Error Metrics for Large-Scale Digitization

Paul Conway
University of Michigan
4427 North Quad
105 S. State Street
Ann Arbor, MI 48109
+1 734 615-1419
pconway@umich.edu

Jacqueline Bronicki
University of Michigan
320 Hatcher North
Ann Arbor, MI 48109
+1 734 764-8742
bronick@umich.edu

ABSTRACT

The paper summarizes the methodology utilized in an ongoing project that is exploring quality issues in the large-scale digitization of books by third-party vendors – such as Google and the Internet Archive – that are preserved in the HathiTrust Digital Library. The paper describes the research foundation for the project and the model of digitization error that frames the data gathering effort. The heart of the paper is an overview of the metrics and methodologies developed in the project to apply the error model to statistically valid random samples of digital book-surrogates that represent the full range of source volumes digitized by Google and other third party vendors. Proportional and systematic sampling of page-images within each 1,000-volume sample produced a study set of 356,217 page images. Using custom-built web-enabled database systems, teams of trained coders have recorded perceived error in page-images on a severity scale of 0-5 for up to eleven possible errors. The paper concludes with a summary of ongoing research and the potential for future research derived from the present effort.

Categories and Subject Descriptors

I.4.1 [**Image Processing & Computer Vision**]: Digitization and Image Capture

General Terms

Measurement, Verification

Keywords

digitization quality; error model; Google Books; HathiTrust

1. INTRODUCTION

From Project Gutenberg to Google Books, the large-scale digitization of books and serials is generating extraordinary collections of intellectual content that are transforming the way society reads and learns. Questions are being raised, however, regarding the quality and usefulness of digital surrogates produced by third-party vendors and deposited in digital repositories for preservation and access. For such repositories and their communities of users to trust digital documents, repositories must validate the quality of these objects and their fitness for the uses envisioned for them. Information quality should be an important component of the value proposition that digital preservation repositories offer their stakeholders and users. [4]

The quality of digital information has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. The literature on information quality, however, is relatively silent on how to measure quality attributes of very large collections of digitized books and journals, created as a combination of page images and full-text data by third party vendors. Lin [10] provides an excellent review of the state of digital image analysis (DIA) research within the context of large-scale book digitization projects and establishes a "catalog of quality errors," adapted from Doermann. [8] His research 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, even though quality assurance today depends in large measure upon manual visual inspection of digitized surrogates or the original book volumes. [9]

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 argues that "the repository is far less useful when it's incomplete for whatever task the user has in mind." [11, p. 54] 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." [2, p. 2] Certain fundamental, baseline capabilities of digital objects span disciplinary boundaries and can be predicted to be important to nearly all users. [7] Use-cases articulate what stakeholders and users might accomplish if digital content was validated as capable of service-oriented functions. [6] Individual users construct scenarios that articulate their requirements for digital content. [1]

For this research project, we define quality as the absence of errors in scanning and post-scan processing relative to expected uses. [5] Within the context of a large-scale preservation repository, the research adapts Stvilia's [12] model of intrinsic quality attributes and Lin's [10] framework of errors in book surrogates derived from digitization and post-scan processing. The overall design of the three-year research project consists of three overlapping investigative phases. Phase one defines and tests a set of error metrics (a system of measurement) for digitized books and journals. Phase two applies those metrics to produce a set of statistically valid measures regarding the patterns of error (frequency and severity) in multiple samples of volumes drawn from strata of HathiTrust. Phase three (ongoing) will engage stakeholders and users in building, refining, and validating the use-case scenarios that emerge from the research findings.

The research project utilizes content deposited in the HathiTrust Digital Library, which is a digital preservation repository launched in October 2008 by a group of research universities, including the Committee on Institutional Cooperation [the Big Ten universities and the University of Chicago] and the

University of California system. At present [August 2012] HathiTrust consists of 10.4 million digitized volumes ingested from multiple digitization sources (primarily Google). HathiTrust is supported by base funding from its 66 institutional partners, and its governing body includes top administrators from libraries and information offices at investing institutions. [13][14] 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. The findings of the research are broadly applicable to the challenges in duplication, collection development, and digital preservation that are common to all digital libraries.

2. ERROR MODEL

A three-tiered hierarchical error typology and associated value definitions are the keystones of the study. The error model **(Figure 1)** identifies error at the data, page, and volume levels and establishes hypotheses regarding the cause of each error (source, scanning, post-scan manipulation). Data and page-image errors are individually identifiable errors that affect the visual appearance of single bitmap pages. A particular error may be confined to a single page or repeated across a sequence in a volume. Whole volume-level errors apply to structural issues surrounding the completeness or accuracy of the volume as a whole, such as missing pages, duplicate pages, and ordering of pages. The development process for the error model was deeply iterative and involved substantial testing of individual error items and the meaning of narrative error definitions. The goal was to create a validated error model with clearly defined errors that could be repeatedly and consistently identified by coding staff in multiple settings.

2.1 Sources of error

The error model implies causality regarding one of three factors: the physical qualities of the source volume, the cluster of scanning activities that create a master bitmap image of two pages in an open book, and the suite of post-scan manipulation processes that produce the final deliverable image that users consult. One of the primary objectives of the data collection process is to gather data on errors without assuming the cause of error. Coders were instructed to "code what you see" rather than speculate on the cause of error.

2.2 Severity of error

The research team developed a severity scale for each of the eleven page-image errors to capture a more granular rating of each error. In order to train coding staff to uniformly assign severity, the team outlined four main definitions for coders to reflect upon when assigning severity: original content, error, reading ability, and inference. *Original Content* is defined as the text or image content on the page created through the original printing process. Original content excludes marginalia, annotations, and other library-added content (bar codes, call numbers, book plates, circulation aids) added by users after the acquisition of the volume by the library. *Error* is defined as variations from the expected appearance of Original Content. *Reading ability* is designated as the ability of a reviewer to interpret the letters, illustrations, and other information contained in the Original Content of a page. *Inference* is the degree to which an average reviewer cannot detect Original Content, but must use contextual information to determine letters, words, or other information that compose the Original Content. Using this understanding, the coder is expected to apply a level of severity from zero to five for all errors detected on the page upon review. **Figure 2** displays the operative severity scale used by the 12 part-time coders working in teams at the University of Michigan and the University of Minnesota.

3. METRICS FOR DIGITIZATION ERROR

The research 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 development phase, the research explored how to specify the gap between benchmark and digitized volumes in terms of detectable error. The project developed a highly reliable and statistically sound data gathering and analysis system to measure error-incidence in HathiTrust volumes. The research team focused initially on sampled page-images within a digitized volume, followed by physical review of sampled volumes, and culminating with a whole volume review of the same sampled volumes. The scope of the project included review of 356,217 individually sampled pages from four distinctive samples, plus a second-stage review of entire volumes totaling 691,972 page-images.

3.1 Page-level data collection

A key component of our study is efficient coding of each digital page-image with an easy to use web application **(Figure 3).** The project built a highly efficient web-based application that could be used in multiple remote locations. The web application has a user interface that populates to a backend database with complex controls to minimize data entry error. The database records all coded values per sequence number relating to a unique volume, identified by a unique HathiTrust ID.

3.2 Physical book inspection

To supplement the data gathered on page-level and whole volume errors, the research team designed a process for inspecting physical volumes and correlating material and bibliographic characteristics with detected errors. A physical review of each sampled volume was conducted by current UMSI students. The physical review model was developed by the principal investigator based on prior standards and variables used by the preservation community to review physical volumes for damage and deterioration. The independent variables and their values were crafted into a brief online questionnaire and student volunteers were trained to identify and capture physical characteristics of the volume under the supervision of the principal investigator. The survey featured 11 questions regarding the quality of the book as a whole, 12 bibliographic data fields to be confirmed by the reviewer, and 4 metadata fields populated by the project programmer.

The project programmer created a stand-alone web-based interface designed with efficiency and mobility as the central features **(Figure 4).** The interface connects to a backend SQL database where a unique identifier could be used to map data gathered in physical inspection to page-level and whole volume error data. Reviewers were able to access the interface from various locations through a secure internet connection after they were authenticated by the system.

3.3 Whole volume error

The error model identifies five distinct whole volume error categories related to scanning and processing of digital volumes that relate to completeness and integrity of the volume. The five major binary error types are: missing page(s), duplicate page (s),

out of order page(s), false page(s), and fully obscured page(s). No severity level is assigned to whole volume as the condition either exists or it does not.

A secure web-based application **(Figure 5)** has been developed to capture error coding at the whole volume level. All coded errors are captured in a central database for statistical analysis. To control for HathiTrust interface effects, the application was designed to have a minimalist thumbnail view interface while maximizing data collection efficiency. Each data coder is authenticated using unique ID and login, thus allowing the detailed logging of coding activity. The coder has access to an entire volume as sequenced in HathiTrust along with relevant metadata to enhance the ability to code error. The coder inspects several parts of the digital image as well as aspects of vendor-supplied metadata to determine if an error exists: page number as seen in digital image, page number as provided by vendor in the metadata, context of the text from page to page, and context of the volume as a whole.

4. APPLYING THE METRICS

4.1 Representativeness (two tier sampling)

The purpose of sampling is to gather a representative group of volumes to test and refine the error definition model and to make projections about error in a given strata population. The issue of representativeness was addressed in the sampling techniques applied during data collection phase. Under direction from the team statistician, the programmer developed a systematic random sampling algorithm to pull random samples from the HathiTrust Library with pre-determined sample parameters. The project co-PI, who is a distinguished scholar of statistical process control, determined that 1,000 volumes would be representative of sampling pools within HathiTrust and would allow for statistical comparison of sub-populations with small frequencies in important variables.

Within each 1,000 volume sample, the project team extracted a systematic random sample of 100 pages within each volume to predict the distribution of error within the volume as a whole. The sampling algorithm is applied to the image sequence number, the complete set of which serves as a proxy for the total number of pages in a given volume, cover to cover. The algorithm divides the total number of images within a volume by one hundred to establish a number that determines the sequential sampling interval value. A random number generator establishes where in the volume (between sequence number 1 and 10) to begin sequential sampling. This method ensures that the sample will be representative of the images at the front and ends of the volumes. Sequential sampling then selects pages according to the sampling interval value, rounded up or down accordingly, to determine which whole-sequence-number image should be chosen.

4.2 Data reliability and tests of significance

The research adapts 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 entails multiple reviewers coding the severity of errors in the same volumes. Collapsing severity to a two-point scale (severe/not) allows for 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 utilize the Chi Square test of significance. The outcome of these analyses supports improved training of coders and establish the lower threshold of coding consistency in a distributed review environment.

4.3 Data gathered in the study

The project team established two data gathering teams, one group of four part time staff at the University of Minnesota and another group of between four and eight part time staff. The Project Manager developed training materials and a training routine to establish a consistent pattern of review behavior. **Table 1** displays the total number of volumes and pages reviewed by the combined coding teams and estimates the size of the populations represented by the random samples.

5. ONGOING RESEARCH

5.1 Cost of manual inspection

The Project Coordinator tracked very closely the expenditure of time and resources by paid coding staff. Additionally, the web-based review systems recorded the time spend by individual coders on page level and volume level review. This data will be processed to yield an assessment of the total cost of manual review processes as well as a comparison of the cost of the separate approaches to quality review (page-level versus whole volume).

5.2 Validating results from users

Ongoing research with two populations of users of digitized volumes seeks to validate the statistical findings with end-user needs and expectations. The two populations of study are digital humanities scholars (faculty and doctoral students), whose research requires close reading of published books; and library collection development staff who expect to use digitized volumes as replacements for or surrogates of physical volumes. The goal of the research is to identify needs-based thresholds of acceptance of detected error.

5.3 Potential for automatic error detection

Findings from page-level and volume level error will yield a prioritized list of scanning and post-scan procedures that result in error. Future research will explore the extent to which the most frequent and the most offending errors can be detected and corrected using image processing algorithms. Preliminary research has identified potentially valuable processing procedures for duplicate page images, and for warped or skewed page images. Fixing text anomalies might also be possible in certain cases. The challenges of correcting scanning artifacts in book illustrations are more problematical.

5.4 Tagging and rating error

A supplemental goal of the project is to address a priority need within the HathiTrust community of stakeholders: namely a tool for the efficient review of individual volumes on demand and the rating of these volumes in terms of the presence or absence of critically important errors. This work is ongoing and will become one of the principal deliverables of the grant project.

6. ACKNOWLEDGMENTS

Project planning supported by the Andrew W. Mellon Foundation; project implementation supported by a grant from the Institute for Museum and Library Services [LG-06-10-0144-10]. We wish to thank Ryan Rotter, Systems Programmer; Ken Guire, Statistician;

Jeremy York, Associate Librarian; and co-PI Edward Rothman, Professor of Statistics, for their indispensible work on the project.

7. REFERENCES

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

[2] Baird, H. (2004). "Difficult and Urgent Open Problems in Document Image Analysis for Libraries," *Proc. Of International Workshop on Document Image Analysis for Libraries*? (?): 25-32.

[3] 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.

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

[5] Conway, P. (2011). "Archival Quality and Long-term Preservation: A Research Framework for Validating the Usefulness of Digital Surrogates." *Archival Science* 11 (3): 293-309. [DOI: 10.1007/s10502-011-9155-0]

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

[7] Crane, G. and 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, November 28, 2007.* Washington, D.C.: Council on Library and Information Resources.

[8] 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.

[9] 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), 23-24 Jan., Palo Alto, California, pp. 2-24.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] York, J. J. (2010). Building a future by preserving our past: The preservation infrastructure of HathiTrust digital library." *76th IFLA General Congress and Assembly, 10-15 August*, Gothenberg, Sweden.

.

Level of Abstraction	*Possible Cause of Error*
LEVEL 1: DATA/INFORMATION	
1.1 Text: thick text [fill, excessive]	Source or post-processing
1.2 Text: broken text [character breakup]	Source or post-processing
1.3 Illustration: scanner effects [moiré patterns, gridding]	Scanning or post-processing
1.4 Illustration: tone, brightness, contrast	Scanning, post-processing, or source
1.5 Illustration: color imbalance, gradient shifts	Scanning, post-processing, or source
LEVEL 2: ENTIRE PAGE	
2.1 Blur [distortion]	Scanning or source
2.2 Warp [text alignment]	Post-processing
2.3 Skew [page alignment]	Scanning, post-processing, or source
2.4 Crop [gutter, text block]	Source or post-processing
2.5 Obscured [portions not visible]	Scanning or post-processing
2.6 Colorization [text bleed, low contrast]	Source or post-processing
LEVEL 3: WHOLE VOLUME	
3.1 Fully obscured [foldouts]	Scanning
3.2 Missing pages [one or more]	Original source or scanning
3.3 Duplicate pages [one or more]	Original source or scanning
3.4 Order of pages	Original source or scanning
3.5 False pages [not part of original content]	Scanning or post-processing

Figure 1. Model of error in large-scale digitization

0 - Default - Error is undetectable on the page.

1 - Error exists but has a negligible effect on the Original Content.

2 - Error clearly alters appearance of Original Content, but has a negligible effect on reading ability.

3 - Error clearly alters appearance of Original Content and has a clear negative impact on reading ability.

4 - Nearly unable to decipher Original Content in affected area of the page; significant inference required by reviewer to obtain legibility and meaning.

Figure 2. Severity scale

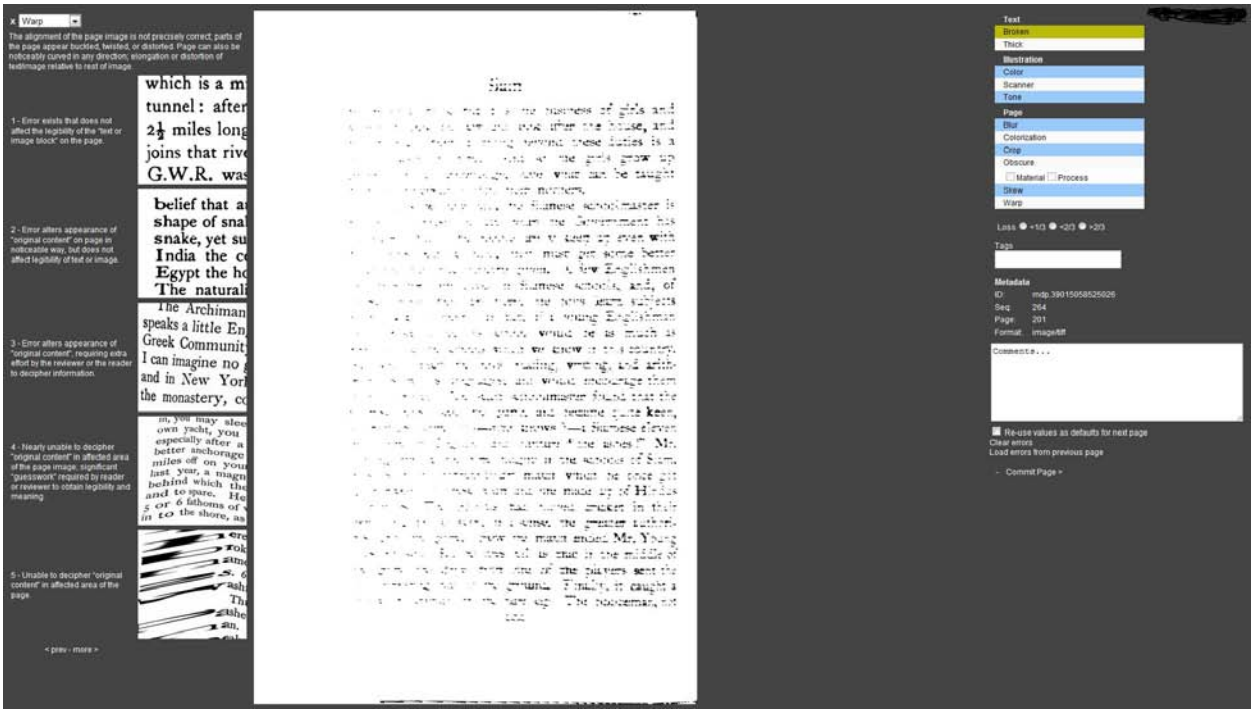

Figure 3. Interface for coding page-level error

Figure 4. Physical review interface (partial view)

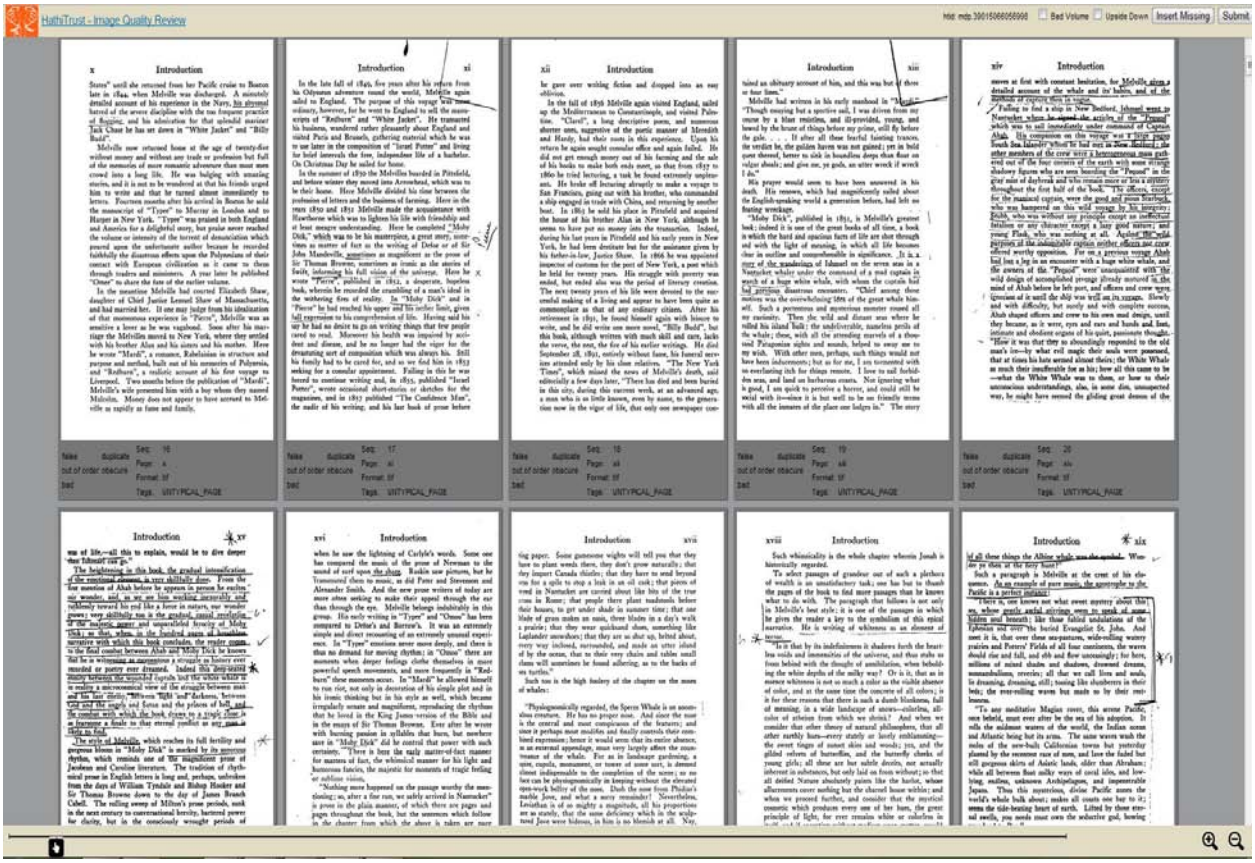

Figure 5. Interface for coding whole volume errors

Table 1. Summary of sample sizes

Sample Name	Criteria for Sample Selection	Sampling Pool Size	Number of Volumes Reviewed	Number of Pages Reviewed
	Page-Level Samples			
Production Run #1	Google-Digitized, Publication Date ≤ 1923, English Language	1.3 Million Volumes	1,000	93,858
Production Run #2	Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	1,000	86,439
Production Run #3	Internet Archive Digitized, Publication Date ≤ 1923, English Language, Monograph	850,000 Volumes	1,000	84,539
Production Run #4	Non-Roman Language/Script Digitized Content in HathiTrust 4 Main Language/Script Categories: Arabic, Asian, Cyrillic, Hebrew	1.29 Million Volumes	1,000	91,381
	Whole Volume Error Samples			
Production Run #1a	Same Sampled Volumes from Production Run #1 Google-Digitized, Publication Date ≤ 1923, English Language	1.3 Million Volumes	1,000	397,467
Production Run #2a	Same Sampled Volumes from Production Run #2 Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	1,000	294,505
	Physical Review Samples			
Production Run #1b	Same Sampled Volumes from Production Run #1 Google-Digitized, Publication Date ≤ 1923, English Language	1.3 Million Volumes	906	-
Production Run #2b	Same Sampled Volumes from Production Run #2 *Only University of Michigan Owned Volumes Google-Digitized, Publication Date > 1922, English Language, Monograph	6.5 Million Volumes	584	-