Paul Conway*

Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in HathiTrust

Abstract: Large-scale digitization efforts by third-party firms are the subject of no small amount of controversy and criticism, as is especially the case with Google Books. This article reports some of the findings and important implications of a rigorous multi-year quantitative and qualitative assessment of the images representing a sizable proportion of the digital surrogates created by Google and deposited in the HathiTrust, which is one of the most important large-scale preservation initiatives to emerge in higher education in the past fifty years. The population of study described here consists of English-language books and serials published before 1923 that were scanned and processed by Google between 2004 and 2010. At the time the data for the study were gathered (2011), this population consisted of approximately 1.25 million volumes or roughly 12 percent of the HathiTrust corpus. The findings suggest that the imperfection of digital surrogates is an obvious and nearly ubiquitous feature of Google Books and that such imperfection has become and will remain firmly ensconced in collaborative preservation repositories.

***Paul Conway:** Associate Professor, University of Michigan School of Information, e-mail: pconway@umich.edu

1 Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in HathiTrust

The HathiTrust Digital Library is one of the most important large-scale preservation initiatives to emerge in higher education in the past fifty years. Its sixty-plus research library members have joined their resources, built a robust and sustainable digital storage and delivery platform, and established a governance structure with a mission "to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge" (HathiTrust 2012a). Behind this commitment to a longstanding mandate of research libraries is a simple reality: HathiTrust is now and is likely to be for the foreseeable future primarily a repository for digitized library volumes from Google's foray into large-scale digitization. HathiTrust now (2013) contains well over 10 million digitized volumes, 96.4 percent of which have been produced by Google from the contents of at least 18 library collections (York 2010). The digital surrogates in HathiTrust encompass 429 languages across the spectrum of library classification and the history of books and printing since Gutenberg (HathiTrust 2012b). In terms of collection size, HathiTrust now ranks approximately 10th among the 126 members of the Association of Research Libraries (ARL 2012).

Large-scale digitization efforts by third-party firms are the subject of no small amount of controversy and criticism, as is especially the case with Google Books. Charles Bailey maintains a wide-ranging bibliography of writings on Google's digitization program that includes 350 items published between 2004 and 2011 (Bailey 2011). Among the major concerns expressed in the accumulated news coverage, scholarly articles, and books for the general and specialized reader are the dangers of corporate control of research resources (Darnton 2010), the legality of wholesale digitization (Proskine 2006), inadequate and incomplete coverage of intellectual disciplines (Lavoie 2005; Jones 2011), poor search-and-discovery results (Nunberg 2009), and the secrecy that surrounds Google's digitization workflows (Leetaru 2008). Oya Rieger (2008) explored the preservation implications of four large-scale projects, including Google Books, and concluded that some of the most serious problems have to do with the quality of the page images for the reader, the metadata associated with digital surrogates, and the underlying full-text data that make text searchable. A litany of complaints from scholars, librarians, archivists, and technologists about image quality leaves the impression that Google has privatized a vital public resource and has delivered to readers an inadequate digital product that fails to meet the needs of scholars, students, or the general public. In the debate over the appropriateness of large-scale book digitization, it seems that everyone has a stake in digital book surrogates because everyone knows what a book is and how a digital version ought to be represented online.

The commitment to preserve digital surrogates from Google Books dates from the contract that the University of Michigan negotiated with Google at the start of what has become a world-wide digitization effort (Courant 2006; Karle-Zenith 2006). HathiTrust has emerged since 2008 as a large-scale exemplar of a preservation repository containing digitized content with intellectual property rights owned by a variety of external entities, created by multiple digitization vendors for access, and deposited and preserved collaboratively (York 2009, 2010). The HathiTrust project is also an example of an emerging trend in preservation repositories to accept digital content with quality assurance largely limited to checking, during the ingest process, that image files render properly in current software. For such repositories and their communities of users to trust digital documents, repositories must assess the quality of the information content of the preserved digital objects and validate their fitness for the many uses envisioned for them. Information quality should be an important component of the value proposition that digital preservation repositories offer their stakeholders and users (Conway 2010).

Some key findings and important implications of a rigorous quantitative and qualitative assessment have emerged concerning the quality of the images representing a sizable proportion of the digital surrogates created by Google and deposited in HathiTrust. This report is one component of a multi-year, multi-method research project consisting of three overlapping investigative phases. Phase one defined and tested a set of error metrics (a system of measurement) for digitized books and journals. Phase two (part of which is reported here) applied those metrics to produce a set of statistically valid measures regarding the patterns of error (frequency and severity) in multiple samples of volumes drawn from strata of HathiTrust. Phase three engaged stakeholders and users in building, refining, and validating the use-case scenarios that emerge from the research findings. The three-year research program has been supported by the Andrew W. Mellon Foundation and the Institute for Museum and Library Services. The design of the study and summary of the quantitative methodology are published elsewhere (Conway 2011; Conway and Bronicki 2012).

2 Background and Relevant Research

The quality of digital information writ large has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. Stuart Madnick and his colleagues (2009) review the evolution and landscape of information quality research and call for research on the quality of large-scale image databases as one among a number of important recommendations. Building on Doermann (2003) and Le Bourgeois (2004), Lin (2006) provides an excellent review of the state of digital-image-analysis research in the context of large-scale book-digitization projects. He establishes a "catalog of quality errors" that distinguishes errors that take place during digitization (e.g., missing or duplicated pages, poor image quality, poor document source) from errors that arise from post-scan data processing (e.g., image segmentation, text recognition errors, and document-structure-analysis errors). The literature on information quality, however, is relatively silent on how to measure quality attributes of large collections of digitized books and journals, created as a combination of page images, full-text data, and underlying XML.

Over the seventy-year period since the words "archival quality" first appeared together in professional and research literature, the term has been used as a simple metaphor for three complex but interrelated concepts: properties of archival records, characteristics of storage media, and the processes that preserve the essential nature of artifacts when copied or transferred to another medium (Conway 2011). "Archival quality" became shorthand for a suite of processes and policies designed to extend the life expectancy of archival materials, thereby distinguishing them from information resources of lesser value (Conway 1989). Trust and archival quality have become most closely associated through the preservation management of digital surrogates, beginning with research at Cornell University to adapt quality-measurement techniques, from microfilm to digital bit mapping (Kenney & Chapman 1996). In doing so, archivists, librarians, and preservation administrators mutually reinforced a particular perspective on quality oriented toward defining thresholds of digital-image characteristics adapted from image science (FADGI 2010). Initial critical commentary on Google Books, for example, focused on the failure of large-scale digitization to adhere to well established digital reproduction specifications, such as those promulgated by the Digital Library Federation (DLF 2002).

Little systematic research has been completed on the digitization quality of Google Books. Scholars Robert Townsend (2007) and Paul Duguid (2007) attempt to reach general conclusions about digitization quality from a close inspection of a favorite volume. Ryan James (2010) conducted a small random-sample study of text legibility and found about one percent of the 2,500 pages reviewed had errors severe enough to affect readability, such as text blurring, obstructed content, and missing pages. Scott McEathron (2011) evaluated a random sample of

180 volumes on geology from a population of over 2,500 volumes in HathiTrust. He found a 2.5 percent rate of scanning errors thinly but widely distributed through 63 percent of the sample.

The work of historian Alan Gevinson (2010) stands out for his attempt to reach beyond personalized, impressionistic treatment of image error. Gevinson, a scholar of American intellectual history, started his investigation with a well accepted list of 200 influential books in the field. Searching for each of them in the Google Books interface, which contains nearly identical versions of those deposited in HathiTrust, he reported on his success in finding digital versions and on the problems he encountered with the books located and examined. He found a low incidence of error in volumes published since 1922 but a host of problems with older volumes, including 21 percent with pages missing, 16 percent with blurred or thin text, and 19 percent with cropped or obscured text. Gevinson's study suffers from challenges he had in finding and viewing specific titles in HathiTrust, and from a lack of clarity about error definitions and the little effort put forth to distinguish between minor and critical error. For example, Gevinson judges 32 percent of the pre-1923 volumes to be of "poor" quality, without providing a definition of the term. Gevinson's research, however, points the way toward a systematic and predictive study of quality.

In the context of large-scale digitization, in which thousands or millions of objects are scanned against a single digitization technical specification in a factory-like workflow (Leetaru 2008), digitization leaves traces that are visually detectable to varying degrees—artifacts of the scanning process itself (clamps and fingers, skewed objects) and artifacts of the post-scan image manipulation processes (moiré patterns, visual distortions). In this study the quality of large-scale digitization is not defined as a property shared between source and the resultant scan, but instead as the absence of visible artifacts in the digital page-image surrogate, in the form of process and processing errors that interfere with use**.** The absence of error in a digitized volume may be absolute, in that a given volume and its representative page images are *perceived to be* free of errors. The absence of error may also be defined relative to expected uses, whereby perceived error may or may not have an impact on the usefulness or usability of the original content transformed to digital form. Other errors may not interfere with use but nevertheless may impinge on the overall acceptance of the digital surrogate relative to the original source or to other digital surrogates produced at a higher standard. The assessment of quality in large-scale digitization thus must begin with the definition of error and the measurement of absolute error in a given population of digital surrogates. When the extent of absolute error is understood with respect to reliability, it then becomes possible to assess the impact of error on use, on acceptance of surrogacy itself, and ultimately on the trust in the entire repository and its preserved content. This article is thus a presentation of evidence on the presence or absence of absolute error in a large sample of digitized books and an assessment of why such errors occur.

3 Study Methodology

The population of study described here consists of English-language books and serials published before 1923 that were scanned and processed by Google between 2004 and 2010. At the time the data for the study were gathered (2011), this population consisted of approximately 1.25 million volumes or roughly 12 percent of the HathiTrust corpus. This sampling was chosen for this first study because all of these volumes are in the public domain, potentially physically accessible for inspection, reviewable without special language skills, and fully viewable through the HathiTrust user interface (http://www.hathitrust.org/) as well as through search and viewing tools made available by Google Books (http://books.google.com/). Future reporting will assess error in three other sample populations: Google-digitized books published after 1922; books in the public domain digitized under the auspices of the Internet Archive; and books in HathiTrust printed in four non-Roman scripts.

In the assessment of error in a large dataset, statistical sampling has two purposes: to test and refine an error-definition model; and to predict the incidence of error for all or part of the general population from which the sample is drawn. The project team addressed the issue of representativeness in the sampling techniques applied during the data-collection phase. Under direction from the team statistician, a programmer developed an algorithm to select an appropriately sized random sample from the HathiTrust. Project co-Principal Investigator Edward Rothman, a distinguished scholar of statistical process control, determined that 1,000 volumes would be representative of sampling pools in HathiTrust and would allow for statistical comparison of sub-populations with potentially small frequencies in important variables (Jovanovic and Levy 1997).

From the 1,000-volume sample, the project team extracted a systematic random sample of approximately

100 pages from each volume.[1] This method insured that the sample fully represents the sequencing of page images in a given volume while giving equal treatment to volumes with widely varying numbers of pages. The review thus began with a total of 93,858 page images from 1,000 volumes representing 1.25 million digitized volumes.

The research team focused initially on sampled page images in a digitized volume, followed by a page-by-page review of the entire digitized volume, culminating, as explained below, in a physical review of the exact-match volume originally scanned by Google. A three-tiered hierarchical model hypothesizes error at the levels of text/illustration, page image, and whole volume and assigns one or more potential causes for each error (source volume, scanning, post-scan manipulation) (Conway and Bronicki 2012). Page-image errors are individually identifiable attributes that affect the visual appearance of single bitmap pages, such as thick or broken text, distortions in accompanying illustrations, and warped or cropped pages. A particular error may be confined to a single page or repeated across a sequence in a volume. Whole volume-level errors apply to structural issues surrounding the completeness or accuracy of the volume as a whole, such as missing pages (including foldouts not digitized), duplicate pages, and ordering of pages. For each of the eleven page-image errors in the model, the research team developed and tested a scale to rate the perceived severity of each error on a scale of 0 to 5, where the most severe rating applies to errors that make all or some portion of the original content in a page image unusable.

Carefully trained reviewers working independently at the University of Michigan and the University of Minnesota visually inspected full-scale page images and manually assigned a severity score from one to five for each error on a given page image. A default-data value of zero represents no perceived error for a given error type. When a reviewer detected an error at the highest level of severity (5), an additional variable provided for the assignment of a code representing the proportion of the page affected by the error. The project developed a highly efficient and statistically reliable data-gathering and analysis system to measure error incidence in HathiTrust volumes. Reviewers then re-inspected each volume in the sample using a separate review system that displayed large, zoomable thumbnails of up to 18 page images at a time, in the order stored and presented in HathiTrust. The system supported the binary coding (yes/no) of the presence of five whole-volume errors as well as a count of continuous sequences of volume-level error. Finally, a separately trained and managed team of graduate students retrieved exact-match original source volumes from the University of Michigan and 17 other research libraries represented in the sample. With volume in hand and using a custom-built data-gathering interface, the students inspected the physical volumes for the presence of damage, printing and binding anomalies, and other signs of age or use that might have an impact on the quality of the resultant digital scans. The design of the physical inspection was influenced by the sequence of preservation-condition surveys pioneered by Yale University (Starmer and Rice 2004).

The data-gathering process produced accurate, complete, and well formed data sets for each of the three review methods (page-image, whole-volume-level, physical volume) linked by a unique HathiTrust volume identifier and a physical-volume bar code. This approach to data management allows for the assessment of the frequency and severity of error at the individual page-image level and the aggregation of error measures to the volume level. Linking the three data sets also allows for the correlation of values from page-image data to volume aggregation, as well as correlation of page-image data with the physical characteristics of digitized volumes.

4 Findings on Digitization Error

The following three sections of this article present and interpret the findings from a review of 93,858 page images from 1,000 volumes published in English before 1923 and digitized by Google between 2004 and 2010. The first section covers page-level errors. The second section discusses the relationship between page-level error and the physical characteristics of the original source volumes. The third section covers the five whole-volume errors perceived in a page-by-page review of the same volumes sampled at the page level and inspected physically.

[1] The sampling algorithm was applied to the image sequence number, the complete set of which serves as a proxy for the total number of pages in a given volume, cover to cover. The sampling algorithm divided the total number of image sequences for a given volume by 100 to establish the whole-number sequential sampling interval value. A random number generator established where in the volume (between sequence number 0001 and 0010) to begin sequential sampling. Sequential processing then identified and selected page images according to the sampling interval value, rounded up or down accordingly. For example, for a digitized volume of 435 pages, every fourth page image was chosen from the volume for review; a digitized volume of 789 pages was sampled every eight page images.

4.1 Distribution of Page-level Error

The data set for page-level error contains 1,032,438 data points, representing the coding of eleven possible errors for each of the 93,858 page images reviewed. Eight coders in two teams perceived error at some level of severity 226,851 times, for an overall average of 2.42 error for any given page image where error was detected. Appendix 1 presents the total number and proportion of errors codes for each of the eleven errors across the five levels of severity.

Coders had significant difficulty applying the five-level severity coding scheme to page images with digitized illustrations. The information has been excluded from the following analysis because it does not appear to be a reliable indicator of the perception of digital artifacts from scanning (e.g., moiré patterns) or problems with the tonal contrast or color fidelity of illustrations and graphic material. A separate study of illustration error was conducted subsequent to the completion of the full sample and will be reported separately.

Of the remaining eight errors, five of them (thick text, broken text, warped pages, cropped pages, and obscured content) account for most of the error at any level. Table 1 presents the distribution of the severity of error across these five errors, with a special emphasis on distinguishing between minor and severe error. The five errors account for 96.9 percent of all perceived error at level one, 82.5 percent at severity level four, and 87.9 percent of all error perceived at level five. Severity levels four and five are distinguished by the amount of inference that is required or possible by the reader to render the text intelligible. At severity level four, data coders were nearly unable to decipher the content in the affected area of the page and significant inference was required by the reviewer to obtain legibility and meaning. Severity level five is catastrophic; original content in the affected area of the page cannot be unambiguously deciphered or has been obscured altogether.

The representation of English-language text in Google-digitized page images is problematical at both extremes of severity. Table 1 shows that almost 30 percent of all page images in the sample display some level of text distortion on some portion of the image. Over a quarter of all images reviewed yielded evidence of low-severity thick text or broken text. Thick text appears to the reader as boldfaced in a way that is not typographical. Broken text poses the opposite challenge; readability is challenged by light, thin, or disintegrated text. At its most severe, thick text appears as blobs rather than distinct characters, rendering it difficult or impossible to understand. Broken text can be perceived as a lack of image contrast, where the typography appears washed out. Errors in text rendering may affect users in different ways. Clearly fully indecipherable text on all or part of a page inhibits understanding; but pervasive low-level error in text may affect concentration in online reading or undermine the desirability of digital surrogacy for some users. It is important that future research clarify the impact of low-level error on usability.

Although problems with the rendering of text are common, extreme distortion is rare in the sample. Only 711

Severity Level (excluding levels 2 and 3)

		Severity=0		Severity=1		Severity=4		Severity=5	
				Sev=1					
		Total	Percentage	Total	Percentage	Total	Percentage	Total	Percentage
Error Type	Thick Text	58,233	62.04%	24,086	25.66%	183	0.19%	107	0.11%
	Broken Text	57,252	61.00%	28,124	29.96%	179	0.19%	242	0.25%
	Cropped Page	93,235	99.37%	256	0.27%	20	0.02%	143	0.15%
	Warped Page	27,425	29.22%	56,482	60.18%	33	0.04%	44	0.05%
	Obscured Content	15,847	16.88%	73,257	78.05%	75	0.08%	436	0.46%
	Total Error			182,205		490		972	
	Portion of Total Error			96.9%		82.5%		87.9%	

Table 1: Distribution of Five Most Common Page-level Errors
Google-digitized, pre-1923, English language
Sample size: 93,858 page images from 1,000 volumes

of 93,858 pages reviewed (0.76%) were nearly or completely indecipherable.[2] The large sample size yields a 95 percent confidence level that this small proportion of catastrophic text error represents the predicted severe error in the overall population of Google-digitized volumes published before 1923. This small proportion of severe error to the study population of 1.25 million volumes at an estimated 300 page images per volume suggests that the text on about 1.4 million page images may be indecipherable.

Beyond the challenge of accurately rendering text characters, a review of the sample page images also revealed three important page-level errors. Over 60 percent of all page images do not appear flat. The subtle effect of warping is a byproduct of Google's patented post-scan processing algorithms that attempt to remove the appearance of curvature that results when volumes are scanned in their bindings. When the algorithm fails, an error of severity level four or five results. When the algorithm does not flatten the image completely but does not interfere with intelligibility, reviewers assigned a severity level of one.

The most common page-level error is obscured content, which is observable in over three quarters (78.0%) of all page images reviewed. The most common causes of this error are Google's scanning process or incomplete or failed efforts to repair the image through post-scan processing. Dan Cohen (2007) and other commentators have been quick to comment on the human fingers frequently evident in page images. More common still are the subtle remnants of Google's patented method for processing images to remove fingers and clamps, substituting pixels that are coded to resemble the tone and color of surrounding paper. When this post-scan processing does not affect text or illustration in a page image, reviewers assigned a severity level of one. When fingers or clamps cover text, reviewers assigned a severity level of five. In the study set, 66.25 percent of the 289 page images are assigned a severity rating of five because original content was covered by fingers or clamps.

The proportion of severe error perceived in page images is in keeping with the findings of McEathron (2011) and James (2010) but is far less than the error observed by Gevinson (2010). McEathron's sample was small and not precisely specified, so the direct comparison of findings should be done carefully.

4.2 Contribution to Error of Source-Volume Characteristics

Page images of books digitized at scale without rigorous quality-control processes at each stage of production pose a particular challenge in determining the source of error. Is that blurry text an artifact of scanning or is it a result of poor printing? Is the cropped text a result of a mistake in automated post-scan manipulation or did the text block get re-bound too tightly long ago? The adage "bad books produce bad scans" has been a staple of digitization for preservation for two decades. In gathering data for the research project, reviewers were trained to "code what you see" without speculating on the cause of error. Nevertheless, understanding the source of page-image error establishes a balanced view of large-scale digitization processes.

To support an assessment of the contribution that a source volume makes to patterns of perceived error, the research project team borrowed and inspected 860 exact-match books from the 1000-volume sample. The remaining 140 books were either not available for loan or were too fragile to be inspected. The inspection program required the cooperation of eighteen libraries where the exact-match volumes are housed. The effort to match and obtain the volumes was complex and time consuming; a full description of the inspection methodology will be reported separately. Table 2 summarizes the findings from the physical inspection of exact-match volumes in the sample population. The results are organized into three categories reflecting the issues that bindings, paper quality, and printing technologies present to digitization processes.

Bound volumes can complicate production-level scanning by increasing the time and effort required to handle them. The physical inspection of the sample, whose publication dates range from 1699 to 1922, found that over one-quarter of the volumes have tightly bound text blocks with gutter margins of one centimeter or less. Almost all of the volumes (98.7%) were well produced, showing no signs of warped or skewed text. Reviewers found eleven volumes with cropped text, four of which can be pegged to tight gutters. These findings suggest strongly that digital surrogates that display these characteristics of warping and skewing suffer from poor digitization processing rather than poor source material.

[2] For page images with the most severe error, reviewers coded the page image for the proportion of the page affected by the catastrophic error. The text for 349 page images is completely indecipherable, because of thick or broken characters. Over three-quarters of this indecipherable text is found on one-third or less of a given page image. The text of 38 pages images from the extensive study set is indecipherable on most or all of the page image.

Total volumes reviewed = 860

Binding Issues	Condition	Total Number of Volumes	Proportion of Sample	Chi-square (p-value)
Binding Integrity	**Fully Intact**	**692**	**80.5%**	0.0897
	Loose	119	13.8%	
	Not Intact	43	5.0%	
	Missing All or Part	6	0.7%	
Gutter (Width)	**More than 1.0 cm**	**644**	**74.9%**	0.0017
	Less than 1.0 cm	216	25.1%	
Text Block Integrity	**Fully Intact**	**690**	**80.2%**	0.0431
	Missing Pages	9	1.0%	
	Loose Pages	93	10.8%	
	Parts Missing	18	2.1%	
	Text Block Broken	50	5.8%	
Binding Errors	**None**	**849**	**98.7%**	N/A
	Cropped Text	11	1.3%	
	Skewed Text	0	0.0%	
	Warped Text	0	0.0%	
Paper Issues	**Condition**	**Total Number of Volumes**	**Proportion of Sample**	
Brittle	**Not Brittle**	**390**	**45.3%**	0.9927
	4 Double Folds	117	13.6%	
	2 Double Folds	353	41.0%	
Paper Damage	**Not Damaged**	**769**	**89.4%**	0.1671
	Torn, Ripped	57	6.6%	
	Animal or Insect	0	0.0%	
	Water Damage	33	3.8%	
	Food or Drink	1	0.1%	
Text Bleedthrough	**No**	**785**	**91.3%**	N/A
	Yes	75	8.7%	
Annotation	**No Annotation**	**829**	**96.4%**	0.3137
	Pencil	19	2.2%	
	Ink	9	1.0%	
	Highlighter	1	0.1%	
	Other	2	0.2%	
Printing Issues	**Condition**	**Total Number of Volumes**	**Proportion of Sample**	
Text Printing Errors	**None**	**847**	**98.5%**	N/A
	Thick Text	1	0.1%	
	Broken Text	11	1.3%	
	Blurred Text	1	0.1%	
Substantial Color	**No**	**838**	**97.4%**	0.2198
	Yes	22	2.6%	
Illustrations > 10	**No**	**571**	**66.4%**	0.001
	Yes	289	33.6%	

Table 2: Material Characteristics of Sampled Volumes
Google-Digitized, pre-1923
Volumes Reviewed = 860

Nine of the 860 volumes inspected (1.0%) showed clear evidence of having missing pages and another 18 volumes were severely dog-eared or had portions of the pages missing. These particular findings are problematic for large-scale digitization because a missing page in a source volume leads directly to missing information in the digital file that is impossible to attribute definitively to either the source or the surrogate without side-by-side comparison. Since Google's production-level scanning rarely pauses to notice missing information, the onus for a book's physical integrity must fall on some form of pre-scan collation or post-scan processing.

The findings on physical condition are consistent with findings from 25 years of physical-condition surveys in research libraries (Starmer and Rice 2004). Nearly 20 percent of the volumes have loose, separated, or missing bindings. The same proportion of volumes (19.8%) has problems with the integrity of the text block, including loose or torn pages. Approximately 41 percent of the volumes examined have brittle paper (breaking at 2 double folds) while an additional 13.6 percent of the volumes are on the cusp of brittleness. In the sample, 110 volumes have torn or ripped pages or show evidence of water damage. Few books contain annotations of any form. Text bleed-through from one side of the page to the other could be a problem for 8.7 percent of the sample population. Overall it seems that problems with the physical integrity of scanned volumes could be greater than are problems with paper itself.

Only 13 of the 860 physical volumes reviewed showed evidence of thick, broken, or blurred text. Similarly, few of the volumes (2.6%) printed before 1922 substantially use color in text or illustration, so digitization may not present the same challenge regarding color fidelity in publications after that date.

To show the contribution to digitization error made by the physical characteristics of source volumes, the physical-inspection data in Table 1 were correlated with the incidence of severe error (level 4 or 5) in the digital surrogates for each volume. The analysis counted the number of page images in a digital volume that contained severe page-level error of any type, with a maximum allowable page-image count of 100. The results of this count were separated into two groups: volumes with up to three page images with severe error (514 of 860 volumes or 59.7 %) and volumes with more than seven page images with severe errors (38 of 860 volumes or 4.4%). A statistical analysis then tested the assumption that volumes with a relatively large number of page images with severe error will also show the presence of anomalies in the physical volume (e.g., brittle paper, tight bindings, annotations, etc.). The null hypothesis is that the distribution in the sample of volumes with the most severe error is random (not associated with any particular physical characteristics). Table 2 reports on the distribution of physical anomalies and on the significance of the chi-square test of random distribution.

The table shows that the presence of four characteristics of the physical volume predicts the prevalence of severe page-level error. Three of the four characteristics relate to the condition of the binding and the text block. Books with loose, not intact, or missing parts of the bindings tend to have fewer errors than volumes with fully intact bindings. Books with inner margins less than one centimeter between binding and text block tend to have more severe scanning errors than books with more ample gutters. Books with fully intact text blocks tend to have fewer severe errors than those with loose, broken, or missing text blocks. The implication for digitization of the comparison of physical books with exact-match digital surrogates is that in general the physical and bibliographic characteristics of source volumes have little or no impact on the quality of digitization. Severe error occurs largely independent of physical form. Where physical features do impinge on digitization quality, binding and text block condition are far more important than paper condition or the quality of printing. It is perhaps surprising that books in fragile condition may produce significantly fewer severe errors than do books intact and in good condition. These findings should not be construed to suggest that digitization processes should favor the weakest books in the collection. If anything, the findings argue that large-scale digitization can proceed without regard for the physical condition of the volumes.

4.3 Distribution of Whole-volume Error

In high-volume production scanning, errors may occur in assembling a complete copy of the text when pages are missed or are scanned more than once or poorly. Post-scan manipulation may inadvertently result in page-sequencing errors. Finally, policies and procedures that govern the digitization workflow may result in content that is not captured. None of these errors can be assessed accurately with the use of any in-volume sampling strategy such as was employed for page-image error detection. Given the present limitations of automated image processing, absolute confidence in the detection of whole-volume error requires a manual inspection of the entire digital surrogate, page by page, sometimes with the source volume in hand. The purpose of the whole-volume component of the research project was thus to measure components of the error model that affect the integrity of the digital surrogate as a whole. Secondarily, the whole-volume review project was designed to test the feasibility of identifying severe error (level 4 or 5) on page images without logging the specific nature of such catastrophic error.

For the whole-volume component of the research project, the project team developed a distinctive review interface that was optimized for the rapid review of all the page images in a given digital surrogate, presented in the order that they are stored in the HathiTrust digital repository. The system allowed for the indication of one or more missing pages in a sequence, the flagging of duplicate-page images, the identification of one or more pages out of page-number sequence, and the presence of page images that do not belong in the volume for any reason. One example of a false page is a miscue during scanning that results in an image of the scanning bed (or the lap of the camera operator) being included in the surrogate file. The system also allowed for reviewers to tag page images with indecipherable content. The system automatically tallied the frequency of each of these five errors and presented the tallies and the individual page sequences for statistical analysis. Finally, a time-and-motion study accompanying the review of the sample population measured the time it took each reviewer to complete the coding of a given volume.

Table 3 summarizes the frequency of error for the four whole-volume errors as well as the frequency of page images whose content was fully obscured. Eight reviewers coded whole-volume errors for the identical population of 1,000 digital volumes that was used in the page-image-error study and the physical-inspection project.

The first important finding is that 649 of the 1,000 volumes given rapid, whole-volume review are free of errors and have no page images whose content is obscured at severity levels 4 and 5. These volumes are accurate, from the perspective of pagination, and are reliably intelligible surrogates. Of the 468 volumes with whole-volume error, 117 volumes have more than one type of error.

Table 3 illustrates a low incidence of whole-volume error in terms of number of pages affected by a given error but a relatively high frequency of volumes affected. Eight reviewers working through the volumes identified only 660 missing pages in the total sample population of 397,467 pages reviewed. Only 66 of the 1,000 volumes reviewed indicated missing page sequences, with an average of 10 pages missing for each volume with a missing-page problem. The story is similar for the review of duplicate, out of order, and false pages. Missing pages represent a potentially irrevocable loss of content, while duplicate, out-of-order, and false pages may be primarily an annoyance that prohibits comprehension without ultimately prohibiting the use of the surrogate volume.

The challenge that whole-book error presents to the integrity of the HathiTrust collection arises from three complications. The first issue is that detecting elusive whole-volume errors requires manual inspection of 99 percent of acceptable page images to identify the one percent of the error. Manual inspection can be an efficient process, but given the scale of the HathiTrust collection, it may be impossible to review every volume using specially

	Total # of Page Errors	**Proportion of Total Page Error**	**Number of Volumes with Error Type Coded at Least Once**	**Mean Errors per Volume**
Missing Page	660	0,0017	66	10,0
Duplicate Page	572	0,0014	104	5,5
Out of Order Page	240	0,0006	32	7,5
False Page	41	0,0001	24	1,7
Fully Obscured Page	3307	0,0083	242	13,7
Summary				
Number of Volumes with No Error	649			
Number of Volumes with Volume-level error	468			
Number of Volumes with More than One Type of Volume-Level Error	117			

Table 3: Frequency of Whole-volume Error
Google-Digitized, pre-1923 Volumes
volume n =1,000
page-image n =397,467

trained coders. As review shifts from a skilled trade to one undertaken “by the crowd” in an open-review process, the detection of coding error rises.

The second issue with whole-volume review is that two of the four whole-volume errors (missing or out-of-order pages) can be confidently declared “errors” in digitization processing only with the side-by-side comparison of the source volume with the digital surrogate. Given that the physical inspection of the sample books found a one percent incidence of books with missing pages, it is as likely that missing pages are caused by digitization error as they are by the source volume itself.

The third issue with pagination-related errors in Google-digitized books published before 1922 relates to the presence of duplicate pages in the digital file. Close inspection of volumes coded with duplicate pages exposed Google’s policy to treat as a distinctive “real” page the thin tissues (“offset sheets”) inserted to protect an illustration from bleeding onto the text on the facing page. Google’s scanning technicians scanned a two-page spread twice: once with the tissue on the right (covering either text or illustration) then again with the tissue on the left (covering either text or illustration). Google’s post processing then separated the two two-page scans into four separate page images, two of which appear crisp and clean while the other two appear as broken text and low-contrast illustration. This issue can be fully diagnosed and addressed through a change in scanning policy or by reviewing and choosing the better image.

The most serious information-quality issue with Google Books, and therefore the largest problem for HathiTrust, is fully obscured content. Whole-volume review demonstrates that fully one-quarter of the volumes in the 1,000-volume sample contain at least one page image whose content is unreadable. There are two principal reasons for this: fatal digitization error (severity level 4 or 5) or the failure to digitize a page with folded content, such as maps, charts, or other graphic materials. The latter cause stems from an explicit policy by Google not to pause scanning to digitize foldouts bound into a book (Leetaru 2006). For severe digitization error, page-level review found the incidence of severe digitization error (thick and broken text, cropped and warped pages, obscured content across the entire page) to be 1.18 percent of the sample of 93,858 page images. Whole-volume review that simply sought to identify pages with indecipherable content without regard to error type found a similar incidence of error (0.83%) in the nearly 400,000 pages inspected. A one percent page-level failure rate translates into a large number of unintelligible page images.

5 Implications for the Preservation of Digital Surrogates

Google’s large-scale digitization process is a phenomenally productive method for producing digital surrogates of books and serials. Paul Courant (2006) reports that as provost he was convinced to forge the partnership between Google and the University of Michigan by Google’s assertion that the corporation could digitize the entire Michigan research library holdings in six years when Michigan was on track to do the same job in 1,000 years. Google further agreed to deliver to the university a digital surrogate that conformed to digitization standards promulgated by the Association for Research Libraries (2002). Other research libraries followed suit. Thus began an information-management process that has culminated in the formation and growth of the HathiTrust Digital Library for preserving the surrogates that Google aims to deliver. HathiTrust also has in place a mechanism for re-ingesting digital surrogates after Google has re-processed their digital-source files to reduce and/or eliminate errors.

The project whose partial findings are reported here is seeking to understand the nature of residual error from large-scale digitization and to assess the impact of that error on the uses of digital surrogates of books and serials in research libraries. It is yet premature to present firm conclusions about the relationship of error to usefulness. But the data reported in this article lead to some tentative conclusions.

The first is that minor error that does not limit the readability of digitized text or the intelligibility of digitized illustrations and graphic materials is ubiquitous in the HathiTrust collection. Such low-severity error should be accepted as a part of the price paid for enhanced access. Only a minority of the volumes in HathiTrust that are now in the public domain and are therefore fully viewable are error free at the severity levels one and two. These errors are visible, easily detectable, and so common as to become part of the fabric of digital surrogacy in HathiTrust and Google Books. Low-level-quality errors with text and illustration are not confined to HathiTrust, but also make their way into secondary products, including print-on-demand copies and versions prepared for the Amazon Kindle and other eBook readers. It is likely not feasible and perhaps undesirable to continue to process and reprocess digital surrogates to remove low-level error.

The second and related conclusion is that digital surrogates produced by Google (and likely by other large-scale digitization efforts) carry with them transparent evidence of scanning techniques and post-scan enhancement procedures, including visible fingers and clamps, subtle page warp, and inconsistent typography. In this way, large-scale digitization has established a new ethical norm that varies quite dramatically from the digitization norms pioneered in research libraries over 25 years ago. These norms held firm through a long series of digitization guidelines promulgated by libraries, archives, and museums throughout the world. The existence of millions of digitized volumes presents these organizations with a clear choice: accept these digital surrogates as new intellectual products, rather than as "faithful copies," or re-digitize a substantial portion of the world's research-libraries' holdings of books and serials to create cleaner and more pristine representations of volumes. Coming to terms with the distinction between source and surrogate will be difficult.

A third conclusion is that although minor error could become an acceptable feature of large-scale digitization, fatal error compromises the integrity of large-scale digitization and threatens the long-term trustworthiness of repositories that preserve digital surrogates. The research project has identified five types of error that compromise the trustworthiness of preserved digital surrogates: thick text, broken text, warped pages, cropped text blocks, and fully obscured content. These errors largely exist randomly in the corpus of HathiTrust digitized volumes. They are easily and somewhat reliably detectable through manual, page-by-page review. But given the scale of Google Books and HathiTrust, even small proportions can generate large numbers of catastrophic page and whole-volume error. The long-term usefulness of preservation repositories turns on our ability to review content, flag severe errors, communicate the nature of error to readers, and to fix severely flawed page images. Until the review of content quality becomes a routine function of digital-preservation repositories, questions remain about the advisability of withdrawing from libraries the hard-copy original volumes that are the sources of the surrogates.

Significant questions remain about the impact of the one percent of HathiTrust content that is nearly or completely fatally flawed. Research should proceed on four fronts. The first important area of investigation is the impact of the one percent severe error on the overall acceptance of digital surrogacy. The related question of the ubiquitous low-level artifacts challenges readers to accept digital surrogates. A second avenue of fruitful research involves finding efficient methods to find and tag severe errors and notify readers about them. It is likely that readers will themselves be marshaled as a networked crowd capable of locating the errors that are readily apparent and reporting them. A third critical area for future research is the relationship between severe image error and the quality of the underlying full-text content, which has been created via the processing of images through optical character recognition software. The optimum use for readers of HathiTrust and Google Books will be achieved only when the quality of the images and the underlying text are in synch. A fourth area of future research focuses on the descriptive and structural metadata associated with page images and full-text content. Such research ultimately evaluates the quality of entire book surrogates, not only on page images. These avenues of research converge on what may be the knottiest and most expensive issue for all preservation repositories: When error is found, what is the tradeoff between the costs and benefits of fixing errors, especially when fixing severe error may involve independent action to re-scan or re-process the images from books that are themselves far from perfect?

To preserve the products of large-scale digitization is a decision to preserve imperfection. The findings from one aspect of a multi-faceted investigation into image quality as manifested in the artifacts of error suggest that the imperfection of digital surrogates is a nearly ubiquitous feature of Google Books and that such imperfection will become firmly accepted by preservation repositories. HathiTrust has been designed to hold, protect, and deliver what is essentially becoming an online research library collection in its own right, one that reflects the flaws of the source and introduces new and more complex artifacts. For after all, preserving imperfection is an acknowledgement of the deep relationship between the material nature of our print culture and the equally certain physical aspects of our digital world.

Acknowledgements

Project planning supported by the Andrew W. Mellon Foundation; project implementation supported by a grant from the Institute for Museum and Library Services [LG-06-10-0144-10]. The author wishes to acknowledge the collaborative spirit of the University of Michigan and University of Minnesota libraries and to thank Jacqueline Bronicki, Project Coordinator; Ryan Rotter, Systems Programmer; Kenneth Guire, Statistician; Jeremy York, Associate Librarian; Stacy Maat, graduate student; and co-PI Edward Rothman, Professor of Statistics, for their indispensible work on the project.

Appendix

Distribution of Page-Level Error

		Perceived Severity of Error											
		Severity = 0		Severity = 1		Severity = 2		Severity = 3		Severity = 4		Severity = 5	
		Total	Percent	Total	Percent	Total	Percent	Total	Percent	Total	Percent	Total	Percent
Error Type	**Text**												
	Thick	58,233	62.04%	24,086	25.66%	10,070	10.73%	1.179	1.26%	183	0.19%	107	0.11%
	Broken	57,252	61.00%	28,124	29.96%	7,121	7.59%	940	1.00%	179	0.19%	242	0.25%
	Illustration												
	Scanner	92,802	98.87%	618	0.66%	344	0.37%	90	0.10%	3	0.00%	1	0.00%
	Tone	92,423	98.47%	682	0.73%	512	0.55%	109	0.12%	41	0.04%	91	0.10%
	Color	93,837	99.98%	13	0.01%	4	0.00%	1	0.00%	0	0.00%	3	0.00%
	Page												
	Blur	93,407	99.52%	286	0.30%	115	0.12%	17	0.02%	6	0.00%	27	0.03%
	Warp	27,425	29.22%	56,482	60.18%	9,545	10.17%	329	0.35%	33	0.04%	44	0.05%
	Crop	93,235	99.37%	256	0.27%	139	0.15%	65	0.07%	20	0.02%	143	0.15%
	Obscure	15,847	16.88%	73,257	78.05%	3,872	4.13%	371	0.40%	75	0.08%	436	0.46%
	Colorization	90,535	96.46%	1,580	1.68%	1,453	1.55%	228	0.24%	52	0.06%	10	0.01%
	Skew	90,591	96.52%	2,697	2.87%	538	0.57%	28	0.03%	2	0.00%	2	0.00%
	Total Error	805,587		188,081		33,713		3,357		594		1,106	

Appendix: Distribution of Page-level Error
Google-digitized, English Language, pre-1923
Totals and Proportion by Severity Level for Each Type of Error
n= 93,858 page images

References

ARL. 2012. ARL Statistics 2010-11, Rank order table 1: Volumes in Library. Washington, D.C: Association of Research Libraries.

Bailey, Charles. W., Jr. 2011. *Google Books Bibliography*. Version 7: 8/15/11. Houston: Digital Scholarship, 2005-2011. http://digital-scholarship.org/gbsb/.

Cohen, Dan. 2010. "Is Google Good for History?" *Dan Cohen's Blog*. 7 January 2010. http://www.dancohen.org/2010/01/07/is-google-good-for-history/.

Conway, Paul. 1989. "Archival Preservation: Definitions for Improving Education and Training." *Restaurator* 10.1: 47-60.

---. 2010. "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly* 80.1: 61-79.

---. 2011. "Archival Quality and Long-term Preservation: A Research Framework for Validating the Usefulness of Digital Surrogates." *Archival Science* 11.3: 293-309.

---, and Jacqueline Bronicki. 2012. "Error Metrics in Large-scale Digitization." *Proceedings of the UNC/NSF Workshop Curating for Quality: Ensuring Data Quality to Enable New Science* (NSF III #1247471), September 10-11, 2012, Arlington, VA.

Courant, Paul N. 2006. "Scholarship and Academic Libraries (and their Kin) in the World of Google." *First Monday* 11.8. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1382.

Darnton, Robert. 2010. *The Case for Books: Past, Present, and Future*. Philadelphia: Public Affairs.

DLF. 2002. *Digital Library Federation Benchmark Working Group. Benchmark for Faithful Digital Reproductions of Monographs and Serials.* Version 1. December 2002. http://old.diglib.org/standards/bmarkfin.htm.

Doermann, David, Jisheng Liang, and Huiping Li. 2003. "Progress in Camera-based Document Image Analysis." *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03),* 3.6: 606-16.

Duguid, Paul. 2007. "Inheritance and Loss? A Brief Survey of Google Books." *First Monday* 12.8 (6 August 2007). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847.

FADGI. 2010. Federal Agencies Digitization Guidelines Initiative. Still Image Working Group. Technical Guidelines for Digitizing Cultural Heritage Materials. 24 August 2010. http://www.digitizationguidelines.gov/guidelines/digitize-technical.html.

Gevinson, Alan. 2010. "Results of an Examination of 200 Digitizations [sic] of Books in the Field of American Intellectual History: Summary, Results, Data." In *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*, pp. 106-15. Council on Library and Information Resources, Washington, DC. http://www.clir.org/pubs/abstract/pub147abst.html.

Hahn, Trudi Bellardo. 2008. "Mass Digitization: Implications for Preserving the Scholarly Record." *Library Resources & Technical Services* 52.1: 18-26.

HathiTrust. 2012a. "Mission and Goals." http://www.hathitrust.org/mission_goals.

---. 2012b. "Statistics and Visualizations." http://www.hathitrust.org/statistics_visualizations.

James, Ryan. 2010. "An Assessment of the Legibility of Google Books." *Journal of Access Services* 7.4. 223-28.

Jones, Edgar. 2011. "Google Books as a General Research Collection." *Library Resources and Technical Services* 54.2: 77-89.

Jovanovic, Borko D., and Paul S. Levy. 1997. "A Look at the Rule of Three." *The American Statistician* 51.2 (May 1997): 137-39.

Karle-Zenith, Anne. 2006. "Google Book Search and the University of Michigan." In 26th Annual Charleston Conference, Charleston (US), 8-11 November 2006. e-LIS. http://hdl.handle.net/01760/9011.

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

Lavoie, Brian, Lynn Connaway, and Lorcan Dempsey. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries." *D-Lib Magazine* 11.9 (September). http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

Le Bourgeois, Frank, Éric Trinh, Bénédicte Allier, Véronique Eglin, and Hubert Emptoz. 2004. "Document Images Analysis Solutions for Digital Libraries." *Proceedings of the First International Workshop on Document Image Analysis for Libraries* (DIAL'04), Palo Alto, California: 23-24 January, pp. 2-24.

Leetaru, Kalev. 2008. "Mass Book Digitization: The Deeper Story of Google Books and the Open Content Alliance." *First Monday* 13.10 (October 6). http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2101/2037.

Lin, Xiaofan. 2006. "Quality Assurance in High Volume Document Digitization: A Survey." *Proceedings of the Second International Conference on Document Image Analysis for Libraries* (DIAL'06), 27-28 April, Lyon, France, pp. 319-26.

Madnick, Stuart. E., Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. "Overview and Framework for Data and Information Quality Research." *ACM Journal of Data Information Quality* 1.1, Article 2 (June 2009). http://doi.acm.org/10.1145.1515693.1516680.

McEathron, Scott. 2011. "An Assessment of the Image Quality in Geology Works from the HathiTrust Digital Library." *Proceedings, Geoscience Information Society* 41. http://hdl.handle.net/1808/8301.

Nunberg, Geoffrey. 2009. "Google's Book Search: A Disaster for Scholars." *The Chronicle of Higher Education* 31 (August). http://chronicle.com/article/Googles-Book-Search-A/48245/.

---. 2009. "Google Books: A Metadata Train Wreck." *Language Log, 29 August 2009*. http://languagelog.ldc.upenn.edu/nll/?p=1701.

Proskine, Emily Anne. 2006. "Google's Technicolor Dreamcoat: A Copyright Analysis of the Google Book Search Library Project." *Berkeley Technology Law Journal* 21.1: 213-40.

Rieger, Oya Yildirim. 2008. *Preservation in the Age of Large-scale Digitization: A White Paper*. Washington, D.C.: Council on Library and Information Resources.

Starmer, Mary Ellen, and Dea Miller Rice. 2004. "Surveying the Stacks: Collecting Data and Analyzing Results with SPSS." *Library Resources & Technical Services* 48.4 (October): 263-72.

Townsend, Robert B. 2007. "Google Books: What's Not to Like?" "AHA Today Blog." (April 30) http://blog.historians.org/articles/204/google-books-whats-not-to-like.

York, Jeremy J. 2009. "This Library Never Forgets: Preservation, Cooperation, and the Making of HathiTrust Digital Library." *Proceedings, IS&T Archiving 2009*, Arlington, VA, pp. 5-10.

---. 2010. “Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library.” *76th IFLA General Congress and Assembly, 10-15 August,*, Gothenburg, Sweden.