Page-Image Error in Large-Scale Digitization

Paul Conway; University of Michigan; Ann Arbor, Michigan/USA

Abstract

This paper presents and interprets data on digitization error gathered from four 1,000 volume random samples that represent the full range of source volumes digitized by Google and the Internet Archive over a six year period and deposited in the HathiTrust Digital Library. The paper summarizes the research method for the project and then presents summary findings on the distribution of page-image error. The findings suggest that the imperfection of digital surrogates is a transparent and nearly ubiquitous attribute of large-scale digitization and one that introduces new complexity in preservation repositories. The paper concludes with suggestions for further research.

Introduction

Large-scale digitization efforts by third-party corporations and non-profits are controversial, none more so than Google Books and the Internet Archive [2]. Some of the major concerns expressed in the mix of scholarly and popular media include the dangers of corporate control of research resources [9], the legality of wholesale digitization [21], inadequate and incomplete coverage of intellectual disciplines [13], poor search and discovery results [19], and the secrecy that surrounds Google's digitization workflows [15]. Oya Rieger [22] explored the preservation implications of four large-scale projects and concluded that some of the most serious problems have to do with the quality of the page images displayed to the reader, the metadata associated with digital surrogates, and the underlying full text data that makes text searchable. A litany of complaints from scholars, librarians and archivists, and technologists about image quality fuels an ongoing debate about the appropriateness of large-scale book digitization and the advisability of preserving the resulting products.

The HathiTrust Digital Library, a partnership of sixty-plus research libraries, is the test bed for the research. Its origins and continued vitality as a preservation repository and a service for stakeholders are inextricably tied to large-scale digitization by Google and other third-party digitizers [8]. HathiTrust now (2013) contains well over 10 million digitized volumes, 96.4 percent of which have been digitized by Google from the contents of at least 18 library collections [23, 24]. The digital surrogates in HathiTrust encompass 429 languages across the spectrum of library classification and the history of books and printing since Gutenberg [11]. HathiTrust now ranks among the largest of the 126-member Association of Research Libraries [1].

This paper reports the core findings on the quality of digital page-images in HathiTrust form an assessment of four large random samples that represent much of the digital content presently deposited in HathiTrust. This report is one component of a multi-year, multi-method research project that has established an error model for large-scale digitization of books and serials and applied the model 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. 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 [6, 7].

Research Context

As a field of study, digital information quality has established a strong foundation of research and theoretical scrutiny since at least the mid-1990s [17]. The literature on information quality, however, is relatively silent on how to measure quality attributes of very large collections of digitized books and journals, created as a combination of page images, full-text data, and underlying XML [16, 3]. Little systematic research has been completed on the digitization quality of Google Books. James [12] conducted a small random-sample study of text legibility in Google Books 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. McEathron [18] evaluated a random sample of 180 volumes on geology topics, from a population of over 2,500 volumes in HathiTrust. He found a 2.5 percent rate of scanning errors very thinly but widely distributed through 63 percent of the sample.

American intellectual historian Alan Gevinson [10] reaches beyond the personalized, impressionistic treatment of image error in Google Books, using a list of 200 influential books in the field. He found a very 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 of the volumes with cropped or obscured text. Gevinson's study suffers from challenges he had in finding and viewing specific titles, but more important from a lack of clarity about error definitions and little effort 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. His research points the way toward the possibility of a systematic and predictive study of quality.

In the context of large-scale digitization, thousands or millions of objects are scanned against a single digitization technical specification in a factory like workflow [15]. In this study the quality of large-scale digitization is not defined in terms of the properties of the raster-image surrogate of a book, but instead as the absence of visible artifacts, in the form of process and processing errors, that interfere with using that book in digital form. The assessment of quality in large-scale digitization thus must begin with the measurement of absolute error in a given population of digital surrogates. When the extent of absolute error is understood 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 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 what can be done to address this error.

Methodology

At the heart of the research project is a three-tiered hierarchical model that hypothesizes error at the levels of text/illustration, page-image, and whole volume and that assigns one or more potential causes for each error (source volume, scanning, post-scan manipulation) [5, 6, 7]. 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 in Figure 1, 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.

LEVEL 1: DATA/INFORMATION

1.1 Text: thick text [fill, excessive]
1.2 Text: broken text [character breakup]
1.3 Illustration: scanner effects [moiré patterns]
1.4 Illustration: tone, brightness, contrast
1.5 Illustration: color imbalance, gradient shifts

LEVEL 2: ENTIRE PAGE

2.1 Blur [distortion]
2.2 Warp [text alignment]
2.3 Skew [page alignment]
2.4 Crop [gutter, text block]
2.5 Obscured [portions not visible]
2.6 Colorization [text bleed, low contrast]

LEVEL 3: WHOLE VOLUME

3.1 Fully obscured [foldouts]
3.2 Missing pages [one or more]
3.3 Duplicate pages [one or more]
3.4 Order of pages
3.5 False pages [not part of original content]

Figure 1. Model of error in large-scale digitization

The overall population of study described here consists of books and serials digitized by Google between 2004 and 2010 and by the Internet Archive's Open Content Alliance project over a similar time frame. To help assure the representativeness of the study, the project team used a two tier sampling strategy to draw four 1,000 volume random samples from sub-populations of HathiTrust content: 1) English language books and serials published before 1923 and digitized by Google; 2) English language books published after 1922 and digitized by Google; 3) books in the public domain digitized by Internet Archive; and 4) Google-digitized books from any time period published in four non-Roman scripts: Arabic, Asian, Cyrillic, and Hebrew. Within the 1,000 volume sample, the project team extracted a systematic random sample of approximately 100 pages within each volume. The sample size and in-volume sampling strategy allows for statistical comparison of sub-populations with potentially very small frequencies in important variables [14] and insures 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.

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 perceived to be present 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 severe error. The project developed a highly efficient and statistically reliable data gathering and analysis system to measure error-incidence in HathiTrust volumes. The data gathering process produced accurate, complete, and well formed data sets for each of the four samples. This approach to data management provides 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.

Findings on Page-image Error

Four samples of 1,000 volumes yielded a total of 356,217 page images for manual review. What follows are summary findings on page-image error for portions of the four samples. The first section juxtaposes data on the most common errors in Google-digitized books represented in two samples. The second section presents the distribution of error in the sample of Internet Archive-digitized books. The third section highlights particular issues with books published in Asian languages. The fourth section is an overview of the most severe error across all four samples. Much more data has been gathered than is reported here. The focus is on the distribution of page-image error in four large samples and some additional insights gained from aggregating individual page errors to the volume level.

Google Books

Coders had significant difficulty applying the five-level severity coding scheme to page-images with digitized illustrations. The data 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 embedded illustrations and graphic material. A special 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 between 82.5 and 96.9 percent of all perceived error at any level. Table 1 presents the distribution of the severity of error

Table 1. Most common error in Google Books (n=93,858 page images pre-1923; 86,439 post-1922)

		Severity 0		Severity 1		Severity 4		Severity 5	
		before 1923	after	before 1923	after	before 1923	after	before 1923	after
Error Type	*Thick Text*	62.0%	67.5%	25.7%	21.0%	0.19%	0.40%	0.11%	0.42%
	Broken Text	61.0%	73.4%	30.0%	19.2%	0.19%	0.41%	0.25%	0.36%
	Cropped Page	99.4%	98.9%	0.3%	7.1%	0.02%	0.04%	0.15%	0.25%
	Warped Page	29.2%	45.8%	60.2%	48.9%	0.04%	0.04%	0.05%	0.06%
	Obscured Content	16.9%	56.8%	78.1%	41.7%	0.08%	0.02%	0.46%	0.16%
	Total Error Detected pre-1923			182,205		490		972	
	Proportion at Severity Level			96.9%		82.5%		87.9%	
	Total Error Detected post-1923				113,682		795		1,077
	Proportion at Severity Level				90.5%		87.9%		86.5%

across these five errors, with a special emphasis on comparing minor and very severe error rating. Errors coded at severity level one are perceptible to reviewers but have no impact on information content. 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, on the other hand is catastrophic. Original content in the affected area of the page cannot be unambiguously deciphered or has been obscured.

The representation of English-language text in Google-digitized page images is problematical at both extremes of severity. Table 1 shows that between 19.2 and 30 percent of all page-images in the sample display some level of text distortion on some portion of the image. About a quarter of all images reviewed yielded evidence of low-severity thick text or broken text. Thick text appears to the reader as bolded in a way that is not typographical in nature. Broken text poses the opposite challenge; readability is compromised 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 the words formed by discrete characters. In Google digitized volumes, thick and broken text tend to co-occur on the same page image, but the statistical association is weak. All other errors in the model are statistically independent.

Extreme distortion is very rare in both samples. Less than one-half of one percent of pages reviewed are nearly or completely indecipherable. The large sample size yields a 95 percent confidence level that this very small proportion of catastrophic text error represents the predicted severe error in the overall population of Google-digitized volumes. Errors in text rendering may affect users in different ways. Pervasive low level error in text may affect concentration in online reading or undermine the desirability of digital surrogacy for some users. Psychologists have labeled this phenomenon "cognitive stress" and have demonstrated how such challenges to mental fluency influence the decisions [20].

Beyond the challenge of accurately rendering text characters, a review of the sample page-images also revealed three important page-level errors: warped or cropped pages and obscured content. Over half of all page images do not appear flat when viewed. The subtle effect of warping is a byproduct of Google's patented post-scan processing algorithms that attempts to remove the appearance of curvature that results when volumes are scanned in their bindings. When the algorithm fails completely, 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 the presence of artifacts of Google's scanning process or incomplete or failed efforts to remove these artifacts through post-scan processing, observable in over three quarters (78.0%) of page-images reviewed in sample one and 41.7 percent reviewed in sample two. Dan Cohen [4] and other commentators have been quick to comment on the pink-tipped 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. The proportion of severe error perceived in page-images is in keeping with the findings of McEathron [18] and James [12] but is far less than the error recorded by Gevinson [10].

Internet Archive Books

The research team drew a third random sample from volumes digitized by the Internet Archive under terms of its collaborative Open Content Alliance project. The volumes are all in the public domain and are available online through the Internet Archive interface. Approximately 305,000 digital surrogates are deposited in HathiTrust. Table 2 displays the frequency and severity of error of detected in the sub-sample of 84,539 page images selected systematically from the sample population of 1,000 volumes.

Table 2. Frequency and severity of error in Internet Archive volumes (n=84,539 page images from 1,000 volumes)

Error Type	Severity 0		Severity 1		Severity 4		Severity 5	
Broken Text	69,076	81.7%	9,952	11.8%	158	0.19%	81	0.10%
Thick Text	78,819	93.2%	3,483	4.1%	6	0.01%	2	0.00%
Warped Page	34,772	41.1%	48,184	57.0%	2	0.00%	0	0.00%
Colorized Page	40,204	47.6%	38,230	45.2%	4	0.00%	23	0.03%
Skewed Page	76,291	90.2%	7,890	9.3%	0	0.00%	0	0.00%
Blurred Page	79,654	94.2%	3,618	4.3%	34	0.04%	27	0.03%
Cropped Page	84,211	99.6%	187	0.2%	16	0.02%	63	0.07%
Obscured Content	48,127	56.9%	33,487	39.6%	17	0.02%	60	0.07%

Overall, Internet Archive volumes display significantly fewer errors than those digitized by Google. In particular, virtually no errors exist at severity levels four and five; no single error occurs at these levels more frequently than one-fifth of one percent of the sample. The more commonly occurring errors at severity level one are perceptible to the naked eye when displayed at 100% fidelity on a high resolution monitor but do not affect the readability of textual or illustrated content. The most common low-severity textual error is broken text, which occurs in 11.8 percent of the page-images viewed. Thick text is far less prevalent at low severity levels, perceived in just 4.1 percent of the sample pages.

Page-level digitization errors are more common in the Internet Archive population. In particular, the relatively common occurrence of gently warped (57.0%), skewed (9.3%), and blurred (4.3%) page images reflects the intensely manual scanning procedures that utilize traditional glass-covered book cradles that facilitate non-destructive scanning and the minimal post-scan manipulation. Such scanning techniques also contribute artifacts to the page image that register as severity-level-one obscured content. Over 39 percent of all Internet Archive page images reviewed showed signs of dust, glass smudges, and other imperfections.

The colorization of page images, however, apparently represents a conscious effort by the Internet Archive to reduce the effect of high contrast scanning on the display of digital surrogates. When applied sensitively, colorization adds the impression of ageing to what would otherwise appear to be black text on a white background. When the colorization algorithm appears to be over-applied, reviewers assigned a severity level of one to this error. In the Internet Archive sample, low-severity colorization error occurs on 45.2 percent of all page images reviewed.

Non-Roman Languages

The research team drew a fourth 1,000 volume random sample of books digitized by Google and printed in four non-Roman scripts: Arabic, Asian (Chinese, Japanese, or Korean), Cyrillic, and Hebrew. The overall population of such publications in HathiTrust exceeds 1.29 million volumes. Trained reviewers with native fluency in the appropriate language inspected 250 surrogate volumes for each of the four scripts and coded error using the identical methods utilized for samples of English-language volumes digitized by Google and the Internet Archive.

With one major exception, the frequency and severity of digitization error in the volumes of non-Roman scripts nearly is identical to that of English-language volumes. The same five error types (thick and broken text, warped and cropped pages, and obscured content) dominate the error landscape of these volumes. There is no statistically significant difference in the distribution of the errors across the five-level severity scale between the three samples of English-language and non-Roman scripts.

The exception to this general conclusion pertains to the representation of Asian-language characters. Table 3 presents the data for the five most common errors in the 250 Asian-language volumes included in the random sample. The table shows that 21.8 percent of the 28,952 pages reviewed show evidence of thick text at severity levels four or five and an additional 28.2 percent thick text at severity levels two and three. At the most severe levels, text is nearly or completely unintelligible to native Asian language

Table 3. Frequency and severity of error in Asian-language Google-digitized volumes (n=28,952 page images)

Error Type	Severity 0 or 1		Severity 2 or 3		Severity 4 or 5	
	Total	Percent	Total	Percent	Total	Percent
Broken Text	22,860	79.0%	5,293	18.3%	799	2.8%
Thick Text	14,461	50.0%	8,177	28.2%	6,314	21.8%
Warped Page	26,481	91.5%	2,450	8.5%	21	0.1%
Obscured Content	28,744	99.3%	94	0.3%	114	0.4%

speakers. At moderate levels of severity, readability is affected adversely but the text is intelligible with effort. Only half of the pages images reviewed had text that was not affected by thickening. Similar but less dramatic findings are also evident with the broken-text error. Some 18 percent of the page images show evidence of text break-up at severity levels two or three.

Bad Books

In the presentation of findings above, digitization error is assessed in a population of page-images without regard for their clustering in published volumes. Table 4 presents the findings for digitized volumes that contain page images with error perceived at severity levels four or five. At these severity levels, error renders a page nearly or completely unintelligible. The table shows the number and proportion of volumes in each of the four 1,000 volume samples that have zero, one, and multiple page images with very severe error of any type. Such errors may occur with text on the page, with the page itself, or with obscured content due to digitization or post-scan processing. A test of the sequencing of severe errors in digitized volumes found no statistical relationship between the presence of severe error and location in the volume.

The table demonstrates the contrast in error incidence between books digitized by Google and the Internet Archive. Over 92 percent of the Internet Archive volumes display no severe error. For Google-digitized English-language volumes, 59.5 percent of public domain volumes have no severe error, while a much higher proportion of volumes (69.2%) published since 1922 display no severe error. At the other end of the spectrum, relatively few volumes in three of four samples have more than eleven pages with very severe error. Only four of the 944 books digitized by the Internet Archive display frequent severe error, with the poorest quality volume in the sample measuring 27 page images with severe error. The table also reinforces the apparent challenges that Google digitization presents for books in non-Roman scripts. Just half of the volumes reviewed contain no perceived severe error, while a third of the 250 volumes reviewed are perceived to have multiple page images with very severe error. These results are influenced by the digitization errors with Asian-language text but may also be exacerbated by digitization error for books in other languages.

Implications for Preservation

Large-scale digitization is a phenomenally productive method for producing digital surrogates of books. The partial findings on the frequency and severity of error reported here suggest that high production carries a small risk of random unintelligibility. It is premature to present firm conclusions about the relationship between infrequent but truly serious error and the usefulness of digital volumes as a whole. But the data reported in this article may yield some tentative conclusions about the implications of preserving digital surrogates from large-scale digitization efforts by Google, the Internet Archive, and other sources.

The first conclusion is that minor error that does not limit the readability of digitized text might be accepted as a part of the price of enhanced access. Only a minority of the volumes in HathiTrust are error free at severity levels one and two. These errors are easily

Table 4. Distribution of severe error in four random samples

Sample 1: Google-digitized Volumes

English language, published before 1923

Pages / Volume	Number of Volumes	Proportion of Sample
0	555	59.5%
1	167	17.9%
2 to 10	182	19.5%
11 to 68	28	3.0%
	932	100.0%

Sample 2: Google-digitized Volumes

English language, published after 1922

Pages / Volume	Number of Volumes	Proportion of Sample
0	637	69.2%
1	131	14.2%
2 to 10	115	12.5%
11 to 168	38	4.1%
	921	100.0%

Sample 3: Internet Archive-digitized Volumes

English language, all volumes in public domain

Pages / Volume	Number of Volumes	Proportion of Sample
0	876	92.80%
1	43	4.56%
2 to 10	21	2.22%
11 to 27	4	0.42%
	944	100.00%

Sample 4: Google-digitized Volumes

Arabic, Asian, Cyrillic, Hebrew Scripts

Pages / Volume	Number of Volumes	Proportion of Sample
0	510	51.0%
1	139	13.9%
2 to 10	182	18.2%
11 to 289	169	16.9%
	1000	100.0%

detectable and are so common as to be a part of the fabric of digital surrogacy. Low-level quality errors with text and illustration are not confined to online delivery, but also make their way into secondary products, including print-on-demand copies and versions prepared for eBook readers. It is likely infeasible and perhaps undesirable to continue to process and reprocess digital surrogates to remove low-level error. The existence of millions of digitized volumes presents these same 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 library holdings of books and serials to create cleaner and more pristine representations of source volumes.

A second conclusion is that although minor error could become an acceptable feature of large-scale digitization, extremely severe error compromises the integrity of large-scale digitization and threatens the long-term trustworthiness of repositories that preserve digital surrogates. With the exception of Asian language text, these near fatal errors largely exist randomly and in very small proportions in the corpus of HathiTrust volumes digitized by Google and the Internet Archive. The long-term viability of preservation repositories turns on the ability to review content, flag severe errors, communicate the nature of error to readers, and to set in motion processes to fix severely flawed page images.

Significant questions remain about the impact of the one percent of HathiTrust content that is nearly or completely fatally flawed. Research should proceed on three fronts. The first important area of investigation is the impact of the one percent severe error on the overall acceptance of digital surrogacy. It may be that for most users the value of having millions of books available online overrides the occasional unreadable page image in an otherwise intelligible surrogate volume. Research is also needed on the extent to which ubiquitous low-severity error places unacceptable cognitive strain on the end user, leading them to seek alternative sources [20].

A second avenue of fruitful research involves finding efficient methods to find, tag, and communicate the presence of severe error to readers. It is possible and likely that readers may themselves be marshaled as a networked crowd capable of applying consistent judgment to the quite specific errors that are readily apparent. These two avenues of research converge on what may be the knottiest and most expensive issue facing all preservation repositories: the tradeoff between the costs and the benefits of fixing errors, especially when addressing severe error may involve independent action to re-scan or re-process the images from books that are themselves far from perfect.

A third and potentially fruitful 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. It is possible that the errors present in underlying full text can be used to predict the existence of visible error in the associated page images. Similarly, severe error in pages images detected through manual inspection can point to localized weakness in associated full text. The full potential for users of HathiTrust and Google Books will only be achieved when the quality of the images and underlying text are in synch.

The HathiTrust Digital Library 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, and deposited and preserved collaboratively. The findings from one aspect of a multi-faceted investigation into the quality of the digital surrogates suggest that the imperfection of digital surrogates is a transparent and nearly ubiquitous attribute, one that reflects the flaws of the source and introduces new and more complex artifacts in preservation repositories.

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 thank Jacqueline Bronicki, Project Coordinator; Ryan Rotter, Systems Programmer; Kenneth Guire, Statistician; and Jeremy York, Associate Librarian.

References

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

[2] Bailey, C. W., Jr. 2011. *Google Books bibliography*. Version 7: 8/15/11. Houston: Digital Scholarship, 2005-2011. http://digital-scholarship.org/gbsb/

[3] Baird, H. 2004. "Difficult and urgent open problems in document image analysis for libraries," *Proc. Of International Workshop on Document Image Analysis for Libraries* ? (?): 25-32.

[4] Cohen, D. 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/

[5] Conway, P. 2010. "Preservation in the age of Google: Digitization, digital preservation, and dilemmas." *Library Quarterly* 80 (1): 61-79.

[6] Conway, P. 2011. "Archival quality and long-term preservation: A research framework for validating the usefulness of digital surrogates." *Archival Science* 11 (3): 293-309.

[7] Conway, P. and J. 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.

[8] Courant, P. N. 2006. "Scholarship and academic libraries (and their kin) in the world of Google." *First Monday* 11, no. 8. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1382

[9] Darnton, R. 2010. The case for books: past, present, and future. Philadelphia: PublicAffairs.

[10] Gevinson A. 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. 106-115. Council on Library and Information Resources, Washington, DC. http://www.clir.org/pubs/abstract/pub147abst.html .

[11] HathiTrust. 2012. Statistics and visualizations. http://www.hathitrust.org/statistics_visualizations.

[12] James, R. 2010. An assessment of the Legibility of Google Books. *Journal of Access Services* 7 (4). 223-228.

[13] Jones, E. 2011. Google Books as a general research collection. *Library Resources and Technical Services* 54 (2): 77-89.

[14] Jovanovic, B. D. & P. S. Levy. 1997. "A look at the rule of three." *The American Statistician* 51 (2): 137-139.

[15] Leetaru, K. 2008. Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday* (Online) v. 13 no10 (October 6). http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2101/2037

[16] Lin, X. 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. 319-326.

[17] Madnick, S. E., Y. W. Lee, R. Y. Wang, and H. 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 .

[18] McEathron, S. 2011. An assessment of the image quality in geology works from the HathiTrust Digital Library. *Proceedings, Geoscience Information Society*, vo. 41. http://hdl.handle.net/1808/8301

[19] Nunberg, G. 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/?sid=at&utm_source=at&utm_medium=en ;

[20] Oppenheimer, D.M., 2008. The secret life of fluency. *Trends in Cognitive Sciences* 12, no. 6: 237-241.

[21] Proskine, E. A. 2006. Google's technicolor dreamcoat: A copyright analysis of the Google Book Search library project. *Berkeley Technology Law Journal* 21, no. 1: 213-240.

[22] Rieger, O. Y. 2008. Preservation in the age of large-scale digitization: A white paper. Washington, D.C.: Council on Library and Information Resources.

[23] York, J. J. 2009. This library never forgets: Preservation, cooperation, and the making of HathiTrust digital library. *Proc. IS&T Archiving 2009*, Arlington, VA, pp. 5-10.

[24] York, J. J. 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.

Author Biography

Paul Conway is associate professor in the School of Information at the University of Michigan. His research encompasses archival science, the digitization of cultural heritage resources, particularly photographic archives, and the measurement of image and text quality in large-scale digitization programs. He holds the PhD in information studies from the University of Michigan (1991). He is a Distinguished Fellow of the Society of American Archivists.