MEASURING CONTENT QUALITY IN A PRESERVATION REPOSITORY: HATHITRUST AND LARGE-SCALE BOOK DIGITIZATION

Paul Conway
University of Michigan
School of Information
105 South State Street
Ann Arbor, MI 48109-1285
pconway@umich.edu

ABSTRACT

As mechanisms emerge to certify the trustworthiness of digital preservation repositories, no systematic efforts have been devoted to assessing the quality and usefulness of the preserved content itself. With generous support from the Andrew W. Mellon Foundation, the University of Michigan's School of Information, in close collaboration with the University of Michigan Library and HathiTrust, is developing new methods to measure the visual and textual qualities of books from university libraries digitized by Google, Internet Archive, and others and then deposited for preservation. This paper describes a new approach to measuring quality in large-scale digitization; namely, the absence of error relative to the expected uses of the deposited content. The paper specifies the design of a research project to develop and test statistically valid methods of measuring error. The design includes a model of understanding and recording errors observed through manual inspection of sample volumes, and strategies to validate the outcomes of the research through open evaluation by stakeholders and users. The research project will utilize content deposited in HathiTrust – a large-scale digital preservation repository that presently contains over five million digitized volumes – to develop broadly applicable quality assessment strategies for preservation repositories.

1. INTRODUCTION

The large-scale digitization of books and serials is generating extraordinary collections of intellectual content that are transforming teaching and scholarship at all levels of the educational enterprise. Along with burgeoning interest in the technical, legal, and administrative complexities of large-scale digitization [2], significant questions have risen regarding the quality and fitness for use of digital surrogates produced by third-parties such as Google or the Internet Archive. Until recently, those who built digital repositories also exercised significant control over the creation of digital content, either by specifying digitization best practices [24] or by limiting the range of digital content forms accepted for deposit and long-term maintenance [29]. For an institution and its community of users to trust that individual digital objects created by third parties are accurate, complete, and intact and to know that objects deposited in preservation repositories have the capacity to meet a variety of uses envisioned for them by different stakeholders, repositories must validate the quality and fitness for use of the objects they preserve.

Information quality is an important component of the value proposition that digital preservation repositories offer their stakeholders and users [12]. For well over a decade, the cultural heritage community of libraries, archives, and museums has embraced the need for trustworthy digital repositories with the technical capacity to acquire, manage, and deliver digital content persistently [42]. During the past decade, standards-based mechanisms for building and maintaining repository databases and associated metadata schema have emerged to enable the construction of preservation repositories on a scale appropriate to the preservation challenge at hand [26][19]. Significant progress has been made in establishing the terms and procedures for certifying trustworthiness through independently administered auditing processes [40]. In the new environment of large-scale digitization and third-party content aggregation, however, certification at the repository level alone may be insufficient to provide assurances to stakeholders and end-users on the quality of preserved content. One of the grand challenges of digital preservation is for repositories to establish the capacity to validate the quality of digitized content as "fit-for-use," and in so doing provide additional investment incentives for existing and new stakeholders.

2. LITERATURE REVIEW

Critics of quality: Although large-scale digitization programs have their vocal advocates [13], scholars, librarians, and the preservation community increasingly are raising concerns about the quality and usability of image and full-text products [34]. For example, Bearman [4], Duguid [18], and Darnton [14] cite scanning and post-production errors in early iterations of Google's book digitization program. Tanner [39] finds a high level of error in text conversion of newspapers. S. Cohen [9] suggests that quality issues will arise most strikingly when entire books are printed on demand. Schonfeld [37] concludes that only the full comparison of original journal volumes with their digital surrogates is sufficient before hard copies can be withdrawn from library collections. Attempting to sort through the commentary, D. Cohen [18] identifies a fundamental need for research: "of course Google has some poor scans—as the

saying goes, haste makes waste—but I've yet to see a *scientific* survey of the overall percentage of pages that are unreadable or missing (surely a miniscule fraction in my viewing of scores of Victorian books)."

Information quality definitions: The quality of digital information has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. At a definitional level, Garvin [20] identifies five discrete approaches to understanding quality, two of which – product-based and user-based – are particularly relevant to the proposed research. Models for information quality have emerged from important empirical research on data quality [41] and have been adapted for the Internet context [25] Research derived from business auditing principles [6] and information science theory [35] grounds the analysis of information quality in the language of credibility and trust. Research informed by archival theory has also addressed the importance of information quality [43]. Although the emergent models are quite inconsistent in terminology, they provide a comparable theoretical foundation for research on quality in large-scale digitization. The research design described here joins the relatively objective product-based findings on digitization quality with the more subjective evaluation judgments of a user-based approach.

Fitness for use: Stvilia [38] builds on the commonality that exists in information quality models, and focuses special attention on the challenge of measuring the relationship between the attributes of information quality and information use. In adopting the marketing concept of "fitness for use," he recognizes both the technical nature of information quality and the need to contextualize "fitness" in terms of specific uses. Stvilia establishes and tests a useful taxonomy for creating quality metrics and measurement techniques for "intrinsic qualities" (i.e., properties of the objects themselves). In the context of digitization products, intrinsic quality attributes are objectively determined technical properties of the digitized volume, derived from the results of digitization and post-scan image processing. By distinguishing measurable and relatively objective attributes of information objects from the usefulness of those objects, Stvilia establishes a viable research model that can be applied to the measurement of the quality of digitized books within particular use-cases.

Use-cases: Quality judgments are by definition subjective and incomplete. From the perspective of users and stakeholders, information quality is not a fixed property of digital content [11]. Tolerance for error may vary depending upon the expected uses for digitized books and journals. Marshall [31, p. 54] argues that "the repository is far less useful when it's incomplete for whatever task the user has in mind." Baird makes the essential connection between quality measurement and expected uses in articulating the need for research into *goal directed metrics* of document image quality, tied quantitatively to the reliability of downstream processing of the images." [3, p. 2] Certain fundamental, baseline capabilities of digital objects span disciplinary boundaries and can be predicted to be important to nearly all users. Use-cases articulate what stakeholders and users might accomplish if digital content was validated as capable of service-oriented functions [7].

Error measurement: The literature on information quality 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 and full-text data by third party vendors. Lin [28] provides an excellent review of the state of digital image analysis (DIA) research within the context of large-scale book digitization projects. Because Lin's framework is determined by ongoing DIA research problems, his "catalog of quality errors," adapted from Doermann [17], may be overly simplistic; but his work is most relevant because it distinguishes errors that take place during digitization [e.g., missing or duplicated pages, poor image quality, poor document source] from those that arise from post-scan data processing [e.g., image segmentation, text recognition errors, and document structure analysis errors]. Lin recognizes that, in the future, quality in large-scale collections of books and journals will depend on the development of fully automated analysis routines. The state of the art in quality assurance today depends in large measure upon manual visual inspection of digitized surrogates or the original book volumes [27]. Although the research design is oriented toward the possibility of eventual automated quality assurance, data gathering will be based fundamentally on manual review of statistically valid samples of digitized volumes.

3. HATHITRUST TEST BED

HathiTrust is a digital preservation repository that was launched in October 2008 by a group of 25 research universities, including the Committee on Institutional Cooperation [the Big Ten universities and the University of Chicago] and the University of California system.[1] At present [July 2010] HathiTrust consists of 6.2 million digitized volumes ingested from multiple digitization sources (primarily Google). HathiTrust is a large-scale exemplar of a preservation repository containing digitized content 1) with intellectual property rights owned by a variety of external entities, 2) created by multiple digitization vendors for access, and 3) deposited and held/preserved collaboratively. HathiTrust is also a technological environment for collaboratively addressing challenges in duplication, collection development, and digital preservation that are common to all libraries. The repository is in the midst of a rigorous certification audit by the Center for Research Libraries using the TRAC [40] framework. HathiTrust is supported by base funding from all of its institutional partners, and its governing body includes top administrators from libraries and information offices at investing institutions [44].

HathiTrust is highly organic, posing interesting challenges for quality assessment, and at the same time

[1] HathiTrust. http://www.hathitrust.org/

making it an ideal test-case for quality research. Large portions of HathiTrust can amount to an information quality "moving target," because the repository overlays existing copies of works digitized by Google with improved versions as Google makes those versions available (between 100,000 and 200,000 volumes are improved and replaced in this way each month, on average). HathiTrust also is growing rapidly, having increased in size by a monthly average of 230,000 volumes in 2009. This volubility challenges the assignment of quality projections across the entire repository. HathiTrust, however, possesses the technical infrastructure and the type of digital content required to develop quality metrics, validate those metrics with users, and assess quality changes over time. The findings of this research will be broadly applicable to the current digital repository environment, ranging from smaller and somewhat stable repositories to large-scale evolving digital preservation services such as HathiTrust.

4. DIGITIZATION QUALITY AND ERROR

The research design is innovative in part for its effort to rethink what quality means within the context of preserved digital content. Until very large-scale digitization forced this issue to the forefront, the preservation community attempted to influence digitization quality through adherence to best practices that the community itself promulgated [24]. Successful implementation of guidelines enables the vertical integration of content creation, content delivery, and content preservation at a scale that seemed large ten years ago but which now pales in comparison to the efforts of third party digitizers such as Google. With vertical integration also comes the possibility of controlling digitization workflows that span the entire conversion-to-preservation process.

Today's digital content environment is marked by distributed responsibility for content creation and a trend toward collaborative responsibility for long-term preservation and access [10]. Increasingly, preservation repositories take what they can get, with, at best, assurances from the publisher/creator that the submitted content meets the original purposes or those deemed appropriate by the creator/publisher [30]. In a distributed content creation environment, it may be both infeasible and inappropriate to validate digitization quality against a community "gold standard."[2] Rather, preservation repositories may have to establish benchmarks that represent the best efforts of the content creator. Such a "bronze standard" recognizes the limitations of large-scale digital conversion and reorients quality assurance toward detecting and remedying errors that may occur at stages of the conversion process.

Within the context of a large-scale preservation repository, our research adapts Stvilia's [38] model of intrinsic quality attributes and Lin's [28] framework of errors in book surrogates derived from digitization and post-scan processing. The error measurement model for the project design recognizes that errors originate from some combination of problems with (a) the source volume (original book), (b) digital conversion processes (scanning and OCR conversion), and (c) post-scan enhancement processing. The research design draws on data from four years of quality review compiled by the University of Michigan Library (MLibrary) as part of the ingest of over five million volumes into HathiTrust. The MLibrary quality review manual, which defines and illustrates eight digitization errors evaluated in books deposited in HathiTrust for the past three years, is available online.[3]

Table 1 presents the distribution of critical level of eight errors identified by University of Michigan library staff over a four-year period. A critical error is one whose presence in one or more of a random sequence of 20 pages is sufficiently severe to render the volume unusable. The table shows the total number of volumes ingested into HathiTrust in a given year, the total number and total percentage of volumes inspected for errors using an online logging system built at Michigan. The summary inspection data shows a declining proportion of volumes inspected over time, due to confidence in the inspection process garnered after the first two years of quality assurance work across approximately 70,000 volumes. The table also shows the relatively low rate of critical error and the low absolute number of volumes with critical errors. Errors in post-scan image manipulation (cleaning, colorization, cropping) account for a very large portion of the errors logged. The number of volumes with errors in a given year cannot be totaled, due to the fact that volumes with errors most likely display multiple types of critical error. For example, volumes with warped pages are also likely to have pages with blurred text. The research design adjusts for a flaw in the Michigan model of error inspection, which does not allow for disambiguating error incidence.

The research design builds on the Michigan error detection framework, first by determining the nature and level of intrinsic quality error at three levels of abstraction: (1) data/information; (2) page-image; (3) whole volume as a unit of analysis. Within each level of abstraction exist a number of possible errors that separately or together present a volume that may have limited usefulness for a given user-case scenario. At the data/information level, a volume should be free of errors that inhibit interpretability of text and/or illustrations viewed as data or information on a page. At the page-image level, a volume should be free of errors that inhibit the digital representation of a published page as a whole object. At the whole-volume level, a volume should be free of errors that affect the representation of

[2] Federal Agencies Digitization Guidelines Initiative. http://www.digitizationguidelines.gov/

[3] http://www.hathitrust.org/documents/UM-QR-Manual.pdf

Critical Error Type	*Cause*	*May 2006-April 2007*		*May 2007-April 2008*		*May 2008-April 2009*		*May 2009-April 2010*		*TOTAL*
Thick text	scanning	189	0.57%	70	0.19%	19	0.06%	144	0.81%	422
Broken text	scannng	518	1.57%	121	0.33%	76	0.26%	64	0.36%	779
Blurred text	scanning	252	0.76%	40	0.11%	10	0.03%	54	0.30%	356
Obscured text	source	57	0.17%	35	0.09%	21	0.07%	8	0.04%	121
Warpped page	post-scan	47	0.14%	37	0.10%	14	0.05%	22	0.12%	120
Cropped text block	post-scan	424	1.28%	246	0.67%	100	0.34%	67	0.38%	837
Cleaning	post-scan	208	0.63%	214	0.58%	1256	4.23%	439	2.46%	2117
Colorization	post-scan	3250	9.83%	272	0.74%	35	0.12%	19	0.11%	3576
Volumes ingested		288,044		460,620		2,523,049		1,665,167		4,936,880
Volumes reviewed (20 pages/vol.)		33,047		36,981		29,677		17,850		117,555
Ingested/Received		11.47%		8.03%		1.18%		1.07%		2.38%

Table 1. Incidence of critical error in volumes ingested into HathiTrust, 2006-10.

the digital volume as a surrogate of a book. Errors originate from some combination of problems with the source volume (original book) or digitization (scanning, post-processing).

A major goal of the study is to define meaningful distinctions in severity of error and to validate those distinctions within specific use cases. The project design's error incidence model in Table 2 modifies the Michigan error model (bolded items) by adding reference to possible errors with book illustrations [23], OCR full-text errors, and errors that apply fully to an entire volume. Error detection must account for frequency and severity and be contextualized by level of abstraction. The development of specific judgments of severity of error requires assessment on ordinal scales instead of the binary distinctions between critical and non-critical error utilized presently.

5. RESEARCH MODEL AND METHODOLOGY

The overall design of the research project consists of two overlapping investigative phases. Phase one will define and test a set of error metrics (a system of measurement) for digitized books and journals. Phase two will apply 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. The design of each phase is anchored by a specific research question that drives the associated data gathering, analysis, and user validation activities.

We refer to "validation" in our research model in two ways that expressly bridge the product-based findings and the user-based approaches to quality. First, validation also refers to the procedures that engage users in identifying the distinctive combination of digitization errors that apply to a given use-case. Second, validation refers to the data analysis routines that demonstrate the statistical power of the error analysis to measure the difference between observed and benchmarked volumes. Validation through user-based feedback provides a "reality check" that statistically determined findings on quality properly describe the" fitness for use" of digitized volumes.

LEVEL 1: DATA/INFORMATION

1.1 Image: thick [character fill, excessive bolding, indistinguishable characters]
1.2 Image: broken [character breakup, unresolved fonts]
1.3 Full-text: OCR errors per page-image
1.4 Illustration: scanner effects [moiré patterns, halftone gridding, lines]
1.5 Illustration: tone, brightness, contrast
1.6 Illustration: color imbalance, gradient shifts

LEVEL 2: ENTIRE PAGE

2.1 Blur [movement]
2.2 Warp [text alignment, skew]
2.3 Crop [gutter, text block]
2.4 Obscured/cleaned [portions not visible]
2.5 Colorization [text bleed, low text to carrier contrast]
2.6 Full-text: patterns of errors at the page level (e.g., indicative of cropping errors in digitization processing)

LEVEL 3: WHOLE VOLUME

3.1 Order of pages [original source or scanning]
3.2 Missing pages [original source or scanning]
3.3 Duplicate pages [original source or scanning]
3.4 False pages [images not contained in source]
3.6 Full-text: patterns of errors at the volume level (e.g., indicative of OCR failure with non-Roman alphabets)

Table 2. Error incidence model for digitized book and serial volumes.

5.1. Use Case Scenarios

The aim of user-based validation is to confirm that the metrics we have chosen through statistical analysis and then assigned to use cases resonate with users who specify particular use scenarios for HathiTrust content. The development of use-cases is a method used in the design and deployment of software systems to help ensure that the software addresses explicit user needs. Within broad use-cases, individual users can construct stories or scenarios that articulate their requirements for digital content [1]. The research model utilizes use-case design methods to construct specific scenarios for four general purpose use-cases that together could satisfy the vast majority of uses:

Reading Online Images: A digitized volume is 'fit for use' when digital page-images are readable in an online, monitor-based environment. Text must be sufficiently legible to be intelligible [16][32]; visual content of illustrations and graphics are interpretable in the context of the text [23][5], where the envisioned use is legibility of text, interpretability of associated illustrations, and accurate reproduction of graphics sufficient to accomplish a task.

Reading Volumes Printed on Demand: This case refers to printing volumes (whole or substantial parts) derived from digital representations of original volumes upon request [21]. For a volume to be suitable for a print on demand service, it must be accurate, complete, and consistent at the volume level. A print copy is two steps removed from the original source, yet it serves as a ready reference version of the original.

Processing Full Text Data: Most expansively, this use-case specifies the suitability of the underlying full text data for computer-based analysis, summarization, or extraction of full-text textual data associated with any given volume [15]. For a volume to be acceptable for full-text processing, it must support one or more examples of data processing, including image processing and text extraction (OCR), linguistic analysis, automated translation, and other forms of Natural Language Processing [36], most typically applied in the digital humanities.

Managing Collections: This use-case encompasses collaboration among libraries to preserve print materials in a commonly managed space, as well as the management and preservation of the "last, best copy" of regionally determined imprints [33][37]. For digital surrogates to support collection management decision making, digitized volumes must have a sufficiently low frequency or severity of error that they can serve as replacement copies for physical volume.

5.2. Phase One – Metrics

Research Question 1: What is the most reliable system of measurement (metrics) for determining error in digitized book and serial volumes? As a point of departure, the research design hypothesizes a state of image and text quality in which digitized book and serial benchmark-volumes from a given vendor are sufficiently free of error such that these benchmark-surrogates can be used nearly universally within the context of specific use-case scenarios. In the first phase of the research project, we will explore how to specify the gap between benchmarked and digitized volumes in terms of detectable error. The outcome of the first-phase data gathering and analysis will be a highly reliable, statistically sound, and clearly defined error metrics protocol that can be applied in phase two to measure error-incidence in HathiTrust volumes. Addressing the first research question will require the research team to identify benchmark digitized volumes and create a data model for measuring the presence of error within a given digitized volume.

Identify Benchmark Volumes: The detection and recording of errors will be undertaken in reference to the very best examples of digitized volumes from a given vendor (e.g., Google), rather than in reference to an externally validated conversion standard. Benchmarks are volumes that have no errors that inhibit use in a given use-case. Such "bronze standards" will serve as the basis for developing training materials, establishing the point of departure for coding the severity of error, and validating quality baselines as part of the evaluation strategy.

Draw Samples: A programmer, with the guidance of a statistician, will draw multiple small random samples from selected strata of HathiTrust deposits by manipulating descriptive metadata for individual volumes (e.g., data of publication, LC classification, language). The purpose of sampling is to gather a representative group of volumes to test and refine the error definition model and determine the proper measurement scales for each error, rather than to make projections about error in a given strata population.

Code Errors: Staff and student assistants working in two research libraries [Michigan, Minnesota] will carry out whole-book manual review on the sample volumes, compiling the results initially in a spreadsheet designed by the graduate student research associate. The distinctive data gathering goals are: (1) to determine mechanisms for establishing gradations of severity within a given error-attribute; (2) to establish the threshold of "zero-error" that serves as a foundation for establishing the frequency of error on a given volume-page; and (3) confirm the estimates of error-frequency that determine specifications for the error review system.

Refine Error Data Model: The fundamental units of data in the research design are recorded frequency (counts) and severity (on an ordinal scale) of human-detectable error in either image or full-text data at the page level. The overall data model allows for errors related to image, full-text and illustrations within single pages (e.g., broken text, OCR errors, scanner effects on illustrations), or digitization errors that effect the readability of page images or associated full-text (e.g., blur, excessive cropping), and errors that are counted in pages but applied to entire volumes (e.g., missing or duplicate pages).

Determine Error Co-Occurrence: The research project will test the validity of each error measure in terms of the extent of co-occurrence of pairs of errors. Two measures are completely independent if the two errors never occur together on the same page, whereas two measures are totally dependent if the two errors always occur on the same page. For errors that occur with reasonable frequency, we will test the null hypothesis that error types are independent of each other using a 2 x 2 contingency table and Fisher's exact test for independence. This test for significance is used when the chi square expected frequencies are small. The measure of co-occurrence is a valid way to identify discrete error measures and, possibly, to reduce the number of error measures required to derive an overall measure of quality for a given volume.

5.3. Phase Two – Measurement

Research Question 2: What are the most accurate and efficient measures of error in HathiTrust content, relative to benchmarked digitized volumes? To examine the second question, results based on data analysis for Research Question 1 will be used to create and test measurement strategies for gathering error data from multiple diverse samples of volumes deposited in HathiTrust. Detection of error in digitized content is accomplished through the manual inspection of digital files and sometimes through comparison of digitized volumes with their original sources. The net results of the second phase of the project will be measures of error, aggregated to the volume level, that have as high of a level of statistical confidence as is possible to obtain through manual review procedures. Additionally, the outcome in phase two will be reliable estimates of the distribution of error in the population strata related to the analyzed samples.

Establish Sampling Strategies: The research project will design and implement procedures to draw random samples of volumes for manual inspection and to establish systematic page sampling specifications for review inside any given volume. Data analysis is designed to identify (1) the smallest sample size that can be drawn and analyzed to produce statistically meaningful results; (2) when is it most appropriate to utilize whole-book error analysis as opposed to examining an appropriately sized and identifiable sub-set of page images for a given book; and (3) when is it necessary and appropriate to examine errors in original source volumes as opposed to limiting analysis to digital surrogates. The size and number of volumes and samples depends upon the desired confidence interval (95%) and estimates of the proportion of error within the overall population. Based on three years of error assessment at Michigan, we expect the incidence of any given error to be well below 3%. Given this low probability of error, but where such error may indeed be catastrophic for use, the initial sampling strategy will utilize the medical clinician's "Rule of Three" [22], which specifies that 100 volumes or 100 pages sampled systematically in a typical volume will be sufficient to detect errors with an expected frequency < .03. Larger sample sizes are required for lower estimates of error.

Gather Data from Multiple Samples: Project staff will create, disseminate, and explain training materials to students and staff coders. A coding manual will contain narrative and visual examples of each error in the protocol, along with detailed instructions for coding error in the quality review system. Trained coders in the two participating academic libraries at the universities of Michigan and Minnesota will record the frequency (error counts) and severity (ordinal scale) of error in images and full-text data at the page level, as appropriate. The sampling strategy (outlined above) will determine the coding and analysis procedures in the two libraries. The data gathering design specifies resources in two research libraries sufficient to review and code approximately 5,000 volumes in samples of 100, 200 or 300 volumes per series. Estimates of review productivity, derived from the planning project supported by the Mellon Foundation, call for one hour of analysis and coding per volume, which will generate approximately 40 data values for each page reviewed in each volume. Data from error assessment activities will be collected in a centralized database at Michigan and subjected to data validation, cleaning, and processing routines by the graduate student research associate.

Assess Extent of Inter-coder Consistency: The research will adapt analytical procedures designed to diagnose and address the challenge of detecting and adjusting for the fact that two human beings will see and record the same information inconsistently. The presence of significant levels of inter-coder inconsistency generates error in the statistical evaluation of the findings of quality review undertaken by multiple reviewers in a distributed review environment. One error review procedure will entail multiple reviewers coding the severity of errors in the same volumes. Collapsing severity to a two-point scale (severe/not) will allow the testing of the null hypothesis that the pairs of reviewers code error severity in the same way, using Cohen's Kappa statistic as a measure of agreement. Similar tests assessing the frequency of errors detected will utilize the Chi Square test of significance. The outcome of these analyses will support improved training of coders and establish the lower threshold of coding consistency in a distributed review environment.

Aggregate from Page to Volume and Evaluate Results: The level of detail in error data at the page level will permit statistically significant aggregation of findings from page to volume. Data gathered at the page level for frequency and severity will be aggregated to the volume level to create coordinate pairs that can be plotted for further analysis. Volume-level error aggregation is the foundation for establishing quality scores for digitized volumes based on the relative number and severity of errors across a mix of error attributes. Error aggregates from assembled from samples of volumes will allow reliable projections regarding the distribution of error in HathiTrust strata. Examples of possible strata subject to analysis include

date and place of publication, subject classification, and digitization vendor.

6. CONTRIBUTIONS AND IMPLICATIONS

The research design is a significant contribution to the science of information quality within the context of digital preservation repositories, because the design is grounded in the models and methods pioneered by information quality researchers. The research design and the subsequent research project are innovative in their approach to quality definition and measurement, building specific error metrics appropriate for books and journals digitized at a large-scale. The design is also methodologically advanced through its full integration of (1) tools and procedures for gathering data about quality errors in digitized collections, (2) the rigorous analysis of that data to improve confidence in the measures, and (3) statistically significant conclusions about the nature of error in a large scale repository. Quality review processes conducted across two libraries helps ensure that the research findings may be generalized and not simply refer to one library's digital content. The quality metrics that will be developed in the research project are broadly applicable to collections of digitized books and journals other than those deposited in HathiTrust.

New metrics for defining error in digitized books and journals and new, user validated methods for measuring the quality of deposited volumes could have an immediate impact on the scope of repository quality assessment activities and specific quality assurance routines. Measurements of the quality and usefulness of preserved digital objects will allow digital repository managers to evaluate the effectiveness of the digitization standards and processes employed in producing usable content, and provide guidance on ways to alter digital content to improve the user experience. It will also allow repositories to make decisions about preserving digitized content versus requiring re-digitization (where possible). The ability to perform reliable quality review of digital volumes will also pave the way for certification of volumes as useful for a variety of common purposes (reading, printing, data analysis, etc.). Certification of this kind will increase the impact that digitally preserved volumes have in the broader discussions surrounding the management of print collections, and the interplay between print and digital resources in delivering services to users.

7. ACKNOWEDGEMENTS

Planning support has been provided by the Andrew W. Mellon Foundation. The author thanks HathiTrust executive director John Wilkin and the staff of the University of Michigan Library for providing data and technical support. The research project design was developed collaboratively by a planning team consisting of Jeremy York and Emily Campbell (MLibrary), Nicole Calderone and Devan Donaldson (School of Information), Sarah Shreeves (University of Illinois), and Robin Dale (Lyrasis).

8. REFERENCES

[1] Alexander I. F. & Maiden N.A.M., eds. (2004). *Scenarios, Stories and Use Cases.* New York: John Wiley.

[2] Bailey, C. W. (2010). "Google Book Search Bibliography." Version 6: 4/12/2010. http://www.digital-scholarship.org/gbsb/gbsb.html

[3] Baird, H. (2004). "Difficult and Urgent Open Problems in Document Image Analysis for Libraries," *Proc. of First International Workshop on Document Image Analysis for Libraries* (DIAL '04), Palo Alto, CA, pp. 25-32.

[4] Bearman, D. (2006). "Jean-Noël Jeanneney's Critique of Google." *D-Lib Magazine* 12 (12). http://www.dlib.org/dlib/december06/bearman/12bearman.html

[5] Biggs, M. (2004). "What Characterizes Pictures and Text?" *Literary and Linguistic Computing* 19 (3): 265-272.

[6] Bovee, M., Srivastava, R. and Mak, B. (2003). "A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality." *International Journal of Intelligent Systems.* 18 (1): 51-74.

[7] Cockburn, A. (2000). *Writing Effective Use Cases.* Boston: Addison-Wesley.

[8] Cohen, D. (2010). "Is Google Good for History?" *Dan Cohen's Digital Humanities Blog*. Posting on 12 Jan. 2010. http://www.dancohen.org/2010/01/07/is-google-good-for-history/

[9] Cohen, S. (2009). "Google to reincarnate digital books as paperbacks." *Library Stuff.* Information Today, Inc., 17 September, 2009. http://www.librarystuff.net/2009/09/17/google-to-reincarnate-digital-books-as-paperbacks/

[10] Conway, P. (2008). "Modeling the Digital Content Landscape in Universities." *Library Hi Tech* 26 (3): 342-358.

[11] Conway, P. (2009). "The Image and the Expert User." *Proceedings of IS&T's Archiving 2009*, Imaging Science & Technology, Arlington, VA, May 4-7, pp. 142-50.

[12] Conway, P. (2010). "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly* 80 (1): 61-79.

[13] Courant, P. (2006). "Scholarship and Academic Libraries (and their kin) in the World of Google." First Monday 11 (August). http://131.193.153.231/www/issues/issue11_8/courant/index.html

[14] Darnton, R. (2009). "Google and the New Digital Future," *The New York Review of Books* 56 (20): http://www.nybooks.com/articles/23518

[15] DeRose, S. et al. (1990). "What is Text, Really?" *Journal of Computing in Higher Education* 1 (2): 3-26.

[16] Dillon, A. (1992). "Reading from paper versus screens: A critical review of the empirical literature." E*rgonomics*, 35(10): 1297–1326.

[17] Doermann, D., Liang, J., and Li, H. (2003). "Progress in Camera-Based Document Image Analysis." *Proc. Seventh International Conference on Document Analysis and Recognition (ICDAR'03)*, 3 (6): 606-616.

[18] Duguid, P. (2007). "Inheritance and Loss? A Brief Survey of Google Books." *First Monday* 12 (8).

[19] Gartner, R. (2008). "Metadata for Digital Libraries: State of the Art and Future Directions." JISC *TechWatch Report* TSW0801. Bristol, UK: Joint Information Systems Committee.

[20] Garvin, D. A. (1988). *Managing quality: The strategic and competitive edge*. New York: Free Press.

[21] Hyatt, S. (2002). "Judging a book by its cover: e-books, digitization and print on demand." in Gorman, G.E. (ed.) *The Digital Factor in Library and Information Services*. London: Facet Publishing, 112-132.

[22] Jovanovic, B. D. & Levy, P. S. (1997). "A Look at the Rule of Three." *The American Statistician* 51 (2): 137-139.

[23] Kenney, A.R. et al. (1999). *Illustrated Book Study: Digital Conversion Requirements of Printed Illustrations.* (Report to the Library of Congress Preservation Directorate). Ithaca, N.Y.: Cornell University Library.

[24] Kenney, A.R. & Rieger, O.Y. (2000). *Moving Theory into Practice: Digital Imaging for Libraries and Archives.* Mountain View, CA: Research Libraries Group.

[25] Knight, S. (2008). *User Perceptions of Information Quality in World Wide Web Information Retrieval Behaviour*. (PhD Dissertation). Perth, Australia: Edith Cowan University.

[26] Lavoie, B. (2004**).** *The Open Archival Information System Reference Model: Introductory Guide.* Digital Preservation Coalition Technology Watch Report 04-01. Dublin, OH: OCLC.

[27] Le Bourgeois, et al. (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, pp. 2-24.

[28] 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.

[29] Lynch, C. A. (2003). "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." portal*: Libraries and the Academy* 3 (2): 327-336.

[30] Markey, K., Rieh, S. Y., St. Jean, B., Kim, J., and Yakel, E. (2007). *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub140abst.html

[31] Marshall, C. C. (2003). "Finding the Boundaries of the Library without Walls." In. Bishop, A., et al. (eds.) *Digital Library Use: Social Practice in Design and Evaluation.* Cambridge: MIT Press, pp. 43-64.

[32] O'Hara, K. (1996). "Towards a Typology of Reading Goals. Xerox Technical Report." http://www.xrce.xerox.com/content/download/6681/51479/file/ EPC-1996-107.pdf

[33] Payne, L. (2007). *Library Storage Facilities and the Future of Print Collections in North America.* Online Computer Library Center: Dublin, Ohio. www.oclc.org/programs/publications/reports/2007-01.pdf

[34] Rieger, O. (2008). *Preservation in the Age of Large-Scale Digitization: A White Paper.* Washington, DC: Council on Library and Information Resources.

[35] Rieh, S. (2002). "Judgment of Information Quality and Cognitive Authority in the Web." J*ournal of the American Society for Information Science and Technology* 53 (2): 145-161.

[36] Rockwell, G. (2003). "What it Text Analysis, Really?" *Literary and Linguistic Computing* 18 (2): 209-219.

[37] Schonfeld, R. and Housewright, R. (2009). *What to Withdraw? Print Collections Management in the Wake of Digitization*. New York: Ithaka.

[38] Stvilia, B., et al. (2007). "A Framework for Information Quality Assessment." *Journal of the American Society for Information Science and Technology* 58 (12): 1720-1733.

[39] Tanner, S., Munoz, T., and Ros, P. (2009). "Measuring Mass Text Digitization Quality and Usefulness." *D-Lib Magazine* 15 (July/August): 209. http://www.dlib.org/dlib/july09/munoz/07munoz.html

[40] TRAC. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* Center for Research Libraries and OCLC. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[41] Wang, R. and Strong, D. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems* 12 (4): 5-34.

[42] Waters, D. and Garrett, J. (eds.). (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Washington, DC: Commission on Preservation and Access.

[43] Yeo, G. (2008). "Concepts of Record (2): Prototypes and Boundary Objects." *American Archivist* 71 (Summer): 118-143.

[44] 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.