Arch Sci
DOI 10.1007/s10502-011-9155-0

ORIGINAL PAPER

Archival quality and long-term preservation: a research framework for validating the usefulness of digital surrogates

Paul Conway

Abstract Digital archives accept and preserve digital content for long-term use. Increasingly, stakeholders are creating large-scale digital repositories to ingest surrogates of archival resources or digitized books whose intellectual value as surrogates may exceed that of the original sources themselves. Although digital repository developers have expended significant effort to establish the trustworthiness of repository procedures and infrastructures, relatively little attention has been paid to the quality and usefulness of the preserved content itself. In situations where digital content has been created by third-party firms, content quality (or its absence in the form of unacceptable error) may directly influence repository trustworthiness. This article establishes a conceptual foundation for the association of archival quality and information quality research. It outlines a research project that is designed to develop and test measures of quality for digital content preserved in HathiTrust, a large-scale preservation repository. The research establishes methods of measuring error in digitized books at the data, page, and volume level and applies the measures to statistically valid samples of digitized books, adjusting for inter-coder inconsistencies and the effects of sampling strategies. The research findings are then validated with users who conform to one of four use-case scenarios: reading online, printing on demand, data mining, and print collection management. The paper concludes with comments on the implications of assessing archival quality within a digital preservation context.

Keywords Archival quality · Information quality · Large-scale digitization · Error measurement · Preservation repositories · HathiTrust

P. Conway (✉)
School of Information, University of Michigan, 4427 North Quad,
105 South State St., Ann Arbor, MI 48109, USA
e-mail: pconway@umich.edu

Published online: 21 September 2011

Introduction

For well over a decade, a worldwide cultural heritage community of libraries, archives, and museums has embraced the need for trustworthy digital archives that possess the technical capacity to acquire, manage, and deliver digital content persistently (Waters and Garrett 1996; Hedstrom and Ross 2003). This community has made significant progress toward establishing the terms and procedures for certifying trustworthiness at the repository level through independently administered auditing and risk assessment processes (OCLC 2007; McHugh et al. 2007). In the new environment of large-scale digitization and third-party content aggregation, however, repository certification may be insufficient to provide assurances to stakeholders and end-users about the quality of preserved content. For an institution and its community of users to trust that individual digital objects have archival integrity and to know that objects deposited in preservation repositories have the capacity to meet a variety of uses envisioned for them by different stakeholders, additional assurances may be needed. Archivists, digital curators, and digital repository managers must validate the quality and fitness for use of the objects they preserve and, in so doing, provide additional investment incentives for existing and new stakeholders.

Archival trust and archival quality are most closely associated through the preservation management of digital surrogates. One of the principal barriers to assessing the quality of digitized surrogates of archival records, published books, and other primary source materials is the general absence of viable mechanisms for defining and measuring quality factors in growing and complex digital preservation repositories and then validating these measures in the context of broadly applicable use-case scenarios. Until large-scale digitization by organizations such as Google, Microsoft, and the Internet Archive forced the issue of content quality to the forefront (Rieger 2008; Henry and Smith 2010), the preservation community exercised a form of vertical integration of digitization practice though the development and promulgation of best practices (Kenney and Rieger 2000) combined with a tendency to keep scanning activities close at hand and under curatorial control. Today's digital content environment is marked, increasingly, by distributed responsibility for digitization and collaborative responsibility for long-term preservation and access (Conway 2008). Preservation repositories for digitized content take what they can get, with, at best, assurances from the creator that the submitted content meets the original purposes or those deemed appropriate by the creator (Markey et al. 2007). Even where there is documentation on digitization processes or vendor quality assurance routines, questions *do* arise about the gap between the characteristics of a preservation repository's digital content and the expectations of users about that content's capabilities to satisfy information needs (Ackerman 2000).

This article presents the design of a new research effort to establish user-validated quality metrics for digital surrogates in a very large-scale digital preservation repository of digitized content and contextualizes this design within the literatures of archival and information quality. The concept of "quality" presents significant definitional challenges that have complicated efforts to establish and

validate quality metrics (Vullo et al. 2010). As the scale of digital repository building increases, so too does the need for measurement rigor that may lend itself to the development of computer-assisted quality assessment processes. The point of departure for the research described here is the nascent consensus within an international community of archival and digital preservation scholars that digital surrogacy can communicate archival properties through technological transformation (Ross 2007). Notwithstanding the commitments to supporting understandability by "designated communities" that are embedded in emerging digital repository standards (CCSDS 2002, pp. 1–10), advocates for digital preservation have not articulated the requirements for open communication with users about the qualities of the content in digital repositories. The research plan described here, then, is simultaneously an effort to advance the science of information quality measurement in a digital preservation context and an attempt to make more precise the relatively undeveloped concept of "archival quality."

The research project scoped in this article treats the content deposited in HathiTrust as a large-scale test bed for research on quality metrics and measurement processes. HathiTrust is a digital preservation repository launched in October 2008 by a large group of US 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 (March 2011), HathiTrust consists only of digitized content: 8.4 million digitized books and serial volumes ingested from multiple digitization sources, primarily Google's ongoing investment to digitize substantial portions of the bound collections housed in HathiTrust member research libraries. HathiTrust is an exemplar of a preservation repository containing digitized surrogates (1) with intellectual property rights owned by a variety of external entities, (2) created by multiple digitization vendors for access, and (3) deposited and preserved collaboratively. HathiTrust is also a technological environment for addressing the common challenges of collection development and digital preservation that all libraries and archives confront in an increasingly digital use environment (York 2010). The repository is in the mid of a rigorous certification audit by the Center for Research Libraries using the Trustworthy Repositories Audit and Certification framework (OCLC 2007). HathiTrust is supported by base funding from all 52 of its institutional partners, and its governing body includes top administrators from libraries and information officers at investing institutions (York 2009).

Archival quality and information quality in context

The longstanding but sparse literature on constructs of quality written by archivists barely intersects with a rich research base on information quality, yet the two literatures have much to contribute to the development of metrics for digitized archival content.

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

Archival quality

Over the seventy-year period since the words "archival quality" first appeared together in the archival literature, the term has been used as a simple metaphor for three complex but interrelated concepts: properties of records, characteristics of media, and the processes that preserve the essential nature of records when copied or transferred to another medium. Until the early 1960s, the term "quality" or "qualities" served as a synonym for "properties" as the archival profession struggled to define the distinctive character of archival thought and to distinguish archival practices from the well-established traditions of manuscript curation (Garrison 1939, p. 101). Archival quality could then be seen as synonymous with the concept of record or directly associated with the maintenance of provenance. For example, writing about the appraisal challenges of records centers four decades ago, Fishbein (1970, p. 184) noted the appraisal challenges that records centers faced in their "efforts to winnow the records of archival quality from the chaff of ephemerae."

With the widespread adoption of preservation microfilm in archives in the 1960s, "archival quality" assumed a second identity as a symbol of the physical characteristics that media with long life-expectancies should display, or the absence of such characteristics in other media, especially magnetic tape and computer storage systems containing digital data and electronic records (Poole 1977). New preservation-oriented uses of the term "archival quality" did not completely offset the original association with "recordness." For example, Taylor (1979, p. 425) mused on the archival nature of paintings while proposing that a range of artistic and visual works absorb archival qualities to the extent that they can be associated with "classically archival" records, perhaps as attachments, perhaps merely through physical juxtaposition.

A third use of the term emerged in the 1980s with the growth and distribution of formal preservation programmes in archives, libraries, and museums. "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). As the preservation community embraced "archival quality" as a metaphor for process, O'Toole (1989) urged archivists to abandon their affinity for "permanence" as a characteristic of physical records, asking them instead to focus on the archival quality of the information contained therein. At the virtual dawn of the digital reformatting era in archives, O'Toole implicitly granted permission to archivists to favor information over artifact. The intensity of the digital information environment today transforms the "artifact to information" shift into a de-emphasis on media and a renewed focus on the characteristics of digital content.

The adoption of "archival quality" to signify a subset of archival preservation processes served as a foundation for establishing acceptable digitization practices. Cultural heritage best practices for the digitization of printed books emerged from experimentation in the 1990s at the Library of Congress, Cornell, Yale, and other universities, eventually finding codification as guidance or explicit guidelines (Kenney and Rieger 2000). Experimentation produced technical specifications for a

"gold standard" for master access files (Puglia et al. 2004). Rigorous and well-documented technical standards for digitization, such as those under development by the US Federal Agencies Digitization Initiative, set a bar of quality for digitized books, defining what is and is not acceptable practice.[2] More important, the emergence of best practices that embed "archival quality" as a principal outcome of digitization processes reinforced in the cultural heritage community the possibility of and need for vertical control of the archival reformatting.

Since the turn of the 21st century, the association of "archival quality" with the fundamental nature of the archival record and the archival properties of digital collections has reasserted itself as a principle with increasing specificity. For example, in arguing for the special quality of American appraisal practices, Boles and Greene (1996, p. 304) invoke "archival quality" as a litmus test for establishing the acceptability of any theory that challenges established practices, as well as the long-standing assurances that archival programmes provide to users. For some archival theorists, the term "archival quality" remains synonymous with the existence of inviolable properties inherent in the archival record. In their "Quality Core Model" for interoperable digital libraries and repositories, Vullo et al. (2010) balance the need for precise and consistent policies with content that can be certified as having physical and intellectual integrity in the archival sense of those terms, as well as a documented provenance.

Thomassen (2001, p. 382) is especially lucid in identifying the relevance of "archival quality" to an emergent theory of archival science, namely "information itself and the processes that have generated and structured that information." Archival quality in digital repositories (Thomassen did not make explicit reference to such technologies) is the establishment and maintenance of "the optimal visibility and durability of the records, the generating work processes, and their mutual bond." For Thomassen, archival science is deeply associated with a philosophy of preservation that requires an explicit methodology. Thomassen associates archival quality with the processes that generate and structure archival information and ensure its availability, readability, completeness, relevance, representativeness, topicality, authenticity, and reliability. This list of the traits that comprise archival quality is a useful point of departure for linking archival constructs of quality with similar research-driven theories of information quality.

In defining archival quality as a characteristic of archives, Thomassen mirrors Duranti's classic assertion that reliability is the foundation of authority and trustworthiness of records as evidence. "Degree of completeness and degree of control of the procedure of creation are the only two factors that determine the reliability of records" (Duranti 1995, p. 6). Lauriault et al. (2007) document the tight association between archival quality and the trustworthiness that scientists ascribe to digital data repositories. It is clear that archival scholars are increasingly revisiting older notions of "archival quality" and adapting them to the digital environment, even as these same scholars remain largely unaware of complementary ideas emerging from scholarship on information quality.

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

Information quality

The quality of digital information has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. Models for information quality have emerged from important empirical research on data quality (Wang and Strong 1996) and have been adapted for the Internet context. Knight's dissertation (2008) is important for the way it validates 15 years of research on the dimensions of quality and ties those dimensions to the cyclical process of search and discovery that is at the heart of the scholarly communication process. In an important early exercise, Garvin (1988) categorizes five discrete approaches to understanding information quality: transcendent (timelessness); manufacturing-based (consumer preferences); value-based (cost and price); product-based; and user-based. Of these approaches, product-based and user-based categories are particularly relevant to the present research. The former approach views quality as a precise and measureable variable and lends itself to a hierarchical categorization of quality attributes. The latter approach allows for subjective judgments of quality based on a user's perception relative to need.

Research derived from business auditing principles (Bovee et al. 2003) and information science theory (Rieh 2002) grounds the analysis of information quality in the language of credibility and trust, which are the values that designated communities vest in digital preservation repositories. Wand and Wang (1996), among others, emphasize the importance of understanding the gap between the subjective mental models of quality attributes that users retain, which are driven by a perception of usefulness, and the statistically derived measures of quality errors made without regard to use. The source of such a gap could be experiential or reputational, including bad press, social network chatter, and/or scholarly communication (Kelton et al. 2008). Trust in digital repositories may indeed turn on the extent to which repositories can understand and act upon the perspectives of end-users as they interact with the preserved content.

Stvilia's important work (2007; Gasser and Stvilia 2001) builds on the commonality that exists in information quality models, focusing 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 et al. (2007) develop and test a general model of information quality assessment that factors in the processes of entity creation and entity management over time and space. The model establishes a three-part taxonomy of quality issues (intrinsic, representational, and reputational) and maps these issues to their origins in process activity (representation, decontextualizing, stabilizing, and provenance) and cultural/community norms. He then applies the general model to develop specific measurement schemes (metrics) in two cases, both of which highlight the challenges of establishing and maintaining high-quality metadata. Stvilia's model is an excellent combination of synthesis and specificity, making it possible to use the model as a guide to creating quality measurement processes in domains other than metadata.

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 (2006) provides a comprehensive review of the technical literature on quality assurance (QA) in the large-scale digitization pipeline (cataloguing, image capture, image analysis, and recognition). He focuses on how the processes of digital image analysis (DIA) are being addressed through research and presents a framework for understanding gaps in the research literature. Because Lin's framework is determined by ongoing DIA research problems, his "catalog of quality errors," adapted from Doermann et al. (2003), 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 (Le Bourgeois et al. 2004). Although the research design of the current project is oriented toward the possibility of eventual automated quality assurance, data gathering is based fundamentally on manual review of statistically valid samples of digitized volumes.

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 (Conway 2009). Tolerance for error may vary depending upon the expected uses for digitized books and journals. Marshall (2003, p. 54) argues that "the repository is far less useful when it's incomplete for whatever task the user has in mind." An "incomplete" digitized book could reflect scanning errors, blurred or unintelligible text or illustrations, and artifacts introduced by image processing routines on a scale unimagined 10 years ago. Baird (2004, p. 2) makes the essential connection between quality measurement and expected uses in articulating the need for research into "*goal directed metrics* (emphasis added) of document image quality, tied quantitatively to the reliability of downstream processing of the images." Certain fundamental, baseline capabilities of digital objects span disciplinary boundaries and can be predicted to be important to nearly all users (Crane and Friedlander 2008).

Although the emergent models under the broad umbrella of "information quality" are quite inconsistent in terminology, they provide a comparable theoretical foundation for research on quality in large-scale digitization.

Research model and methodology

The research design described here functions at the intersection of the relatively objective product-based findings on digitization quality and the more subjective evaluation judgments of a user-based approach. The design adopts Stvilia's (2007) general analytical model of information quality and drives deeply, with statistical

rigor, into the characteristics of digitized books and journals as rendered through the HathiTrust user interface. As a point of departure, the research design hypothesizes a state of image and text quality in which digitized benchmark-volumes from a given vendor are sufficiently free of error that these benchmark-surrogates can be used nearly universally within the context of specific use-case scenarios.

The research design and the ongoing research project adopt the notion of "validation" in two distinctive but complementary ways that expressly bridge Garvin's (1988) product-based and user-based approaches to digitization quality. First, drawing on the way computer scientists evaluate the performance of a system, validation is a set of data analysis routines that demonstrate the statistical power of the errors in a multi-variable error model. Second, validation is a set of processes whereby users who identify with a particular use-case scenario for HathiTrust content map the elements of the error model to specific HathiTrust content. Validation of the error model through user-based feedback provides a "reality check" that statistically determined findings on quality derived from samples of content properly describe the "fitness for use" of digitized volumes. In both uses of the term, validation is a mechanism for mitigating, but not entirely eliminating, the subjective nature of value judgments.

In the research design, archival quality is the absence of error relative to a given use scenario. A hierarchical model of error is at the heart of the investigation of archival quality in digitized surrogates. The error model is derived from 4 years of quality review data compiled by the University of Michigan Library (MLibrary). The research project's error incidence model, schematicized in Table 1, modifies the Michigan error model (bolded items) by adding reference to possible errors with book illustrations, OCR full-text errors from optical character recognition (OCR) routines, and errors that apply to an entire volume. The new error model also clarifies the intellectual framework of the Michigan model by clustering intrinsic quality error at three levels of abstraction: (1) data/information; (2) page image; and (3) whole volume as a unit of analysis. Within each level of abstraction exists 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 (e.g., broken text, OCR errors, and scanner effects) 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 (e.g., blur, excessive cropping). At the whole-volume level, a volume should be free of errors that affect the representation of the digital volume as a surrogate of a book (e.g., missing, false, or duplicate pages).

The error measurement model recognizes that perceived errors in the rendering of a digitized volume originate from some combination of problems with (a) the source volume (original book), (b) digital conversion processes (scanning and OCR conversion), or (c) post-scan enhancement processing. The history of the printed book is rife with descriptions of how production processes may introduce variety in the published product. Through its physical life cycle of handling and use, the physical integrity of a given volume may be compromised in any number of ways, including lost pages, rebinding that obscures text, and degradation of the paper that

Table 1 Error model for digitized books

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)

adversely affects readability. The digital scanning process itself captures the physical peculiarities of the source volume and then may introduce other artifacts that compromise the intellectual integrity of the volume. Finally, post-scan image enhancement, undertaken on batches of digitized volumes, provides opportunities for image and text corruption. Together, these three sources of quality errors aggregate and potentially co-relate to render a digital representation that may be significantly less useful than users desire, need, or expect to find.

Use-case scenarios

The aim of user-based validation is to confirm that the quality metrics identified through statistical analysis and then assigned to use-cases resonate with users who specify particular scenarios for using HathiTrust content. Use-cases articulate what stakeholders and users might accomplish if digital content was validated as capable of service-oriented functions (Carroll 2000). 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 (Alexander and Maiden 2004). 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 (Dillon 1992; O'Hara 1996); visual content of illustrations and graphics are interpretable in the context of the text (Kenney et al. 1999; Biggs 2004), 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) upon request from digital representations of original volumes (Hyatt 2002). For a volume to be fit 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. The conditions under which users accept a printed copy as a viable surrogate of the original source are important and as yet unexplored issue in large-scale digitization.
- *Processing Full-Text Data:* Most expansively, this use-case specifies the capability of the underlying full-text data to support computer-based analysis, summarization, or extraction of full-text textual data associated with any given volume (DeRose et al. 1990; Tanner et al. 2009). For a volume to be fit 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 (Rockwell 2003), 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 (Kisling et al. 1999; Payne 2007; Schonfeld and Housewright 2009). Digital surrogates may be fit for use in supporting collection management decision making if they have a sufficiently low frequency or severity of error at the whole-volume level (e.g., missing or duplicate pages, systematic scanning errors), such that they can serve as replacement copies for physical volume.

Establishing archival quality metrics

The 2-year research project (2011–2012) consists of two overlapping investigative phases designed first to specify and test the model of representational error and then apply the model to samples of digitized volumes deposited in HathiTrust. The research will establish processes and test procedures for gathering and analyzing data on the frequency and severity of errors in samples of digitized volumes, present the results to clusters of users who conform to a mix of use-case scenarios, and

adjust the error model based on the perceived significance of various elements of the model.

The first phase of the research project will explore how to specify the gap between benchmarked and digitized volumes in terms of detectable error. 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 one or more use-case scenarios. 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 coding ranges for severity of error, and validating quality baselines as part of the evaluation strategy.

The project team 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, and 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. Staff and student assistants working in two research libraries (University of Michigan, University of Minnesota) will carry out whole-book manual review on the sample volumes. 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 four distinctive data gathering goals are the following: (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; (3) confirm the estimates of error-frequency that determine specifications for statistical analysis of the error data set; and (4) test the validity of each error measure in terms of the extent of co-occurrence of pairs of errors. The outcome of the first-phase data gathering and analysis is a highly reliable, statistically sound, and clearly defined error metrics protocol.

Measuring and evaluating archival quality

In the second phase of the research project, the benchmarking tests of the error model established in phase one serve as the basis of measurement strategies for gathering error data from multiple diverse samples of volumes deposited in HathiTrust. The research goal of this phase is to identify the most accurate and efficient measures of error in HathiTrust content, relative to benchmarked digitized volumes. Detection of error in digitized content will be accomplished through the manual inspection of digital files and sometimes through comparison of digitized volumes with their original sources. The project design calls for the manual review

and error coding of approximately 5,000 volumes in samples that range from 100 to 1,000 volumes per series. The review will generate approximately 40 data values for each page in each volume. Data from error assessment activities will be collected in a centralized database at Michigan, aggregated automatically, and then subjected to data validation, cleaning, and processing routines.

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 subset 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 number of volumes in a given sample and the number of samples to be analyzed depend upon the desired confidence interval (95%) and estimates of the proportion of error within the overall population. Based on 3 years of error assessment at Michigan, we expect the incidence of any given error to be well below three percent. 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" (Jovanovic and Levy 1997), 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.

Coders trained in participating academic libraries at the universities of Michigan and Minnesota will record the frequency and severity of error in sample images and full-text data at the page level. In one component of the research protocol, coders will utilize double-blind data entry procedures to allow statistical analysis to detect and adjust for the fact that two human beings may see and record the same information inconsistently. The level of detail in error data at the page level will permit statistically significant aggregation of findings from page to volume. 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. The net result of the second phase of the project is measures of error, aggregated to the volume level, that have as high a level of statistical confidence as is possible to obtain through manual review procedures. An additional outcome from phase two will be reliable estimates of the distribution of error in the population strata related to the analyzed samples.

Given the extraordinary attention that large-scale digitization has garnered in the scholarly, professional, and popular press (Bailey 2009) and in the "blogosphere," the core of the research project's validation strategy encompasses exposing the findings to engagement with and feedback from anyone with an interest in the specific research or large-scale digitization in general. The project will integrate and make available two open-source social software tools, MediaWiki and WordPress, to support the development of use-case scenarios linked to the findings on digitization error and validate the concept of internal benchmarking. The interactive project site will provide detailed descriptions of the four use-cases and will solicit,

compile, and synthesize scenarios from end-users for each case that include explicit descriptions of quality requirements and quality limitations. For example, librarians may specify an explicit scenario for the transfer of volumes to off-campus shelving based on quality assumptions in digitized volumes. The project site will make available for review examples of the benchmark-volumes chosen by the project team to serve as "bronze standards" for zero-error in digitization, OCR processing, and post-scan processing activities. The interactive project site will display preliminary research findings on error measurement and co-occurrence of errors and will elicit and analyze user input on the relative importance of individual errors for particular use-cases. For example, users may rank a two percent incidence of severe page blur as a more significant problem for print-on-demand uses than for reading online. The project site will display in a readily interpretable form the statistical findings regarding the distribution of error in HathiTrust strata, along with visual examples of volumes that conform to the research project's error definition model.

Implications

The research design makes a 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. One might be tempted to question the grounding of research on an archival construct such as "archival quality" in the products of large-scale book digitization. And yet, as more and more historical source materials shift *en masse* to online access, out-of-print and sometimes quite hard to locate books take their place alongside photographs and other archival records as primary sources for end-users. Taylor (1987, p. 27), in quoting Kaplan (1964, p. 297) is explicit about the need to recognize books as an important part of the historical cultural record. 'A library… is first of all an archive or repository in which society can find what it has already learned'. This is written by a librarian with, at first sight, rather a curious use of the term 'archive', yet a library might be considered as a printed 'archives' of countless authors recounting what they have learned, because books are 'about' primary materials." Over time, digital repositories will preserve a wider and wider variety of digitized content drawn from archives and special collections.

The findings of the research described here will be broadly applicable to the current digital repository environment, ranging from smaller and somewhat stable repositories to large scale evolving digital preservation services. The research outlined here will establish new metrics for defining error in digitized books and journals and new, user-validated methods for measuring the quality of deposited volumes. The findings 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 archivists and curators to evaluate the effectiveness of the digitization standards and processes

they employ to produce usable content. Research findings may also help digital repository managers 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.

If Thomassen is correct that archival science can in large measure be defined by a methodology aimed in particular at establishing and maintaining the archival quality of process-bound information (2001, p. 382), then such a methodology must be developed and tested, not simply asserted as a first principle. Research on large-scale cultural heritage digitization maps the re-emergent theoretical construct of "archival quality" to an explicit, replicable methodology derived from distinctive work in the community of information quality research. The philosophical underpinnings of this research extend beyond the objective mechanics of statistical analysis to encompass how end-users judge the validity of statistically derived truths. In this way, the research design creates a bridge between the content-oriented mandates of digital preservation and the user-oriented expectations of access systems. In reformulating access as an archival paradigm, Menne-Haritz (2001, p. 81) explicitly makes the same claim. She grounds access to archives in the expectations and needs of users, not simply on the processes of preparing archival records for use. "The completeness of knowledge depends on the demand and not on the input of sources and it can only be measured by the needs and not by the scope of sources prepared in advance. So, the responsibility for the quality and the completeness of knowledge is attributed to the user and not to a provider." The development of an archival theory, thus, depends upon continuing the movement away from terminological absolutes while reinforcing the need for methodological rigor.

Beyond the development of a reliable model for measuring quality in large-scale digital archives, the research serves as a mechanism for continuing a conversation on the nature of archival quality as applied to large collections of surrogates of archival holdings. Because the issue of archival quality has tended over the years to devolve into a debate between media permanence, on the one hand, and the retention of archival properties through the information life cycle, on the other, the archival nature of digital surrogates has received short shrift. As users depend (and expect, require, or demand) direct online access to archival holdings, archivists must calibrate closely the characteristics of the digital content for which they assume responsibility in persistent repositories.

Acknowledgments The Andrew W. Mellon Foundation provided support for planning and project development. The U.S. Institute for Museum and Library Studies [LG-06-10-0144-10] is supporting the implementation of the research design. 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 was designed collaboratively by a planning team consisting of Jeremy York and Emily Campbell (MLibrary), Nicole Calderone and Devan Donaldson (University of Michigan School of Information), Sarah Shreeves (University of Illinois), and Robin Dale (Lyrasis).

References

Ackerman M (2000) The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. Human Comput Interact 15:179–203

Alexander IF, Maiden NAM (eds) (2004) Scenarios, stories and use cases. Wiley, New York

Bailey CW (2009) Google book search bibliography. Version 6 (4 Dec 2010). http://www.digital-scholarship.org/gbsb/gbsb.htm. Accessed 14 March 2011

Baird H (2004) Difficult and urgent open problems in document image analysis for libraries. In: Proceedings of first international workshop on document image analysis for libraries (DIAL' 04), Palo Alto, CA, pp 25–32

Biggs M (2004) What characterizes pictures and text? Lit Linguist Comput 19(3):265–272

Boles F, Greene MA (1996) Et tu Schellenberg? Thoughts on the dagger of American appraisal theory. Am Arch 59(3):298–310

Bovee M, Srivastava R, Mak B (2003) A conceptual framework and belief-function approach to assessing overall information quality. Int J Intell Syst 18(1):51–74

Carroll J (2000) Making use: scenario-based design of human-computer interactions. MIT Press, Cambridge

Consultative Committee for Space Data Systems (2002) Reference model for an open archival information system (OAIS) recommendation for space data system standards; blue book. CCSDS Secretariat, Washington. http://public.ccsds.org/publications/archive/650x0b1.pdf. Accessed 11 March 2011

Conway P (1989) Archival preservation: definitions for improving education and training. Restaurator 10(1):47–60

Conway P (2008) Modelling the digital content landscape in universities. Library Hi Tech 26(3):342–358

Conway P (2009) The image and the expert user. In: Proceedings of IS&T's archiving 2009, imaging science and technology, Arlington, VA, 4–7 May, pp 142–50

Crane G, Friedlander A (2008) Many more than a million: building the digital environment for the age of abundance. Report of a one-day seminar on promoting digital scholarship, 28 Nov 2007. Council on library and information resources, Washington. http://www.clir.org/activities/digitalscholar/Nov28final.pdf. Accessed 21 March 2011

DeRose S et al (1990) What is text, really? J Comput High Edu 1(2):3–26

Dillon A (1992) Reading from paper versus screens: a critical review of the empirical literature. Ergonomics 35(10):1297–1326

Doermann D, Liang J, Li H (2003) Progress in camera-based document image analysis. In: Proceedings seventh international conference on document analysis and recognition (ICDAR' 03), vol 3(6), pp 606–616

Duranti L (1995) Reliability and authenticity: the concepts and their implications. Archivaria 39:5–10

Fishbein MH (1970) A viewpoint on appraisal of national records. Am Arch 33(2):175–187

Garrison C (1939) The relation of historical manuscripts to archival materials. Am Arch 2(2):97–105

Garvin DA (1988) Managing quality: the strategic and competitive edge. Free Press, New York

Gasser L, Stvilia B (2001) A new framework for information quality (ISRN UIUCLIS-2001/1 + AMAS). Univ of Illinois, Champaign

Hedstrom M, Ross S (2003) Invest to save: report and recommendations of the NSF-DELOS working group on digital archiving and preservation. Prepared for national science foundation's (NSF) digital library initiative and the European Union under the fifth framework programme by the network of excellence for digital libraries (DELOS)

Henry C, Smith K (2010) Ghostlier demarcations: large-scale text digitization projects and their utility for contemporary humanities scholarship. In the idea of order: transforming research collections for 21st century scholarship, pp 106–115. Council on Library and Information Resources, Washington. (See also the supplemental online report and data by A. Gevinson (2010) Results of an examination of

200 digitizations [sic] of books in the field of American intellectual history: summary, results, data. http://www.clir.org/pubs/abstract/pub147abst.html. Accessed 21 March 2011)
McHugh A, Ruusalepp R, Ross S, Hofman, H (2007) Digital repository audit method based on risk assessment (DRAMBORA). Digital curation centre (DCC) and digital preservation Europe (DPE). http://www.repositoryaudit.eu/. Accessed 21 March 2011
Hyatt S (2002) Judging a book by its cover: e-books, digitization and print on demand. In: Gorman GE (ed) The digital factor in library and information services. Facet Publishing, London, pp 112–132
Jovanovic BD, Levy PS (1997) A look at the rule of three. Am Stat 51(2):137–139
Kaplan A (1964) The age of symbol: a philosophy of library education. Libr Quart 34(4):295–304
Kelton K et al (2008) Trust in digital information. J Am Soc Inf Sci Technol 59(3):363–374
Kenney AR, Rieger OY (2000) Moving theory into practice: digital imaging for libraries and archives. Research Libraries Group, Mountain View
Kenney AR et al (1999) Illustrated book study: digital conversion requirements of printed illustrations. Cornell University Library, Ithaca
Kisling V Jr, Haas S, Cenzer P (1999) Last copy depository: cooperative collection management centres in the electronic age. Collect Manag 24(1):87–92
Knight S (2008) User perceptions of information quality in World Wide Web information retrieval behaviour. PhD Dissertation, Edith Cowan University, Perth, Australia
Lauriault T, Craig B, Fraser Taylor DR, Pulsifer P (2007) Today's data are part of tomorrow's research: archival issues in the sciences. Archivaria 64:123–179
Le Bourgeois F, Trinh E et al (2004) Document images analysis solutions for digital libraries. In: Proceedings of the first international workshop on document image analysis for libraries (DIAL' 04), Palo Alto, California, pp 2–24
Lin X (2006) Quality assurance in high volume document digitization: a survey. In: Proceedings of the second international conference on document image analysis for libraries (DIAL'06), 27–28 April, Lyon, France, pp 319–326
Markey K, Rieh SY, St. Jean B, Kim J, Yakel E (2007) Census of institutional repositories in the united states: MIRACLE project research findings. Washington Council on Library and Information Resources, Washington. http://www.clir.org/pubs/abstract/pub140abst.html. Accessed 21 March 2011
Marshall CC (2003) Finding the boundaries of the library without walls. In: Bishop A et al (eds) Digital library use: social practice in design and evaluation. MIT Press, Cambridge, pp 43–64
Menne-Haritz A (2001) Access—the reformulation of an archival paradigm. Arch Sci 1(1):57–82
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. Accessed 21 March 2011
O'Toole JM (1989) On the idea of permanence. Am Arch 52(1):10–25
OCLC-National Archives and Records Administration (2007) Trustworthy repositories audit and certification: criteria and checklist (TRAC) Ver 1.0. Center for Research Libraries, Chicago. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf. Accessed 21 March 2011
Payne L (2007) Library storage facilities and the future of print collections in North America. Online Computer Library Center, Dublin. http://www.oclc.org/programs/publications/reports/2007-01.pdf Accessed 21 March 2011
Poole FG (1977) Some aspects of the conservation problem in archives. Am Arch 40(2):163–171
Puglia S, Reed JA, Rhodes E (2004) Technical guidelines for digitizing archival materials for electronic access: creation of production master files–raster images. Digital Library Federation, Washington
Rieger O (2008) Preservation in the age of large-scale digitization: a white paper. Council on Library and Information Resources, Washington
Rieh S (2002) Judgment of information quality and cognitive authority in the web. J Am Soc Inf Sci Technol 53(2):145–161
Rockwell G (2003) What is text analysis, really? Liter Linguist Comput 18(2):209–219
Ross S (2007) Digital preservation, archival science and methodological foundations for digital libraries. Keynote address at the 11th European conference on digital libraries (ECDL), Budapest, 17 Sept 2007
Schonfeld R, Housewright R (2009) What to withdraw? Print collections management in the wake of digitization. Ithaka, New York
Stvilia B et al (2007) A framework for information quality assessment. J Am Soc Inf Sci Technol 58(12):1720–1733

Tanner S, Munoz T, Ros P (2009) Measuring mass text digitization quality and usefulness. D-Lib Magazine, 15 July/August 2009. http://www.dlib.org/dlib/july09/munoz/07munoz.html. Accessed 21 March 2011

Taylor HA (1979) Documentary art and the role of the archivist. Am Arch 42(4):417–428

Taylor HA (1987) Transformation in the archives: technological adjustment or paradigm shift? Archivaria 25:12–28

Thomassen T (2001) A first introduction to archival science. Arch Sci 1:373–385

Vullo G, Innocenti P, Ross S (2010) Towards policy and quality interoperability: Challenges and approaches for digital libraries. In: Conference proceedings IS&T archiving, 1–4 June 2010. Society for Imaging Science and Technology, The Hague, pp 33–38

Wand Y, Wang R (1996) Anchoring data quality dimensions in ontological foundations. Commun ACM 39(11):86–95

Wang R, Strong D (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12(4):5–34

Waters D, Garrett J (eds) (1996) Preserving digital information: report of the task force on archiving of digital information. Commission on Preservation and Access, Washington

York JJ (2009) This library never forgets: preservation, cooperation, and the making of HathiTrust digital library. Proc. IS&T Archiving 2009, Arlington, pp 5–10

York JJ (2010) Building a future by preserving our past: the preservation infrastructure of HathiTrust digital library. 76th IFLA general congress and assembly, 10–15 August, Gothenberg, Sweden. http://www.ifla.org/files/hq/papers/ifla76/157-york-en.pdf. Accessed 21 March 2011

Author Biography

Paul Conway is associate professor in the School of Information at the University of Michigan. He holds a PhD from the University of Michigan. His research encompasses the digitization of cultural heritage resources, particularly photographic archives, the use of digitized resources by experts in a variety of humanities contexts, and the measurement of image and text quality in large-scale digitization programs. He has extensive research, teaching, and administrative experience in archives and preservation fields and has made major contributions over the past 30 years to the literature on archival users and use, preservation management, and digital imaging technologies. He has held positions at the National Archives and Records Administration (1977–1987; 1989–1992), the Society of American Archivists (1988–1989), Yale University (1992–2001), and Duke University (2001–2006). In 2005, Conway received the American Library Association's Paul Banks and Carolyn Harris Preservation Award for his contributions to the preservation field. He is a Fellow of the Society of American Archivists.