



OCLC
21,1

40

FEATURES

Practical quality control procedures for digital imaging projects

Jenn Riley and Kurt Whitsel

Indiana University, Bloomington, Indiana, USA

Abstract

Purpose – Formal quality review processes are a necessary part of any digital imaging workflow. This article illustrates a set of quality review processes implemented in the Indiana University Digital Library Program's Digital Media and Image Center.

Design/methodology/approach – A methodology for automatic batch review of large numbers of images is presented, along with rationale and procedures for supplemental visual review. The initial stages of an effort to further automate and centralize image quality control at Indiana University are described.

Findings – Automation of checks for objective image criteria, together with formal visual review of a sample of digitized images, is an effective means of implementing a quality review process.

Originality/value – The methodologies described can be used as a model for other institutions performing digital imaging projects of any size.

Keywords Digital libraries, Quality control, Operations and production management, Document image processing

Paper type General review

Most people who plan digital imaging projects are aware of the important role a quality review process plays in the success of the project. These same people are often much less sure how to integrate such a process into their daily workflow. Guides to quality review of digital images, such as Oya Rieger's chapter "Establishing a Quality Control Program" in *Moving Theory into Practice* (Rieger, 2000) and Franziska Frey's contribution to RLG's *Guides to Quality in Visual Resource Imaging*, "Measuring Quality of Digital Masters" (Frey, 2000) are often written from a highly theoretical perspective. While those planning digital imaging projects must have at the very least a basic understanding of the issues raised in these influential quality guides, many practitioners, especially in libraries and archives, have difficulty translating the ideas presented in these works into actual procedures to be used in daily workflows.

Digitization is a habitual process. Any quality review a digitizer might do before saving an image just created would be insufficient, just as delete confirmation boxes are insufficient on their own (Raskin, 2000, p. 22). It is extremely important, therefore, that the quality review process be performed by people *other* than the ones doing the original digitizing. Digitizing staff in the University environment, who are often student employees, should be trained to understand enough about the digitization process to be able to catch many quality problems as they happen. Ultimately, however, the responsibility for ensuring images meet project specifications lies with project planners and supervisors.



A quality review process is necessary, both for images scanned in-house and for those outsourced to a digitization vendor. In both cases, project staff must ensure images meet project specifications. Contracts with vendors for digitization services generally include provisions for rejection and re-digitization of images that do not meet agreed-on specifications, and institutions contracting with vendors must have a means of identifying these images quickly and accurately. When derivative images are purchased from the vendor in addition to the master image files, these derivatives should be subject to quality review as well.

At the Digital Media and Image Center (DMIC) (www.dlib.indiana.edu/dmic/) in the Indiana University Digital Library Program (www.dlib.indiana.edu/), we have developed a practical quality control workflow that has been applied in some way to most of our digital imaging projects since 2001. Our process consists of a combination of automated checks of all images produced to test objective image quality criteria, and visual checks of a sample of images produced to test subjective image quality criteria.

Automated checks

The first stage of the quality control process is an automated check of all files digitized in a specified time frame. How frequently an automated check is run varies among digitizing projects. If we are outsourcing digitization, we run the automated checks whenever we receive a batch of files from the vendor. If we are digitizing an archival collection, neatly organized into discreet boxes that take us a day or two each to scan, we run an automated check on all files scanned when a box is complete. If the collection is more continuous, without natural breaking points, we will run automated checks weekly or even daily. For some collections, all images located in a temporary “holding” directory are checked, then moved to a permanent location after they pass the review. For others, only certain images within a directory are checked at a given time.

The core of the automated checks, as they are currently implemented, are Perl scripts running on a server running the Unix (specifically, IBM’s AIX brand of Unix) operating system. Images are saved either to this server or to a Windows server accessible via an NFS mount to the Unix server where the quality review scripts reside. As the vast majority of the images we check are TIFF images, the Perl scripts call the Unix utility `tiffdump`[1] in order to evaluate metadata stored in the TIFF Image File Directory (IFD, also commonly referred to as the “TIFF header”) on each image to be checked[2]. `Tiffdump` outputs each IFD entry on a separate line, for example:

```
ImageLength (257) SHORT (3) 1 < 3032 >
```

Each line of the `tiffdump` output is parsed with this regular expression, which saves the TIFF tag name, number, length, and value for later use:

```
/([a-zA-z]*)s\((\d*)\)s[A-Z]*s\(\d\)s(\d*) < (.*) > /
```

The script then determines if the tag number on the current line is one that requires a certain value for the collection currently being checked, for example, that the resolution unit tag code represents pixels per inch, compares this value to the expected value, and reports any problems found. Occasionally, the appropriate value for a tag is only known in context of other tags (e.g. when the long side of an image needs to be a certain number of pixels, but which is the long side is not known), and these are saved for validation after all lines of the `tiffdump` output for this image have been handled. The script finishes by ensuring all tags it expected to see were actually found.

Currently, we create a new Perl script to check each new collection we digitize, and each subset within a collection whose images have different specifications. Our collections tend to have some checks in common, so code for a previous collection is re-used as a basis for each new collection. These common features include:

- (1) The structure of the filename is correct according to project specifications.
- (2) File format is correct according to project specifications (usually TIFF).
- (3) Compression scheme is correct according to project specifications (usually uncompressed).
- (4) The image is in little-endian (Intel PC) byte order.
- (5) The resolution unit tag is set to pixels per inch.
- (6) Image resolution is correct according to project specifications, often measured both as resolution and long-side dimension.
- (7) Bit depth (bits per sample and samples per pixel) is correct according to project specifications.
- (8) Based on whether the photometric interpretation is grayscale or RGB:
 - bits per-sample and samples per-pixel are correct according to project specifications; and
 - the image has the correct embedded color profile according to project specifications.
- (9) Other TIFF tags required for a valid TIFF image are present with reasonable values.

We also have a number of checks that are customized for each collection we digitize. For outsourced projects, we often ask vendors to add some metadata to the TIFF IFD, including values in TIFF tags 257 (DocumentName), 306 (DateTime), 315 (Artist), and 33432 (Copyright). For printed materials, we can often expect that all images should be in portrait orientation, so we check that the image height is greater than the image width. For multi-page items, such as sheet music, we check that the filenames indicate no missing pages and that each page of a given item has the same pixel dimensions. We also frequently check images present in a given directory against a log indicating which images have been scanned to ensure all items were scanned and no extraneous files are present. All problems found are documented for record-keeping purposes and each is assigned to a staff member for investigation and resolution. Once all problems identified in both the automated and visual reviews have been addressed, the automated check is run a second time to confirm no additional problems have been introduced.

These automated checks are extremely effective mechanisms for catching image quality problems before they escalate. It is extremely easy for small details such as a digit in the resolution box in the scanning software to get changed accidentally, or for a set of radio buttons telling Photoshop whether or not to use compression when saving a TIFF image to get toggled to an improper setting, and for the scanner operator not to notice in a timely manner. With automated checks of every image scanned for a project, these sorts of problems are caught quickly and ultimately eliminated from the final product.

Visual checks

While the automated checks described above are extremely effective, they cannot catch all possible quality problems in a digital imaging project. A staff member who supervises the digitization team performs visual review of a sample of images to supplement the automated check. The images are viewed on a monitor in a standard viewing environment[3] at a magnification such that one screen pixel displays one image pixel. The number of images checked varies among digitization projects, but for projects digitized in-house we generally check around 10 percent of the total number of images scanned. Digitizing staff track any anomalies in original items that might contribute to image quality problems. These items are visually inspected in addition to the random sampling of the entire set of images.

A visual check generally involves verifying any or all of the following information not available with automated methods:

- The filename assigned to the image is correct based on its identification number and project specifications.
- Color matches the original or is somehow “improved”, according to project specifications.
- Decision on color depth was made correctly when it varies within a collection.
- The digital image has the correct orientation.
- Image skew is within tolerance of project specifications.
- Image border is within tolerance of project specifications.
- No physical matter (dust, hair, specks of paper) is present in the scan.
- No digital artifacts appear in the image.
- Pages appear in the correct order for multi-page items.
- Metadata recorded in the scan log is correct for the item.

Examples of subjective image quality problems

Some projects require other visual checks in addition to these common ones (Figures 1-3). For film scanning projects, we must verify that the film was placed in the scanner with the base (non-emulsion) side facing the scanning lens. When scanning multi-page items, we generally do not scan blank pages. Sometimes, it is important for us to know the location of these blank pages, so the visual quality review verifies that these have been recorded correctly. We also verify that no items were skipped between digitization shifts, and that all items scanned have been returned to their proper order and are ready to be returned to their collection manager. As with the automated checks, all problems found are documented for record-keeping purposes, then assigned to a staff member for investigation and resolution.

The visual review certainly does find image quality problems we would not otherwise find until an alert end-user reports them to us; however, it is not practical to do visual review on every image we scan. Instead, reviewing a reasonably-sized sample of images allows us to find some of these problems, but more importantly allows us to identify recurring errors so that we can find ways to prevent them.

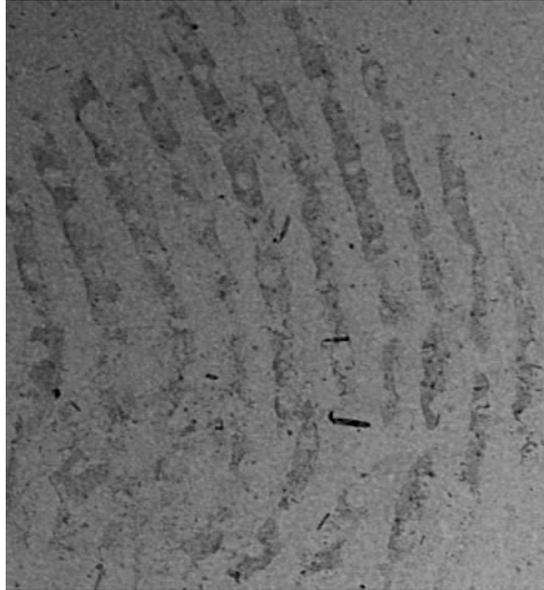


Figure 1.
Fingerprint on film

Moving to a centralized system

The methods described have been effective to date, but they don't work in all cases. The Digital Library Program at Indiana University occasionally hosts content for other University departments that do their own digitizing. These departments do not necessarily have dedicated staff to perform quality review processes, so it was necessary for us to find ways of further automating the process to allow these departments to benefit from quality control with a smaller time investment.

To meet these needs, we developed an image validation and processing system with a web-based interface that departments use for delivery of their digitized images to the Digital Library Program. The system does the same sorts of quality control and validation checks that our Perl scripts do for collections we digitize ourselves. In addition, the image processing system creates derivative images from the ingested master images using ImageMagick (www.imagemagick.org), and archives the master images into Indiana University's Massive Data Storage Service (<http://storage.iu.edu/mdss.html>). We are beginning the process of building a complementary tool to allow departments who use the system to retrieve master images on demand.

The image processing system is built in Perl and uses XML configuration files which allow the same application to be used for a number of different image collections. These XML configuration files are parsed and read using the Perl module, XML::Simple.

We have tied the application into Indiana University's Central Authentication Service so students and employees access the application using their existing IU network password. In order to allow for multiple users to work simultaneously, we needed to have a separate configuration for each workstation, for example:

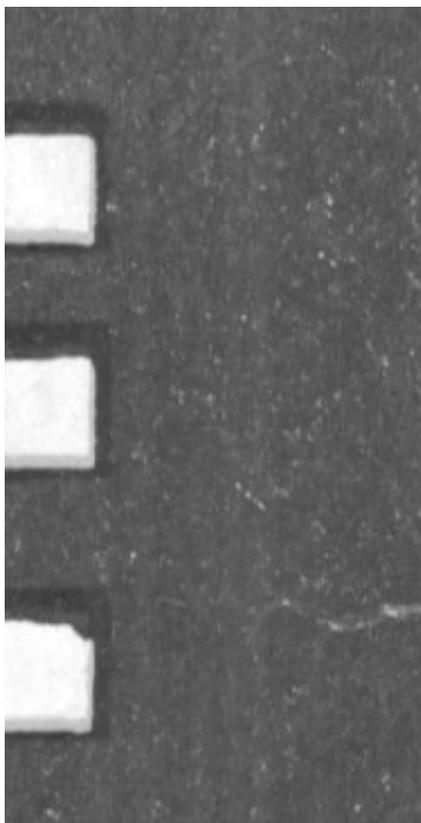


Figure 2.
Scanner malfunction

```
<workstation >  
<stationID > rygel < /stationID >  
<stationname > Rygel < /stationname >  
< !—directory where images are uploaded to, to await processing -- >  
< tiffdir > /digitize/imageproc/upload/slidelib/slidews1 < /tiffdir >  
< !—directory where derivatives are deposited after they are created -- >  
< derivativedir > /digitize/imageproc/dido/derivatives/< /derivativedir >  
< !—directory where master image files are deposited after derivative creation.  
In our case, this is a waiting directory where another process logs and moves  
files to the IU Mass Storage System -- >  
< processeddir > /digitize/hpss/dido/tiff/< /processeddir >  
< /workstation >
```

Once a user has logged in and chosen which workstation he is using, the application will list all of the files found inside that workstation's specified upload directory. The



Figure 3.
Newton Rings on film

user then chooses from a drop down list of possible preset processing rules for derivatives created from that image file. These rules control derivative image specifications such as maximum dimensions, a string to add to the derivative's filename, sharpening level, compression level, and derivative file format. Processing rules can differ for RGB color or grayscale originals, and we often create multiple derivatives for each uploaded master image, as seen here:

```
<rule >
<ruleID > callig </ruleID >
<rulename > Calligraphy </rulename >
<dimension >
<suffix > full </suffix >
<geometry > 1024x1024 </geometry >
<color > +profile iptc -sharpen 1 -quality 90 -format jpg </color >
<grayscale > +profile iptc -sharpen 1 -quality 90 -format jpg </grayscale >
</dimension >
<dimension >
<suffix > screen </suffix >
<geometry > 800x800 </geometry >
<color > +profile iptc -sharpen 1 -quality 90 -format jpg </color >
<grayscale > +profile iptc -sharpen 1 -quality 90 -format jpg </grayscale >
```

```
< /dimension >  
< dimension >  
< suffix > thumb < /suffix >  
< geometry > 150x150 < /geometry >  
< color > +profile iptc -sharpen 1 -quality 90 -format jpg < /color >  
< grayscale > +profile iptc -sharpen 1 -quality 90 -format jpg < /grayscale >  
< /dimension >  
< /rule >
```

While the image processing system has been extremely successful in codifying and enforcing quality standards for imaging projects not directly controlled by the Digital Library Program, it is not a complete quality control solution. We encourage departments using the system to supplement its use with a visual review of some percentage of their digitized images, and are looking for ways to facilitate this within the image processing system.

We are currently evaluating the success of the image processing system, with an eye toward moving some or all of the automated quality review for images created in the Digital Library Program currently checked with standalone Perl scripts to the centralized system. While the system is configurable to a certain extent, it is not currently able to handle all of the collection-specific custom checks described above. We are investigating to what extent we can and should do to extend the current capabilities of the system to accommodate more specialized checks, such as those required for multi-page items.

Conclusion

Having a complete quality review process in place is no substitute for adequate training and ongoing supervision of digitizing staff. It is also only effective if image specifications have been well-defined and tested at the start of the project, based on a thorough understanding of the source material and digital imaging best practices. Similarly, all imaging devices should be thoroughly tested[4] and profiled[5] frequently throughout the duration of imaging projects.

The methods described here can be expanded or altered to work for different imaging environments. Automatic checking of color and tone reproduction could be done for projects where color bars or other standard targets are included in image files. Different software packages or the same packages on different platforms[6] could be used to accomplish the same goals. By automating checks of objective image quality criteria, more quality problems can be found and addressed more quickly, leading to an overall better product in the end. Automated checks, however, are not enough on their own. They must be used in concert with subjective review to achieve a complete quality control solution for digital projects.

Notes

1. Tiffdump is part of the Libtiff package (available at: www.libtiff.org/). Other utilities, such as ImageMagick's (www.imagemagick.org) identify – verbose or tiffinfo, also part of the libtiff package, could be used in a similar manner.

2. A listing of the entries in the TIFF IFD in numerical order can be found in Appendix A of the TIFF 6.0 specification, available at: <http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf>
3. In the DMIC, we use monitors calibrated to a color temperature of 5000°K and a Gamma value of 2.2. Room lighting is also 5000°K. Some digitization operations use monitors calibrated to the sRGB standard of 6500°K and a Gamma value of 2.2, however, the International Organization for Standardization generally recommends 5000°K over 6500°K when comparing images on a screen to hard-copy images. More information is available in Rieger (2000, pp. 67-70), ISO (2000), and ISO (2004).
4. For example, through measurement of the Modulation Transfer Function (MTF), as described in Williams (1998).
5. Color profiling is also referred to as characterization, and can also refer to the related process of device calibration. All of these methods make use of International Color Consortium (ICC) (www.color.org) profiles to some extent, and are achieved through use of color profiling software such as that from Monaco Systems (www.monacosys.com/) and Colorvision (www.colorvision.com/).
6. For example, versions of Perl (www.activestate.com/Products/ASPN_Perl/) and libtiff (gnuwin32.sourceforge.net/packages/tiff.htm) are available for Windows platforms.

References

- Frey, F. (2000), "Measuring quality of digital masters", *Guides to Quality in Visual Resource Imaging*, Council on Library and Information Resources, Washington, DC, available at: www.rlg.org/visguides/visguide4.html
- ISO (2000), "Viewing conditions – graphic technology and photography", *ISO 3664*, International Organization for Standardization, Washington, DC, Vol. 3664.
- ISO (2004), "Graphic technology – displays for colour proofing – characteristics and viewing conditions", *ISO 12646*, International Organization for Standardization, Washington, DC.
- Raskin, J. (2000), *The Humane Interface: New Directions for Designing Interactive Systems*, Addison-Wesley, Reading, MA.
- Rieger, O. (2000), "Establishing a quality control program", in Kenney, A. and Rieger, O. (Eds), *Moving Theory into Practice: Digital Imaging for Libraries and Archive*, Research Libraries Group, Mountain View, CA, pp. 61-83.
- Williams, D. (1998), "What is an MTF . . .and why should you care?", *RLG DigiNews*, Vol. 2 No. 1, available at: www.rlg.org/preserv/diginews/diginews21.html#technical