The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

Cost and benefit of quality control visual checks in large-scale digitization of archival manuscripts

Received 13 January 2013
Revised 29 January 2013
Accepted 3 February 2013

Joyce Chapman
Library Development, State Library of North Carolina, Raleigh, North Carolina, USA, and
Samantha Leonard
Media Services, High Point University Libraries, High Point, North Carolina, USA

Abstract

Purpose – The purpose of this paper is to provide much needed data to staff working with archival digitization on cost and benefit of visual checks during quality control workflows, and to encourage those in the field of digitization to take a data-driven approach to planning and workflow development as they transition into large-scale digitization.

Design/methodology/approach – This is a case study of a cost benefit analysis at the Triangle Research Libraries Network. Data were tracked on time spent performing visual checks compared to scanning production and error type/discovery rates for the consortial grant "Content, context, and capacity: a collaborative large-scale digitization project on the long civil rights movement in North Carolina".

Findings – Findings show that 85 percent of time was spent scanning and 15 percent was spent on quality control with visual checks of every scan. Only one error was discovered for every 223 scans reviewed (0.4 percent of scans). Of the six types of error identified, only half cause critical user experience issues. Of all errors detected, only 32 percent fell into the critical category. One critical error was found for every 700 scans (0.1 percent of scans). If all the time spent performing visual checks were instead spent on scanning, production would have increased by 18 percent. Folders with 100 or more scans comprised only 11.5 percent of all folders and 37 percent of folders in this group contained errors (for comparison, only 8 percent of folders with 50 or more scans contained errors). Additionally, 52 percent of all critical errors occurred in these folders. The errors in larger folders represented 30 percent of total errors, and performing visual checks on the large folders required 32 percent of all visual check time.

Practical implications – The data gathered during this research can be repurposed by others wishing to consider or conduct cost benefit analysis of visual check workflows for large-scale digitization.

Originality/value – To the authors' knowledge, this is the only available dataset on rate of error detection and error type compared to time spent on quality control visual checks in digitization.

Keywords Quality control, Visual checks, Large-scale digitization, Cost benefit analysis, Archives management

Paper type Case study

1. Introduction

While large-scale digitization of books has received significant attention in the past decade (Conway, 2011, p. 304), libraries and archives are also engaging increasingly in

Library Hi Tech
Vol. 31 No. 3, 2013
pp. 405-418

0737-8831
DOI 10.1108/LHT-01-2013-0002

large-scale digitization of manuscripts and other primary source documents, both in place of and alongside smaller, more highly curated primary source digitization projects. This article relates to large-scale digitization of archival manuscripts, not to be confused with large-scale digitization of books. As workflows are evaluated and adjusted to meet the needs of large-scale production and delivery, institutions are experimenting with computer programs and software to assist in the performance of quality control on digital files, with less reliance on visual checks performed by staff. While software can identify certain types of error that the human eye cannot, there are other types of error that are currently best identified by visual checks. But to what degree does the use of visual checks mean that errors – such as image skew (crookedness), unnecessary pages scanned, missing pages, file naming errors, image order, and image quality issues (including blurriness and readability issues) – are not published online? What is the cost and benefit in staff time and in errors caught pre-publication involved in visual checks? In 2012, the Triangle Research Libraries Network (TRLN) performed a cost benefit analysis around these questions. We tracked the type and quantity of errors occurring during scanning in a large-scale manuscripts digitization project, as well as how much time staff spent performing visual checks, and the tradeoff in production capacity. This article uses TRLN's IMLS-funded grant "Content, Context, and Capacity: A Collaborative Large-Scale Digitization Project on the Long Civil Rights Movement in North Carolina" (CCC) as a case study. By gathering data on the time spent scanning, time spent performing quality control, and the number of errors of different types corrected via visual checks, we hope to provide valuable data to colleagues in the field struggling to make cost and value-based decisions about workflows for large-scale digitization.

2. Background

Over the past decade digital storage has become cheaper, digitization technology has become more powerful, and user expectations for access to digital content have increased exponentially (Brown, 2011). In order to meet user demand, practitioners in various fields are taking advantage of these technological advances and increasingly turning to large-scale digitization. For archives working with manuscript primary sources, this usually means concentrating on describing and exposing large quantities of resources at the collection level instead smaller quantities of materials at the item level. A large-scale approach allows staff to put less emphasis on selecting and describing individual items and more emphasis on providing online access (Greene, 2010; Dietz and Ronallo, 2011; Erway and Schaffner, 2007; Southern Historical Collection, 2007). In large-scale digitization of manuscripts, collections or series are often digitized in full, instead of staff selecting a few particular objects from a collection or multiple collections to bring together around a certain topic. In this way, the large-scale approach seeks to replicate the physical reading room online by providing users with the full context for each item within a collection.

To date, large-scale manuscript digitization projects have largely chosen to make digital materials accessible via the traditional tool of the archival finding aid – the same tool used by researchers to search for and find analog materials. Large-scale digital manuscript materials are often viewable directly in finding aids by a digital "folder,"[1] whereas the approach to smaller more curated digital primary source projects has historically involved the creation of online portals developed solely for

those digitized materials. Such curated projects often include creation of item-level metadata for each digitized object, but large-scale digitization tends to provide only the same metadata available to users of physical collections: the information in the existing archival finding aid[2]. Large-scale manuscript digitization introduces many challenges. For example, how should archivists and librarians provide proper online access and incorporate large-scale digitization workflows into existing workflows? Staff need data on the cost and benefit of various processes in order to make informed decisions about allocating time and resources, and develop digital workflows that efficiently publish large quantities of materials online while continuing to meet user needs to the best of their ability.

The CCC project is a multi-year, collaborative large-scale manuscripts digitization project funded through the federal Institute of Museum and Library Services (IMLS) under the provisions of the Library Services and Technology Act (LSTA) as administered by the State Library of North Carolina, a division of the Department of Cultural Resources. The grant began in 2011 and with continuing funding will end in 2014. The CCC project was developed to test collaborative large-scale manuscripts digitization and develop shared workflows between the TRLN partners, which include the libraries of Duke University, North Carolina Central University, North Carolina State University, and the University of North Carolina at Chapel Hill. During the project the four university libraries will digitize approximately 400,000 items at their combined Digital Production Centers on the theme of the Long Civil Rights Movement[3] in North Carolina. The CCC digitized content is freely accessible through all four university library websites, as well as through the shared single search interface, Search TRLN. Digital materials are delivered through finding aids, and digital objects have no additional metadata beyond what can be automatically extracted from the finding aids via scripts or added through batch processes.

TRLN project staff at the Digital Production Center in the Carolina Digital Library and Archives at the University of North Carolina at Chapel Hill – one of the two Digital Production Centers responsible for the bulk of manuscript digitization for the CCC grant – tested and recorded data on visual checks for quality control of large-scale manuscript digitization in 2011-2012, which is presented in this article. By providing information on the amount of time involved in quality control visual checks per number of scans uploaded, the error detection rate in different types of visual checks, the number of errors of different type caught by visual checks, and the percentage of errors detected in different sized folders we hope to help other digitization professionals be able to better gauge the costs and benefits of implementing visual checks into large-scale digitization.

3. Literature review

Though there are many different quality control strategies, it is generally recommended that staff document standards, quality control guidelines, and management approaches[4]. One specific recommended strategy is maintaining workflow logs to track digitization tasks and create data for future evaluation (Anderson and Maxwell, 2004; Lee, 2001). Digital production workflow logs also ensure accountability between all parties, especially in a collaborative digitization project (Anderson and Maxwell, 2004). Quality control is an important part of large-scale digital production and preservation, for both books and primary sources. Conway

(2011, p. 294)argues that validating the quality of digital objects is essential not only to determining fitness for use, but to provide incentives for stakeholders who may contribute to a digital repository or donate materials to an archive.

Quality control comes in the form of both automated checks by computer programs or software and human visual checks. These two types of checks can identify some of the same types of error, but each is also able to identify errors that the other cannot. For example, visual checks might compare digital images to the analog original to check whether the file name assigned to the image is correct, if the color matches the original, if the digital image is in the correct orientation, that no physical matter is included in the image, and that the pages are in the correct order (Riley and Whitsel, 2005, p. 43). Automated checks might review checksums, file type, file naming, or resolution. Open-source programs such as ImageMagick can perform automated checks through the command-line of file names, format, byte order, resolution, bit depth, and more (Belfiore, 2012; Riley and Whitsel, 2005, p. 42). Compared to automated checks, visual checks are immensely time consuming. Mark McFarland notes the difficulty of creating a time-effective quality control workflow with visual checks, discussing "quality problems that took us a great deal of time to isolate and correct" (Library of Congress, 1996-1999)[5]. Despite the effort required, visual checks can uncover scanning errors that automated checks cannot and help identify patterns of errors. Riley and Whistel (2005, p. 43) suggest that visually "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".

The majority of documented manuscripts digitization quality control practices are not specific to large-scale digitization and many recommend visual checks on 100 percent of materials. For example, Anderson and Maxwell (2004, p. 85) recommend checking "each and every paragraph, sentence, word, and punctuation mark to ensure they exactly match their counterpart in the printed document". Documentation for non-large-scale digitization on the website of the University of Alabama Libraries recommends that two rounds of visual checks be performed on every scan; the first by the scanning technician and the second by another colleague, and the online documentation for the Georgia Digital Library also recommend checking each image visually. The University of Maryland recommends that 100 percent of images be visually reviewed for a first batch of digitized materials in a project, and thereafter if 100 percent cannot be achieved, a minimum of 10 percent should be visually reviewed. Alabama's two-person system is recommended by NARA as well, though NARA's guidelines do not require that 100 percent of images be checked. At a minimum, NARA guidelines suggest that 10 percent of each batch of digital images (or ten images, whichever is larger) should be visually reviewed. There are many proponents of sampling methodologies for visual checks as well: Stuart D. Lee recommends visual quality control checks be performed only on a sample of materials in his 2001 book. The University of Alabama's online guidelines for large-scale digitization call for visual spot checks on every twentieth scan[6].

There is little available data related to time required to perform visual checks or the quantity and type of errors detected through visual checks, despite the fact that quantitative analysis and benchmarking are valuable evaluation tools for digitization managers. Several projects in the wider field of large-scale digitization have gathered data on error rates and type of error, and one project in the field of manuscripts

digitization has published data on the time spent performing quality control. To our knowledge, no study other than the current one has produced data combining rates of error detection and error type with data on time spent performing visual checks.

Researchers at the University of Michigan School of Information are currently working on an IMLS-funded grant, "Validating Quality in Large-Scale Digitization." By analyzing books digitized by vendors and deposited in the HathiTrust, the group seeks to establish valid definitions of error, develop statistically valid methods for measuring error through visual checks, develop a multi-institutional, distributed quality assurance process for digitized books and journal content. The group developed both an error typology and a proposed error severity scale, categorizing different types of error and a numeric scale to categorize errors of different severity's effect on decipherability (Conway *et al.*, 2012). Findings from this study will be published in 2013[7], but a sample of preliminary data shared by principle investigator Paul Conway shows severe scanning errors occurring in less than half of 1 percent of scans for Google digitized books and serials published before 1923 in English (Conway, 2013a).

The second set of research documenting error rates for large-scale digitization is part of a 2010 report out of the Council on Library and Information Resources titled "The idea of order: transforming research collections for 21st century scholarship." While the report itself does not contain data on error rate detection, additional reference materials created by Gevinson (2010) over the course of the research do. In the summary of efforts to quantify current correctable errors with large-scale book digitization, Gevinson sampled hundreds of books digitized for three mass digitization projects. While he reports on numerous types of error rates that are unrelated to issues of scan quality or scanning error that are the subject of this paper, Gevinson also reports scan quality error rates for Google Book Search, Microsoft Live Search Books, and ACLS Humanities Ebooks. For example, during his sampling of Google Book Search, 16 percent of volumes were found to have missing pages in the digital copy, 5 percent had pages that were out of order, 10 percent had duplicate pages, 7.5 percent had 1 or more illegible pages that were not a fault of the original copy, 15 percent had pages that were partially cut off or obscured, 25 percent had pages in which letters were not sharp, and 16 percent had light/dark/blurry pages that would adversely affect Optical Character Recognition reading[8].

The only published data the authors were able to find on time collection for visual checks is a time study on manuscripts digitization by Deriddler *et al.* (2012). They tracked time for various digitization workflows – including quality control – for the University of Alabama's item-level digitization workflow as well as for a new large-scale digitization workflow. They then extrapolated the associated costs to compare the old and new digitization methods. DeRiddler *et al.* (2012) found that by the old workflows they spent ten minutes on quality control per each 100 scans versus six minutes per each 100 scans with the new workflow, which translates to savings of 86 cents per each 100 scans for their institution (DeRiddler *et al.*, 2012, p. 158).

In terms of unpublished data, various institutions doubtless retain in-house records of staff time spent on different stages of the digitization process, but this data is not publically available. Duke University is one such institution, and also a TRLN consortial partner. Their Digital Production Center was happy to share examples of quality control time data for manuscripts digitization with the authors. Duke's Digital

Production Center tracks the number of hours per month spent on about a dozen tasks related to digitization by each staff member, for example, collection analysis, conservation consultation, derivative creation, scanning, and quality control. Duke also tracks the number of scans that were reviewed during quality control, so like the University of Alabama study, data is available on how many scans are visually reviewed per unit of time spent performing review. However, neither of these institutions specifically tracks either the number of errors that are uncovered during visual checks or the type of error discovered. While data on errors is important to the type of cost benefit analysis the authors are attempting to pursue here, it is not necessarily useful from the management perspective of those tracking time for other administrative purposes.

4. Methodology and workflows

In order to understand the time commitments involved in performing visual checks within manuscripts large-scale digitization, the volume of errors corrected with visual checks, and the trade off in terms of time lost toward production rates, time data was tracked on visual checks and scanning and data was also gathered on the volume of error discovery for a three-month period (May 25, 2012 to August 29, 2012). During this time, one grant staff member and four students scanned materials at the University of North Carolina at Chapel Hill's Digital Production Center, while one staff and one student performed visual checks. When errors were found, the employee performing the visual check immediately fixes the errors, so the quality control time tracking includes time spent on correcting errors.

Manuscript scanning was performed almost entirely on Zeutschel 12000C Overhead Color scanners using Omniscan software. When performing quality control, the staff and students reviewed the following:

- image quality, including checking for color matching, physical matter or digital artifacts, and readability of print;
- image orientation, skew, and image order;
- missing or duplicated scans; and
- file naming, size, dimension, and type.

Data tracked for quality control visual checks included:

- dates the visual checks were performed;
- identification of the operator performing checks;
- filename of the digital folder being reviewed;
- quantity of digital scans in the folder;
- type of visual check performed (analog or digital check);
- amount of time spent on the check (in minutes); and
- quantity and type of errors found during check.

The total number of scanning hours was calculated based on employee and student work schedules and the total number of scans was tracked by technical metadata files automatically exported from the scanners. Data was analyzed in Microsoft Excel and charts were created with Microsoft Excel.

The visual check workflow used during the data collection period was performed on 100 percent of digital files. Additionally, over half of the physical folders and items within were compared to their digital surrogates, while the rest of the digital files were reviewed alone without their analog equivalents. To perform analog to digital checks, the student or staff member had the physical folder open on a desk while the digital file was open on a computer monitor in order to check for inconsistencies such as image order or missing scans. Other than checking for image order and missing scans, all quality control checks were the same between the analog and digital reviews. For consistency, throughout this paper we will call the comparison of analog to digital materials "analog checks" and will call the quality control of only the digital material "digital checks."

5. Findings and discussions

During the three-month data collection period, 40,624 pages associated with 705 physical manuscript folders from eight manuscript collections and three consortial partners were scanned and reviewed. During that same period 315 hours were spent on scanning and 57 hours were spent performing quality control[9]. Of the total time spent on scanning and on quality control (372 hours), 85 percent of all time was spent on scanning and 15 percent on quality control (see Table I). During quality control review, only 182 errors were identified, or one error for every 223 scans. In other words, an error was found for less than half of 1 percent of all scans (0.4 percent). These errors occurred in 88 folders (12.5 percent of all folders scanned), while 617 folders (87.5 percent) had no scanning errors. Scanning errors were committed by fulltime staff, student workers, and by those who were completely trained. The idea that errors cease to occur once scanning technicians are fully trained was not found to be true.

Errors were classified into six categories:

(1) image skew (crookedness) and rotation;

(2) unnecessary pages scanned (duplicates);

(3) missing pages;

(4) file naming errors;

(5) image order; and

(6) severe image quality issues (such as blurriness, lines through scans, and other readability issues).

Of these six types of error, half can cause serious user experience issues: file naming errors and missing pages both cause the item to not appear online, and if scan quality errors great enough to warrant correction occur, the image is not legible or viewable. We will refer to these three types of error as "critical errors." There were 58 critical errors, representing 32 percent of all errors. The remaining three types of error – comprising

Table I.
Time spent on scanning and quality control during data collection period

	Count	Hours of time spent on activity	Count per hour of time spent	Combined time (%)
Scans produced	40,624	315	129	85
Errors found with quality control	182	57	3.2	15

124 errors, or 68 percent of total errors – cause less serious problems for the user. We will refer to these as "non-critical errors." These include skewed or incorrectly rotated images, unnecessary pages (duplicate scans), and image order. Image order and skew errors – both non-critical errors – occurred more frequently than other types of errors, together representing 42 percent of all errors (see Figure 1 and Table II).

To put this into perspective as a cost benefit equation, if we consider only critical errors (missing images, scan quality, and file naming errors), then only one critical error was found for every 700 scans, or in one-tenth of 1 percent of all scans (0.1 percent), and only one critical error was found per roughly every hour of quality control visual checks performed. If all the time spent performing visual checks were instead spent on scanning, production would increase by around 18 percent.

A basic cost and benefit proposition could therefore be posed as follows:

- Which provides greater benefit and lower cost within the parameters of this project?
- The detection and correction of serious errors in 0.1 percent of scans and non-serious errors in 0.3 percent of scans.
- An 18 percent increase in production alongside undetected serious errors in 0.1 percent of scans and non-serious errors in 0.3 percent of scans.

This is an interesting question within the parameters of the CCC large-scale manuscripts digitization project. While digital production volume is usually highly valued by libraries, this particular project does not have strong incentives to increase production capacity due to issues such as surpassing the digital storage requested and allotted to the project, agreed upon materials preparation timelines at participating institutions, and pre-determined digital file transfer timelines. On the other hand, the

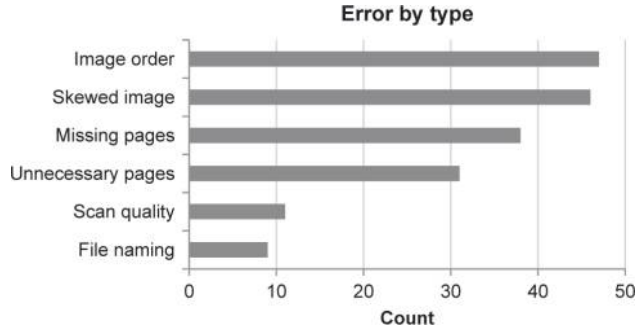

Figure 1.
Graph of error count by type

Type	Count	Critical?	Total (%)
Skew errors	46	No	25
Unnecessary pages	31	No	17
Missing pages	38	Yes	21
File-naming	9	Yes	5
Image order	47	No	26
Image quality	11	Yes	6

Table II.
Error by type and percentage of total errors

approach to large-scale digitization championed by the CCC grant does not support high-touch, item-level work in general. For example, the only metadata provided per digital object is the metadata that can be extracted by scripts from existing finding aids (collection or folder-level metadata). Within this framework, can we rationalize item-level quality control performed by humans instead of relying solely on the more limited error detections possible through automated processes, user-reported errors, or visual spot checks?

In terms of deciding whether to spot check or sample digital images visually, this study uncovered further information that may be of use. During the data collection period, we reviewed 451 folders using analog checks (64 percent of all folders reviewed) and 254 using digital checks (36 percent of all folders reviewed). Analyzed by number of scans contained within the folders, 69 percent of all scans were reviewed with analog checks and 31 percent with digital. Errors were detected during both types of check, and corresponded fairly evenly to the quantity of material reviewed per type of check: 75 percent of errors were detected during analog checks and 25 percent of errors were detected with digital checks. Perhaps more interesting is that 19 percent of folders reviewed with analog checks were found to contain an error, versus 11 percent of those reviewed with digital checks. It is therefore likely that analog checks are of greater assistance in detecting errors than digital checks and that some errors were overlooked during digital-only checks. For example, missing scans or image order errors are quite obvious in analog checks but not in digital checks. If spot checks or sampling are to be performed, we believe that a higher number of errors would be detected with analog to digital comparisons.

We also compared error rates to the quantity of scans per folder to determine whether more scanning errors occurred when scanner operators are working with larger folders. We hypothesized that scanning a larger folder increases the length of time that a scanning technician is performing a repetitive task before breaking to perform other tasks, and may contribute to more mistakes. We chose two break points to measure: folders containing more and less than 50 scans and folders containing more and less than 100 scans (see Table III).

We found that three-quarters of all errors occur in folders with more than 50 scans, and about half of all folders contained more than 50 scans. Only 8 percent of folders with 50 or more scans contained errors. A higher rate of errors per folder is found in those folders with 100 or more scans. While folders with 100 or more scans comprise only 11.5 percent of all folders, 37 percent of this group contained errors. Even so, the set of errors found in these folders represents only 30 percent of total errors. In terms of time, the 11.5 percent of all folders that contained 100 or more scans required 32 percent of all the visual check time over the three-month data collection period.

	Folders *n*	Folders with errors *n*	Total errors *n*	Folders with errors %	All errors %	Folders with errors %	All folders %
< = 50 pages	337	1	42	9	23	9	48
> 50 pages	366	60	140	16	77	8	52
< = 100 pages	621	71	128	11	70	4	88
> 100 pages	82	19	54	23	30	37	12

Table III.
Rate of errors for folders of varying sizes

We then examined critical error rates by folder size. We found that 52 percent of all critical errors occurred within the 11.5 percent of all folders that contained 100 or more scans. Of particular interest was that close to 60 percent of all image quality and missing pages errors were found in folders with 100 or more pages (Figure 2).

The higher rate of error found with large folders may indicate that scanning operators are more likely to make mistakes during longer periods of time performing repetitive tasks. In terms of implementing a sampling system for visual checks, reviewing larger folders more frequently than small folders would increase CCC's "bang for the buck" in terms of time spent on quality control per error discovered. It would also provide a higher rate of detection for critical errors than a simple percentage-based sampling of all folders. According to the data gathered in this case study, visually reviewing only folders with 100 or more scans would account for around half of all critical errors, and would take one-third the time of performing visual checks on every folder. However, the saying that "perfection is the enemy of good" should also be considered during this cost benefit analysis: the rate of critical error detected for the project is so low (one critical error for every 700 scans, or 0.1 percent of scans) that we must ask ourselves whether performing visual checks at all beyond the initial training period for new scanning technicians is a workflow that makes sense for our philosophy and approach to large-scale digitization.

There are aspects of visual checks on which it would be interesting to gather more data: for example, the current study did not track errors or time by type of manuscript material, such as pamphlets, correspondence, or newspapers. Such data could provide even more useful guidance to managers on where to focus visual check efforts.

6. Conclusions

In summary, we found that when scanning and visual checks were considered together, 85 percent of our time is spent scanning while a full 15 percent is spent on quality control. If all the time spent performing visual checks were instead spent on scanning, production would increase by around 18 percent. Despite the extensive time spent performing visual quality control, only one error was discovered for every 223 scans reviewed (0.4 percent of scans), and of all errors detected, only 32 percent caused critical user experience issues, or one critical error per each 700 scans (0.1 percent of scans). Folders with 100 or more scans comprised only 11.5 percent of all folders and 37 percent of folders in this group contained errors, while comparatively only 8 percent of folders with 50 or more scans contained errors. Additionally, 52 percent of all critical

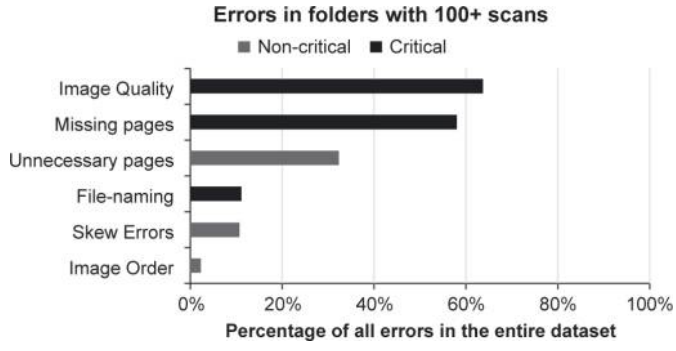

Figure 2.
Percentage of critical and non-critical errors in folders with 100 or more scans

errors were found in the 11.5 percent of folders with more than 100 scans, and performing visual checks on these large folders required 32 percent of all visual check time. Two scenarios that the Digital Production Center at UNC may consider for the CCC project moving forward include performing quality control visual checks only on folders with 100 or more items, or halting visual checks all together after the initial training period for a scanning technician.

Beyond the contribution that we hope our data will make to the field of manuscripts digitization, studying time and quality control data for the CCC large-scale digitization grant will allow the TRLN partners to better understand the cost and value of existing workflows and to reflect on future improvements. Partners now have the information they need to decide on modifications to visual check and other quality control workflows, as well as the impetus to explore automated quality control possibilities. The data gathered for this project is also useful beyond the scope of cost benefit analysis: for example, time data on scanning rates and quality control has proven critical to accurate planning for storage capacity and projections for timelines to goal completion for the CCC grant. Data on the types of errors produced during scanning and on who is producing the errors has been helpful to improving training for individual scanning technicians and developing general training materials.

Though visual checks – particularly those comparing analog materials to digital surrogates – are by far the most effective way to discover certain types of scanning errors, they are time intensive and rely on human effort. Many institutions use combinations of visual and automated checks in manuscripts digitization and have begun relying on sampling instead of performing quality control on 100 percent of materials. However, there is not much documented information available on best practices for quality control in the age of large-scale manuscripts digitization. As large-scale manuscripts digitization rapidly becomes a larger player in archives and libraries, it is critical that we do not simply transfer existing digitization workflows for smaller-scale projects over to large-scale efforts. We must gather and analyze data so that we can make informed decisions that take into account both the costs and the benefits of our workflows, and choose those workflows that best meet our institution's needs and capabilities.

Notes

1. Examples of institutions providing large-scale digital delivery through finding aids include the Archives of American Art at the Smithsonian Institute, available at: www.aaa.si.edu/collections/online, the University of North Carolina at Chapel Hill: http://dc.lib.unc.edu/cdm/archivalhome/collection/ead, Princeton University: http://findingaids.princeton.edu/, among many others.
2. See these articles for further discussion of large-scale digitization: Greene (2010); Ranger (2006); Dietz and Ronallo (2011); Erway and Schaffner (2007); Southern Historical Collection (2009).
3. The term "Long Civil Rights Movement" was coined by historian Jacquelyn Dowd Hall (2005) article, "The Long Civil Rights Movement and the political uses of the past." The Long Civil Rights Movement takes the traditional narrative of the civil rights movement – commonly acknowledged as the period between the Brown v. Board of Education US Supreme Court decision of 1954 and The Voting Rights Act of 1965 – and recasts it, extending the classical chronology of the period at both ends back to the 1930 s and forward to the 1990 s. The new narrative also broadened and deepened the classical storyline of racial injustice to include the struggles against economic, social, and environmental injustice that continue even today.

4. For more articles on the importance of documenting standards, quality control guidelines, and management approaches, see: Aikens *et al.* (2009); Anderson and Maxwell (2004); Conway (2000); Chapman (2000a); Chapman (2000b); Lee (2001); Frey and Reilly (1999); Smith (2000); Greenstein and Thorin (2002); Jordan (2006); Rieger (2000); Smith (2000); Maroso (2005).
5. For more first-person accounts from various universities of specific quality control problems with digitization projects, see The Library of Congress/Ameritech, "Lessons Learned: National Digital Library Competition," Library of Congress, available at: http://lcweb2.loc.gov/ammem/award/lessons/workflow.html
6. University of Alabama large-scale digitization guidelines: available at: www.lib.ua.edu/wiki/digcoll/index.php/Quality_Control and regular digitization guidelines: www.lib.ua.edu/wiki/digcoll/images/0/0a/General_QC_guidelines.docx; Georgia Digital Library guidelines: http://dlg.galileo.usg.edu/guide.html#07; University of Maryland guidelines: www.lib.umd.edu/dcr/publications/best_practice.pdf; NARA's guidelines: www.archives.gov/preservation/technical/guidelines.pdf; University of Alabama's guidelines on sampling for large-scale digitization: www.lib.ua.edu/wiki/digcoll/index.php/Septimus_D._Cabaniss_Papers
7. The first article that will publish the project's findings will be: Conway (2013b).
8. For more information on Gevinson's findings or to see error rates for the other two mass digitization efforts reviewed by Gevinson, available at: www.clir.org/pubs/reports/pub147/reports/pub147/data2Gevinson.pdf or read the entire CLIR report at: www.clir.org/pubs/abstract/reports/pub147
9. This number of hours does not equal the total work hours of either students or staff. The Digital Production Manager only scans ten hours a week, and scanning and related workflows are not the only duties of the student assistants (for example, students also prepare materials by performing condition review and removing fasteners, perform outreach and promotion, train, post process files, and preparing files for long term storage). The total number of work hours for the Digital Production manager and CCC students working in the UNC Digital Production Center during this time was 820.

References

Aikens, B., Weiss, K. and Reiter, T. (2009), "Building a large scale digitization program at the Archives of American Art", paper presented at Moving from Projects to a Program: The Sustainability of Large-Scale Digitization of Manuscript Collections, Chapel Hill, NC, March 12, available at: www.lib.unc.edu/mss/archivalmassdigitization/download/aikens_weiss_reiter.pdf (accessed 13 September 2012).

Anderson, C.G. and Maxwell, D.C. (2004), *Starting a Digitization Center*, Chandos Publishing, Oxford.

Belfiore, D. (2012), "Case study: using Perl and CGI scripts to automate a quality control workflow for scanned congressional documents", *Code4lib*, No. 17.

Brown, L. (2011), "The trickle, the firehose, and the bucket: large-scale manuscript digitization at UNC-Chapel Hill", unpublished paper presented on the panel "More Access to More Content: The EAD Finding Aid and Other Effective Tools for Large-Scale Digitization" at Archives 360: 75th Annual Meeting of the Society of American Archivists, Chicago, IL, 22-27 August.

Chapman, S. (2000a), "Considerations for project management", in Sitts, M. (Ed.), *Handbook for Digital Projects: A Management Tool for Preservation and Access*, Northeast Document Conservation Center, Andover, MA, pp. 16-30.

Chapman, S. (2000b), "Working with printed text and manuscripts", in Sitts, M. (Ed.), *Handbook for Digital Projects: A Management Tool for Preservation and Access*, Northeast Document Conservation Center, Andover, MA, pp. 104-110.

Conway, P. (2000), "Overview: rationale for digitization and preservation", in Sitts, M. (Ed.), *Handbook for Digital Projects: A Management Tool For Preservation And Access*, Northeast Document Conservation Center, Andover, MA, pp. 1-15.

Conway, P. (2011), "Archival quality and long-term preservation: a research framework for validating the usefulness of digital surrogates", *Archival Science*, Vol. 11 No. 3, pp. 293-309.

Conway, P. (2012), *Validating Quality in Large-Scale Digitization*, available at: http://hathitrust-quality.projects.si.umich.edu/ (accessed 13 January 2013).

Conway, P. (2013a), "Re: missing online data set references in your grant proposal", E-mail Message to J. Chapman (joyce.chapman@ncdcr.gov) Sent 1/3/13.

Conway, P. (2013b), "Preserving imperfection: assessing the incidence of digitization error in HathiTrust", *Preservation, Digital Technology & Culture*, Vol. 1 No. 1.

DeRiddler, J.L., Presnell, A.A. and Walker, K.W. (2012), "Leveraging encoded archival description for access to digital content: a cost and usability analysis", *The American Archivist*, Vol. 75 No. 1, pp. 143-170.

Dietz, B. and Ronallo, J. (2011), "Automating a digital special collections workflow through iterative development", paper presented at Association of College and Research Libraries Annual Conference, Philadelphia, PA, April 2, available at: www.ala.org/ala/mgrps/divs/acrl/events/national/2011/papers/automating_digital_s.pdf (accessed 13 September 2012).

Dowd Hall, J. (2005), "The long Civil Rights Movement and the political uses of the past", *The Journal of American History*, Vol. 91 No. 4.

Erway, R. and Schaffner, J. (2007), "Shifting gears: gearing up to get into the flow", paper presented at OCLC Programs and Research, available at: www.oclc.org/resources/research/publications/library/2007/2007-02.pdf (accessed 13 September 2012).

Frey, F.S. and Reilly, J.M. (1999), "Digital imaging for photographic collections: foundations for technical standards", National Endowment for the Humanities, Image Permanence Institute, Rochester Institute of Technology, Rochester, New York, NY, available at: www.ica.org/download.php?id=626 (accessed 13 September 2012).

Gevinson, A. (2010), Summary, Additional research materials for "The idea of order: transforming research collections for 21st century scholarship", Digital Library Federation, Council on Library and Information Resources, available at: www.clir.org/pubs/reports/pub147/reports/pub147/sumGevinson.pdf (accessed 30 December 2012).

Greene, M.A. (2010), "MPLP: it's not just for processing anymore", *The American Archivist*, Vol. 73 No. 1, pp. 175-203.

Greenstein, D. and Thorin, S.E. (2002), "The digital library: a biography", Digital Library Federation, Council on Library and Information Resources, pp. 1-70, available at: www.clir.org/pubs/reports/pub109/pub109.pdf (accessed 13 September 2012).

Jordan, M. (2006), *Putting Content Online: A Practical Guide for Libraries*, Chandos, Oxford.

Lee, S. (2001), *Digital Imaging: A Practical Handbook*, Neal-Schuman Publishers in association with Library Association Publishers, New York, NY.

Library of Congress and Ameritech National Digital Library (1996-1999), *Lessons Learned: Workflow and Project Management*, accessible at: http://lcweb2.loc.gov/ammem/award/lessons/workflow.html (accessed 13 September 2012).

Maroso, A.L. (2005), "Educating future digitizers: The Illinois Digitization Institute's basics and beyond digitization training program", *Library High Tech*, Vol. 23 No. 2, pp. 187-204.

Ranger, J. (2006), "More bytes, less bite: cutting corners in digitization", unpublished paper presented at Midwest Archives Conference symposium, Fall, Omaha, NE, available at: www.archivists.org/conference/sanfrancisco2008/docs/session701-ranger.pdf (accessed 13 September 2012).

Rieger, O. (2000) in Kenney, A.R. and Rieger, O.Y. (Eds), *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, Research Libraries Group, Mountain View, CA, pp. 61-83.

Riley, J. and Whitsel, K. (2005), "Practical quality control procedures for digital imaging projects", *OCLC Systems and Services: International Digital Library Perspectives*, Vol. 21 No. 1, pp. 40-48.

Smith, S.D. (2000), "Cooperative imaging: scans well with others", in Sitts, M. (Ed.), *Handbook for Digital Projects: A Management Tool for Preservation and Access*, Northeast Document Conservation Center, Andover, MA, pp. 136-139.

Southern Historical Collection (2007), "Extending the reach of Southern sources: proceeding to large-scale digitization of manuscript collections", Final Grant Report for the Andrew W. Mellon Foundation, June 2009, available at: www.lib.unc.edu/mss/archival massdigitization/download/extending_the_reach.pdf (accessed 13 September 2012).

Further reading

Fichter, D. (1999), "Saskatchewan digital library collections: enhancing access to the province's information", *Library Hi Tech*, Vol. 17 No. 2, pp. 172-180.

Gladney, H.M. and Lotspiech, J.B. (1997), "Safeguarding digital library contents and users: interim retrospect and prospects", *D-Lib Magazine*, Vol. 4 No. 1.

Jewell, T.D. (2001), "Selection and presentation of commercially available electronic resources: issues and practices", Digital Library Federation, Council on Library and Information Resources, Washington, DC, pp. 1-61.

McGill, T.M. (2004), "Rapid implementation of a large-scale text digitization project: Colorado State University Libraries' experience", *Colorado Libraries*, Vol. 30 No. 1, pp. 29-31.

About the authors

Joyce Chapman is the Consultant for Communications and Data Analysis at the State Library of North Carolina where she provides state-wide support to public libraries for data analysis, assessment, advocacy, and communications. Until recently she served as Project Manager for the consortial grant "Content, Context, and Capacity" at the Triangle Research Libraries Network, and led data analysis efforts for the case study on quality control visual checks.

Samantha Leonard is the Media Services Librarian at Highpoint University Libraries where she manages the media services lab and assists students and faculty with media needs such as film editing, production, design, and photography computer software. Until recently she served as the Digital Production Manager for the consortial grant "Content, Context, and Capacity" at the Triangle Research Libraries Network, and led data collection efforts for the case study on quality control visual checks.