions should be well thought out; adding 001, 002, etc. to the base file name to indicate different versions is an option; however, if 001 and 002 already denote page numbers, a different approach will be required.

Naming Derivative Files

The file naming system should also take into account the creation of derivative image files made from the master files. In general, derivative file names are inherited from the masters, usually with a qualifier added on to distinguish the role of the derivative from other files (i.e., “pr” for printing version, “t” for thumbnail, etc.) Derived files usually imply a change in image dimensions, image resolution, and/or file format from the master. Derivative file names do not have to be descriptive as long as they can be linked back to the master file.

For derivative files intended primarily for Web display, one consideration for naming is that images may need to be cited by users in order to retrieve other higher-quality versions. If so, the derivative file name should contain enough descriptive or numerical meaning to allow for easy retrieval of the original or other digital versions.

Quality Management

Quality control (QC) and quality assurance (QA) are the processes used to ensure digitization and metadata creation are done properly. QC/QA plans and procedures should be initiated, documented and maintained throughout all phases of digital conversion. The plan should address all specifications and reporting requirements associated with each phase of the conversion project, including issues relating to the image files, the associated metadata, and the storage of both (file transfer, data integrity). Also, QC/QA plans should address accuracy requirements for and acceptable error rates for all aspects evaluated. For large digitization projects it may be appropriate to use a statistically valid sampling procedure to inspect files and metadata. In most situations QC/QA are done in a 2-step process- the scanning technician will do initial quality checks during production followed by a second check by another person.

Inspection of Digital Image Files

Inspection of image files involves two processes.

The first is a technical inspection of the file assuring correct imaging parameters were used when imaging, and that the file is valid. JHOVE is a software tool widely used by cultural heritage institutions to validate conformance to image technical specifications. The use of this tool is highly recommended as a part of an inspection program.

Information about JHOVE can be found here: http://jhove.openpreservation.org/

Image quality must then be visually inspected on a profiled graphics workstation by a trained technician, using the following procedure:

A visual review of thumbnails of all images should be done to assure completeness and consistency of the imaging. The initial review is followed by a detailed examination of a subset of the project. This visual evaluation of the images shall be conducted while viewing the images at a 1 to 1 pixel ratio or 100% magnification on the monitor.

We recommend, at a minimum, 10 images or 10 % of each batch of digital images, whichever quantity is larger, should be inspected for compliance with the digital imaging specifications and for defects in the following areas:

File Related

- Files open and display
- Proper format
  - TIFF, JPEG 2000
- Compression
  - Compressed if desired
  - Proper encoding (LZW, ZIP)
- Color mode
  - RGB
  - Grayscale
  - Bitonal
- Bit depth
  - 24-bits or 48-bits for RGB
  - 8-bits or 16-bits for grayscale
  - 1-bit for bitonal
- Color profile (missing or incorrect)
- Paths, channels, and layers (present if desired)

Original/Document Related

- Correct dimensions
- Spatial resolution
  - Correct resolution
  - Correct units (inches or cm)
- Orientation
  - Document- portrait/vertical, landscape/horizontal
  - Image- horizontally or vertically flipped
- Proportions/Distortion
  - Distortion of the aspect ratio

        - Distortion of or within individual channels
    - Image skew
    - Cropping
        - Image completeness
        - Targets included
    - Scale reference (if present, such as engineering scale or ruler)
    - Missing pages or images

Metadata Related - **see below for additional inspection requirements relating to metadata**

- Named properly
- Data in header tags (complete and accurate)
- Descriptive metadata (complete and accurate)
- Technical metadata (complete and accurate)
- Administrative metadata (complete and accurate)

Image Quality Related

- Tone
    - Brightness
    - Contrast
    - Target assessment – aimpoints
    - Clipping – detail lost in high values (highlights) or dark values (shadows) – not applicable to 1-bit images
- Color
    - Accuracy
    - Target assessment – aimpoints
    - Clipping – detail lost in individual color channels
- Aimpoint variability
- Saturation
- Channel registration
    - Misregistration
    - Inconsistencies within individual channels
- Quantization errors
    - Banding
    - Posterization
- Noise
    - Overall
    - In individual channels
    - In areas that correspond to the high density areas of the original
    - In images produced using specific scanner or camera modes
- Artifacts
    - Defects
    - Dust
    - Newton's rings
    - Missing scan lines, discontinuities, or dropped-out pixels
- Detail
    - Loss of fine detail
    - Loss of texture
- Sharpness
    - Lack of sharpness

        - Over-sharpened
        - Inconsistent sharpness
    - Flare
    - Evenness of tonal values, of illumination, and vignetting or lens fall-off (with digital cameras)

This list has been provided as a starting point, it should not be considered comprehensive.

Quality Control of Metadata

Quality control of metadata should be integrated into the workflow of any digital imaging project. Because metadata is critical to the identification, discovery, management, access, preservation, and use of digital resources, it should be subject to quality control procedures similar to those used for verifying the quality of digital images. Since metadata is often created and modified at many points during an image's life cycle, metadata review should be an ongoing process that extends across all phases of an imaging project and beyond.

As with image quality control, a formal review process should also be designed for metadata. The same questions should be asked regarding who will review the metadata, the scope of the review, and how great a tolerance is allowed for errors (if any, as errors can have a deleterious effect on the proper discovery and retrieval of digital resources).

Practical approaches to metadata review may depend on how and where the metadata is stored, as well as the extent of metadata recorded. It is less likely that automated techniques will be as effective in assessing the accuracy, completeness, and utility of metadata *content* (depending on its complexity), which will require some level of manual analysis. Metadata quality assessment will likely require skilled human evaluation rather than machine evaluation. However, some aspects of managing metadata stored within a system can be monitored using automated system tools (for example, a digital asset management system might handle verification of relationships between different versions of an image, produce transaction logs of changes to data, produce derivative images and record information about the conversion process, run error detection routines, etc.). Tools such as checksums (for example, the MD5 Message-Digest Algorithm) can be used to assist in the verification of data that is transferred or archived.

Although there are no clearly defined metrics for evaluating metadata quality, the areas listed below can serve as a starting point for metadata review. Good practice is to review metadata at the time of image quality review. In general, we consider:

- *Adherence to standards set by institutional policy or by the requirements of the imaging project.*

Conformance to a recognized standard, such as Dublin Core for descriptive metadata and the NISO Data Dictionary – Technical Metadata for Digital Still Images for technical and production metadata, is recommended and will allow for better exchange of files and more straightforward interpretation of the data. Metadata stored in encoded schemes such as XML can be parsed and validated using automated tools. However, these tools do not verify accuracy of the content, only accurate syntax. We recommend the use of controlled vocabulary fields or authority files whenever possible to eliminate ambiguous terms, or the use of a locally created standardized terms list.

- *Procedures for accommodating images with incomplete metadata.*

Procedures for dealing with images with incomplete metadata should be in place. The minimal amount of metadata that is acceptable for managing images (such as a unique identifier, or a brief descriptive title or caption, etc.) should be determined. If there is no metadata associated with an image, does this preclude the image from being maintained over time?

- *Relevancy and accuracy of metadata.*

How are data input errors handled? Poor quality metadata means that a resource is essentially invisible and cannot be tracked or used. Check for correct grammar, spelling, and punctuation, especially for manually keyed data.

- *Consistency in the creation of metadata and in interpretation of metadata.*

Data should conform to the data constraints of header or database fields, which should be well defined. Values entered into fields should not be ambiguous. Limit the number of free text fields. Documentation such as a data dictionary can provide further clarification on acceptable field values.

- *Consistency and completeness in the level at which metadata is applied.*

Metadata is collected on many hierarchical levels (file, item, series, collection, etc.), across many versions (format, size, quality), and applies to different logical parts (item or document level, page level, etc.). Information may be mandatory at some levels and not at others. Data constants can be applied at higher levels and inherited down if they apply to all images in a set.

- *Evaluation of the usefulness of the metadata being collected.*

Is the information being recorded useful for resource discovery or management of image files over time? This is an ongoing process that should allow for new metadata to be collected as necessary.

- *Synchronization of metadata stored in more than one location.*

Procedures should be in place to make sure metadata is updated across more than one location. Information related to the image might be stored in the TIFF header, the digital asset management system, and other databases, for example.

- *Representation of different types of metadata.*

Has sufficient descriptive, technical, and administrative metadata been provided? All types must be present to ensure preservation of and access to a resource. All mandatory fields should be complete.

- *Mechanics of the metadata review process.*

A system to track the review process itself is helpful. This could be tracked using a database or a folder system that indicates status.

Specifically, we consider:

- *Verifying accuracy of file identifier.*

File names should consistently and uniquely identify both the digital resource and the metadata record (if it exists independently of the file). File identifiers will likely exist for the metadata record itself in addition to identifiers for the digitized resource, which may embed information such as page or piece number, date, project or institution identifier, among others. Information embedded in file identifiers for the resource should parallel metadata stored in a database record or header. Identifiers often serve as the link from the file to information stored in other databases and must be accurate to bring together distributed metadata about a resource. Verification of identifiers across metadata in disparate locations should be made.

- *Verifying accuracy and completeness of information in image header tags.*

The application Bridge in Adobe Photoshop CS can be used to display some of the default header fields and IPTC fields for quick review of data in the header. However, the tool does not allow for the creation or editing of header information. Special software is required for editing header tags. The following link provides a comparison of metadata editors.

https://en.wikipedia.org/wiki/Comparison_of_metadata_editors

- *Verifying the correct sequence and completeness of multi-page items.*

Pages should be in the correct order with no missing pages. If significant components of the resource are recorded in the metadata, such as chapter headings or other intellectual divisions of a resource, they should match up with the actual image files. For complex items such as folded pamphlets or multiple views of an item (a double page spread, each individual page, and a close-up section of a page, for example), a convention for describing these views should be followed and should match with the actual image files.

- *Adherence to agreed-upon conventions and terminology.*

Descriptions of components of multi-page pieces (i.e., is “front” and “back” or “recto” and “verso” used?) or descriptions of source material, for example, should follow a pre-defined, shared vocabulary.

Testing Results and Acceptance/Rejection

If more than 1% of the total number of images and associated metadata in a batch, based on the randomly selected sampling, are found to be defective for any of the reasons listed above, the entire batch should be re-inspected. Any specific errors found in the random sampling and any additional errors found in the re-inspection should be corrected. If less than 1% of the batch is found to be defective, then only the specific defective images and metadata that are found should be redone.

Storage Recommendations

We recommend that master image files be stored on hard drive systems with a level of data redundancy, such as RAID drives, rather than on optical media, such as CD-R or DVD. An additional set of images with metadata stored on an open standard tape format (such as LTO) is recommended (CD-R or DVD as backup is a less desirable option), and a backup copy should be stored offsite. Regular backups of the images onto tape from the RAID drives are also recommended. A checksum should be generated and should be stored with the image files.

We do not recommend that CD-ROMs or DVDs be used as a long-term storage medium. However, if images are stored on these media, we recommend using high quality or "archival" quality media that has passed ISO 10995 certification. The term "archival" indicates the materials used to manufacture the CD or DVD (usually the dye layer where the data is recording, a protective gold layer to prevent pollutants from attacking the dye, or a physically durable top-coat to protect the surface of the disk) are reasonably stable and have good durability, but this will not guarantee the longevity of the media itself. All disks need to be stored and handled properly. We do not recommend using inexpensive or non-brand name CDs or DVDs, because generally they will be less stable, less durable, and more prone to recording problems.

All disks need to be stored and handled properly. Two (or more) copies should be made. One copy should not be handled, and should be stored offsite. Most importantly, a procedure for migration of the files off of the disks should be in place. In addition, all copies of the CDs or DVDs should be periodically checked using a metric such as a CRC (cyclic redundancy checksum) for data integrity. For large-scale projects or for projects that create very large image files, the limited capacity of CD storage will be problematic. DVDs may be considered for large projects, however, DVD formats are not as standardized as the lower-capacity CD formats, and compatibility and obsolescence is likely to be a problem.

Digital Repositories and the Long-Term Management of Files and Metadata

Digitization of archival records and creation of metadata represent a significant investment in terms of time and money. It is important to realize that the protection of these investments will require the active management of both the image files and the associated metadata. Storing files to CD or DVD and putting them on a shelf will not ensure the long-term viability of the digital images or the continuing access to them.

We recommend digital image files and associated metadata be stored and managed in an ISO 16363 compliant digital repository, see https://en.wikipedia.org/wiki/Trustworthy_Repositories_Audit_%26_Certification.

Resources

Introduction

General Resources

Moving Theory into Practice, Cornell University Library, http://www.library.cornell.edu/preservation/tutorial

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials, Humanities Advanced Technology and Information Institute, University of Glasgow, and National Initiative for a Networked Cultural Heritage, available at http://www.ninch.org/guide.pdf

Project Management Outlines

"NDLP Project Planning Checklist," Library of Congress, http://lcweb2.loc.gov/ammem/prjplan.html

DPC Digital Preservation Handbook 2.0

http://www.dpconline.org/advice/preservationhandbook