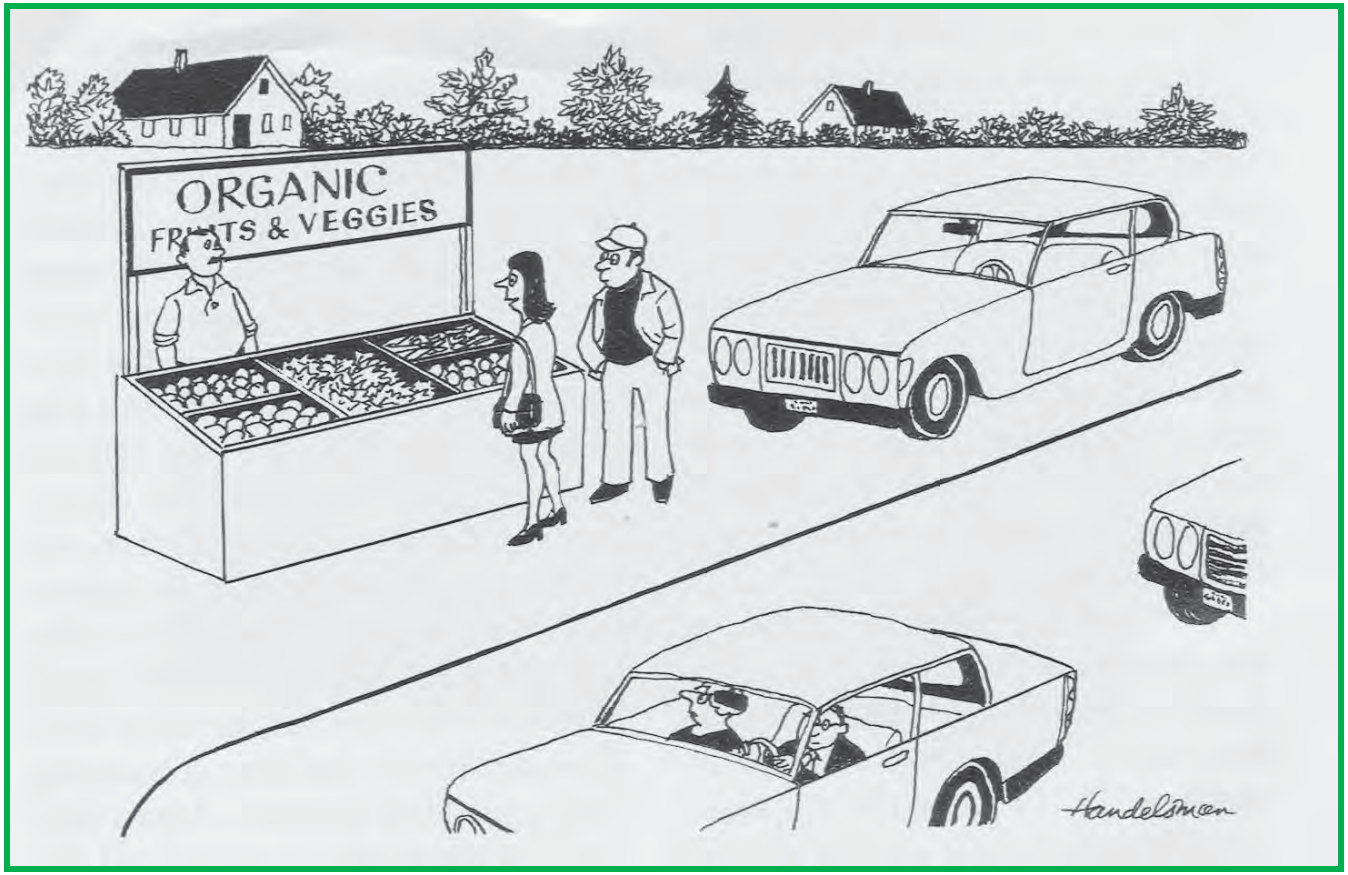
Archival Quality and Long-Term Preservation

Research to Validate the Usefulness of Digital Surrogates



“Organic is nice, but haven’t you got anything digital?”

Outline of Presentation

- Concepts of quality
- Research design to measure quality
- Implications for a theory of archival quality

Archival Quality – A Value Proposition

- Archival nature
  - 1939: Distinguishing characteristic [Garrison]
  - 1970: Permanent records [Fishbein]
- Preservation media
  - 1961: technical characteristics of microfilm [H.G. Jones]
  - 1977: magnetic media & electronic records [Poole]
- Preservation procedures
  - 1989: protection against loss [Conway]
  - 1989: Eschewing “permanence” [O’Toole]
  - 2000: Digital surrogacy [Kenney & Rieger]

Quality and Archival Principle

- Associational value
  - Visual resources as “attachments” [Taylor 1979]
- Reliability [InterPARES]
  - Completeness and process control [Duranti 1995]
- Archival science
  - Processes that generate and structure archival information [Thomassen 2001]
    - Availability, readability, completeness, relevance, representativeness, topicality, authenticity, reliability
- Significant properties
  - Migration of essential elements [Hedstrom & Lee 2002]
- Inviolable properties of the record
  - Physical and intellectual integrity [Vullo et al. 2010]

Information Quality

- IQ framework of attributes and clusters
  - Wang & Strong (1996) – MIS
  - Bovee (2003) – Accounting
  - Stvilia (2007) – Information Science
  - Knight (2008) – IQ/DQ community
- Measuring quality
  - Baird (2004) – Digital image analysis (DIA) for libraries
  - Lin (2006) –Applies DIA to large scale digitization
  - Le Bourgeois (2004) – Need for manual inspection because of weaknesses of image processing

Digital Library Evaluation

- Digital library evaluation establishes [mostly weak] end-user evaluation models and methods
  - Evaluation [Saracevic 2004]
  - Relevance [Saracevic 2007]
  - Few image-based user studies [Harley 2004; Pisciotta 2005]
  - Certification at the repository level [CRL 2007]

- Use case scenarios
  - Adapted from system and interface design [Carroll 2000]
  - Stories articulate requirements [Alexander 2004]

Outline of Presentation

- Concepts of quality
- Research design to measure quality
- Implications for a theory of archival quality

Research Environment

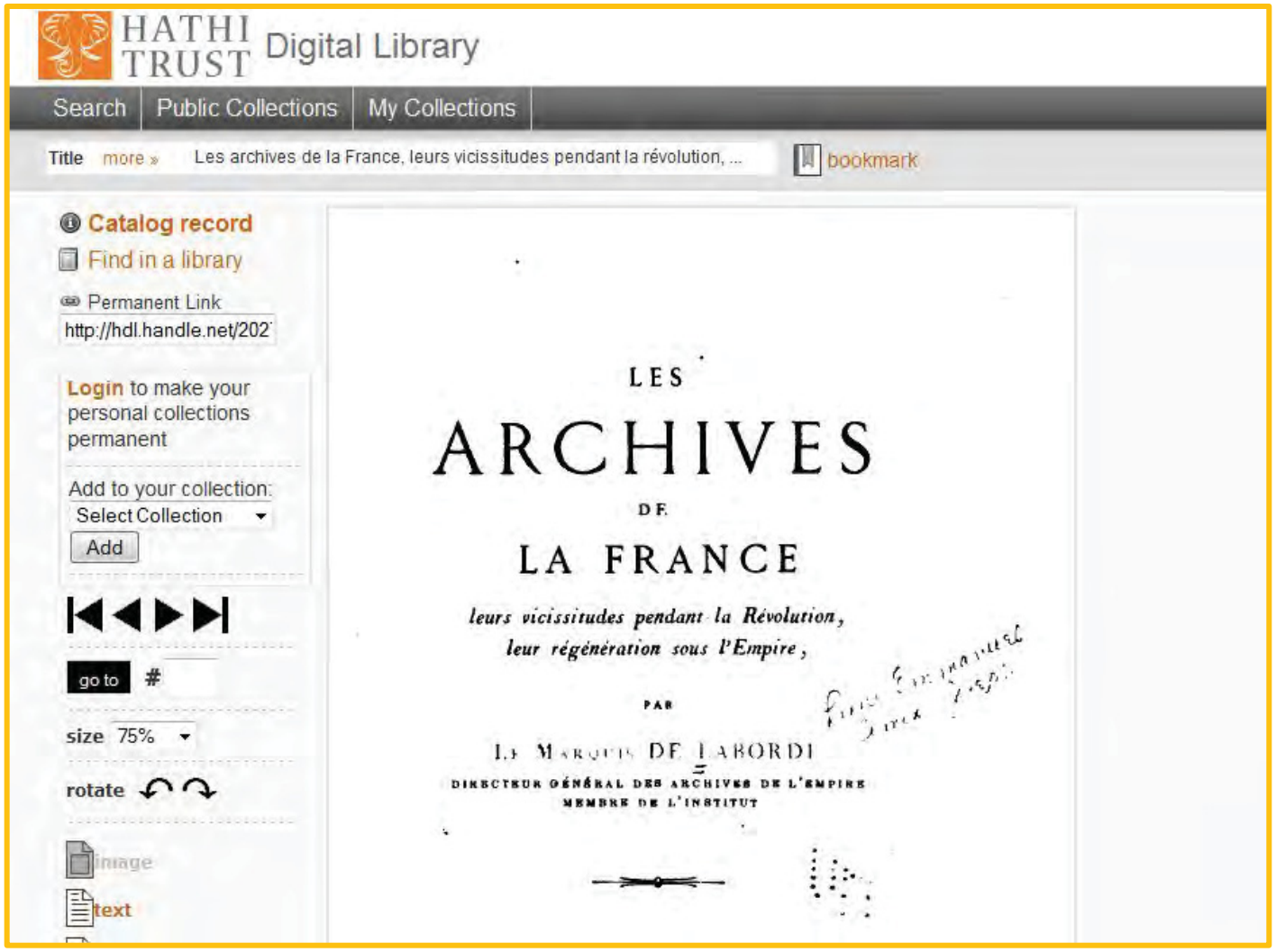
- From vertical integration to distributed management
  - "take what we can get"

- HathiTrust http://www.hathitrust.org/
  - 52 partners
  - 8 million+ volumes
  - Infrastructure, business model, TRAC certification

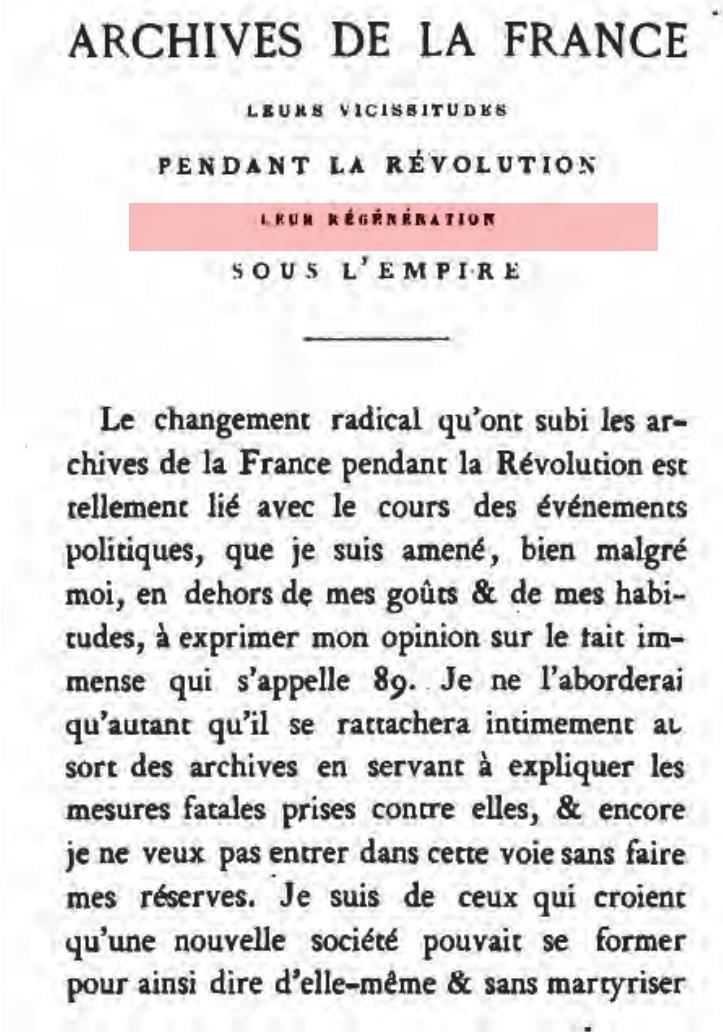
York (2010) "Building a Future by Preserving Our Past"

HATHI TRUST

Les Archives de La France [Laborde, 1867]



Google Book Search: Image and Text



ARCHIVES DE LA FRANCE

LEURS VICISSITUDES

PENDANT LA RÉVOLUTION

LEUR RÉGÉNÉRATION

SOUS L'EMPIRE

Le changement radical qu'ont subi les archives de la France pendant la Révolution est tellement lié avec le cours des événements politiques, que je suis amené, bien malgré moi, en dehors de mes goûts & de mes habitudes, à exprimer mon opinion sur le fait immense qui s'appelle 89. Je ne l'aborderai qu'autant qu'il se rattachera intimement au sort des archives en servant à expliquer les mesures fatales prises contre elles, & encore je ne veux pas entrer dans cette voie sans faire mes réserves. Je suis de ceux qui croient qu'une nouvelle société pouvait se former pour ainsi dire d'elle-même & sans martyriser

LES
ARCHIVES DE LA FRANCE
I. EUKS \iCISSITUDKW
PENDANT LA RÉVOLUTION
1 K U H I! É ', V. K K R A I i O N
sous L'EMPIRE
Le changement radical qu'ont subi les archives de la France pendant la Révolution est tellement lié avec le cours des événements politiques, que je suis amené, bien malgré moi, en dehors de mes goûts & de mes habitudes, à exprimer mon opinion sur le fait immense qui s'appelle 89. Je ne l'aborderai qu'autant qu'il se rattachera intimement au sort des archives en servant à expliquer les mesures fatales prises contre elles, & encore je ne veux pas entrer dans cette voie sans faire mes réserves. Je suis de ceux qui croient qu'une nouvelle société pouvait se former pour ainsi dire d'elle-même & sans martyriser

Research Question 1

What is "intrinsic quality" within the context of digitized books and serials? [or anything bound]

- Hierarchy of information errors based on prior research (IQ/DQ + UM, Google)
- Define and test measures of attribute error
  - Frequency and severity on ordinal scales
- Define and measure correlation effects across measures (co-occurrence)
- Build and test IQ indexes (accuracy, consistency, completeness, redundancy)
  - Cluster and factor analysis

- Outcome: valid quality metrics + indices

Mass Digitization Hysteria

- Data-poor reaction to a variety of socio-political-technical phenomena
  - Viral blogosphere
  - Personalization of quality judgments
  - Opposition to commercialization [Darnton]
- Weaknesses of data analysis
  - University of Michigan – contract compliance review
  - CLIR Mass Digitization Report [2008]
    - Comparative review of four large projects
  - CLIR Study "Idea of Order" [2010]
    - All digital library; costs; large-scale digitization

Ghostlier Demarcations

- Recurring Problems in Mass Digitization [Henry 2010]
  - Alan Gevinson (American Intellectual History)

Google books

	Pre-1923		Post 1922	
	Number	Percent	Number	Percent
Poor Rating	20	32.3%	1	5.6%
Missing Pages	13	21.0%	0	0.0%
Out of Order	4	6.5%	0	0.0%
Duplicate Pages	8	12.9%	0	0.0%
Illegible Pages	6	9.7%	0	0.0%
Obscured Pages	12	19.4%	0	0.0%
Distorted Lines	1	1.6%	1	5.6%
Blurry Pages	13	16.3%	0	0.0%
Not Sharp Pages	9	14.5%	11	61.1%
Total Reviewed	62		18	

http://www.clir.org/pubs/abstract/pub147abst.html

Incidence of Critical Error in HathiTrust

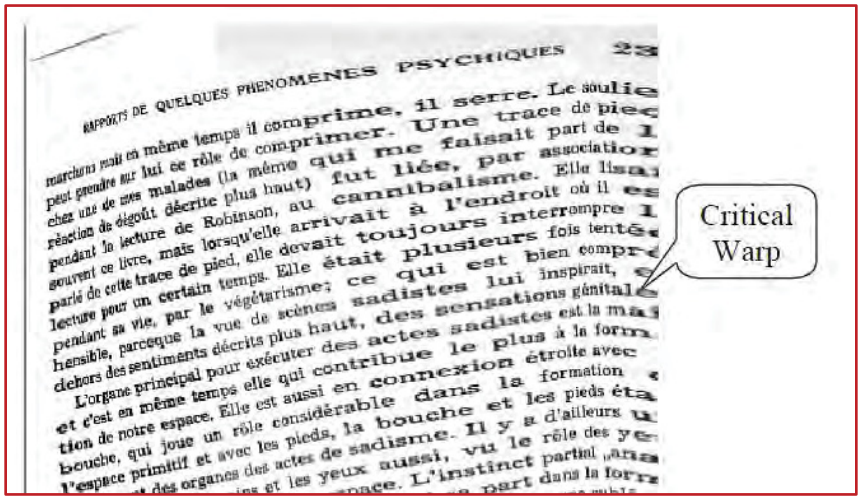
University of Michigan Quality Review, 2006-10

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%

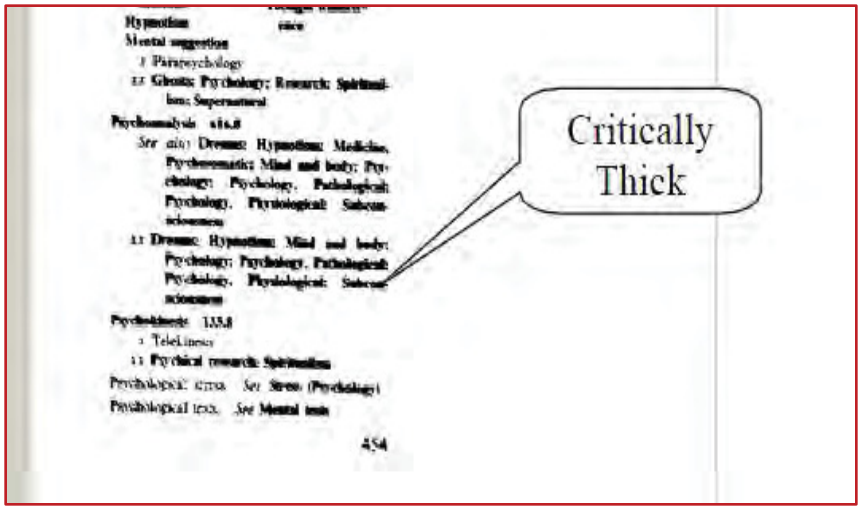
Two Examples ["… flattening & thickening of meaning…"]

Heather MacNeil

Warped Page

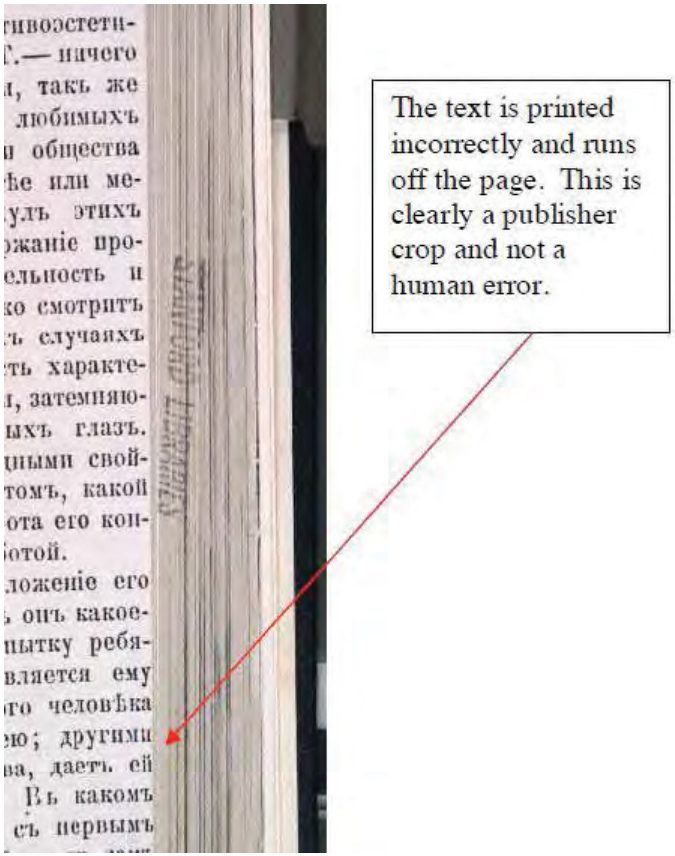


Thick Text

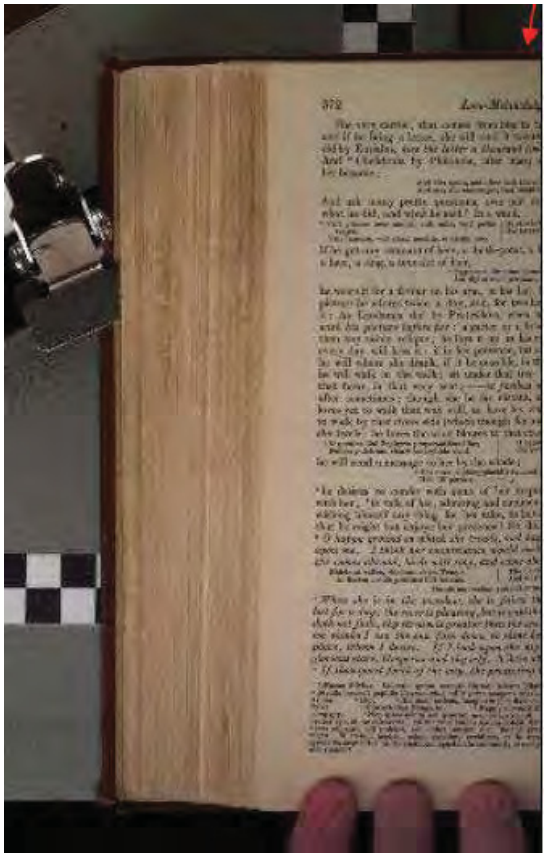


Errors in Source or Scanning

Source Crop



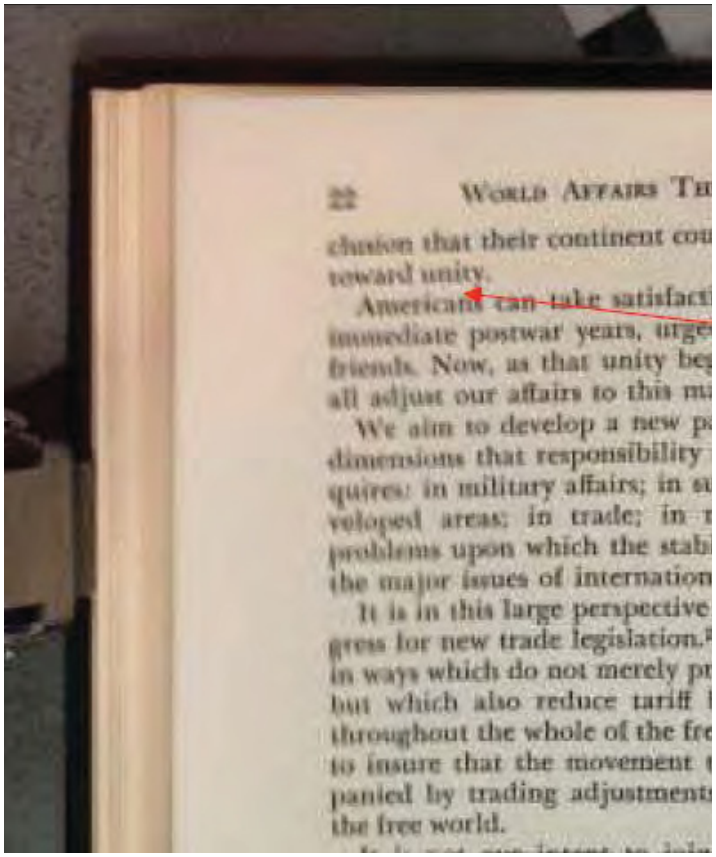
Scan Crop



Errors in Source or Scanning

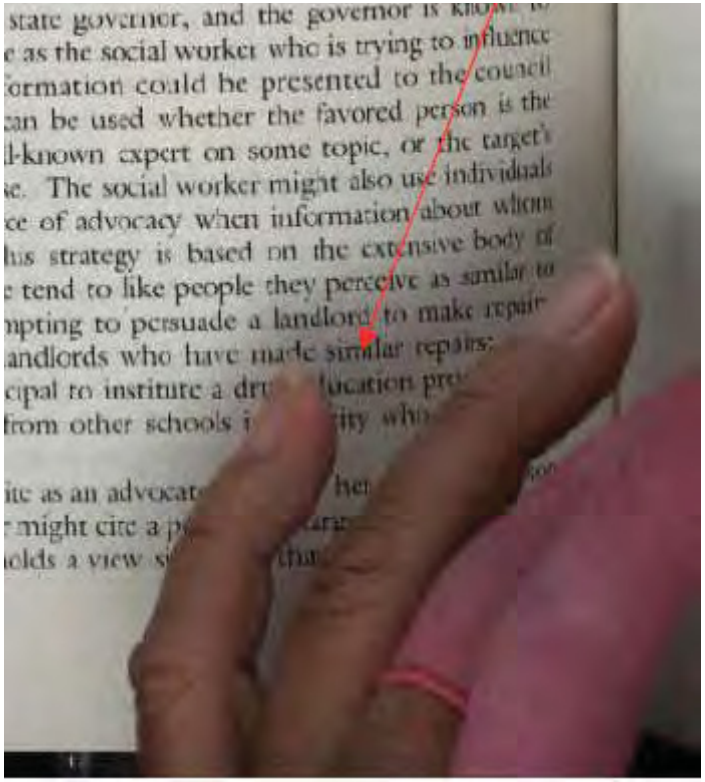
Source Blur

Scanning Blur

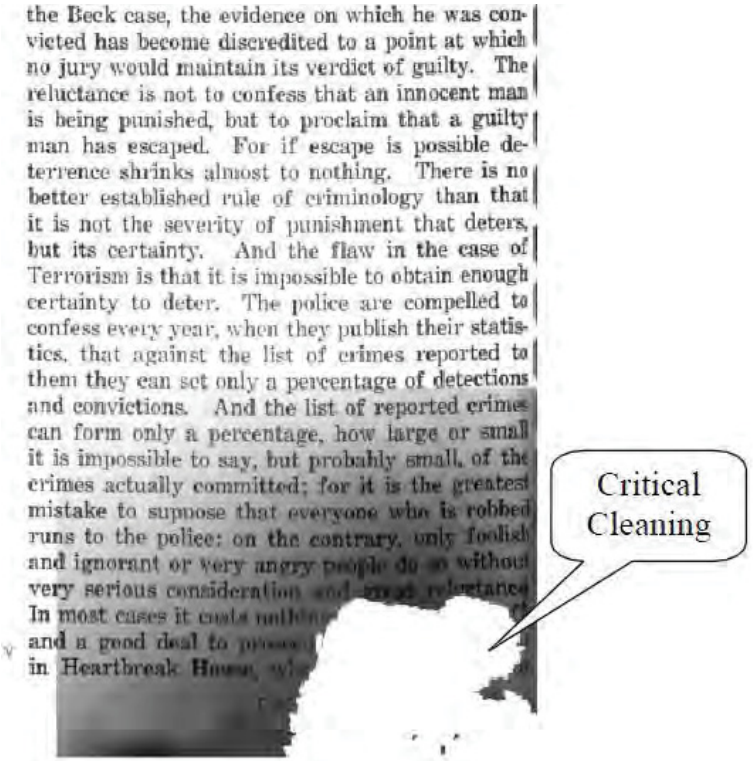


Fingers in Manual Scanning

Traces of human error



Traces digitally cleaned



Error Model

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)

Research Question 2

What is the estimated error-incidence in various clusters of HathiTrust content?

- Apply measures and indices (Q1) within selected strata
  - E.g., pub date; illustrations; source of digitization
- Extensive manual review of many random samples (some including original digitized books)
  - Examine differences between examining entire volume and samples from digital volumes
  - Compare digitized book with original book
- Assess and manage inter-coder inconsistencies in a distributed review model

- Outcome: costs and limits of manual review
- Outcome: identify potential for automated processing of quality review
- Outcome: mechanisms for branding quality using PREMIS metadata framework

Two Views of Validation

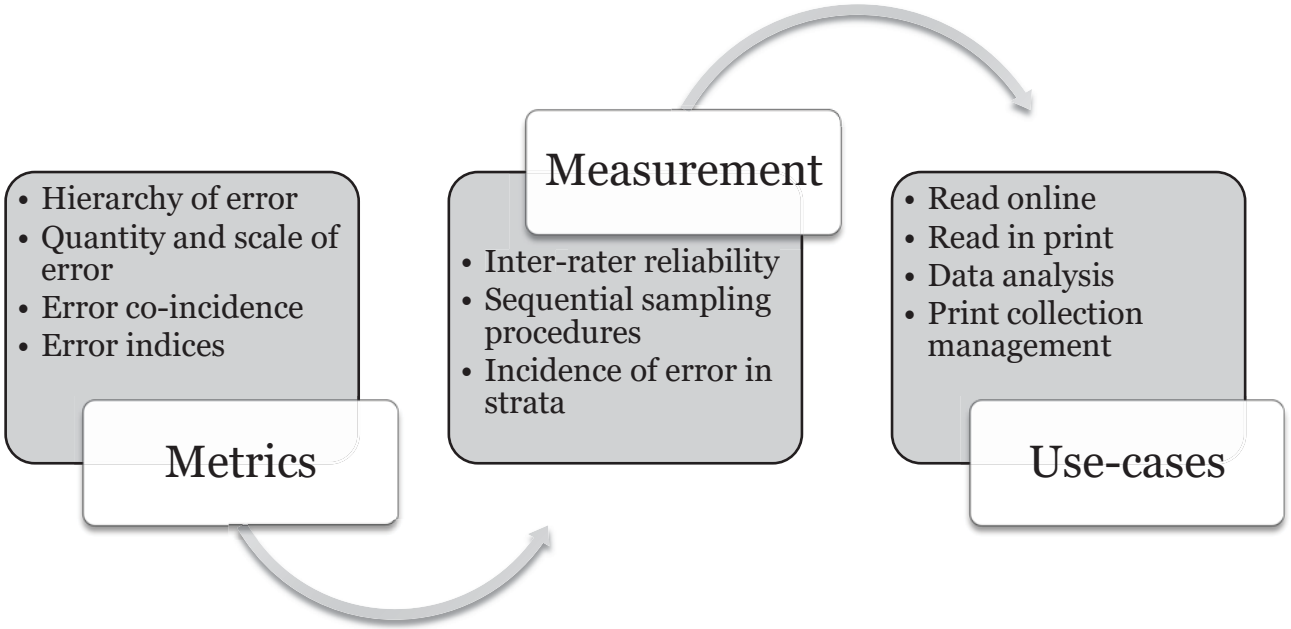
- Objective measurement of phenomena
  - Definition of metrics
  - Testing of metrics
  - Statistical verification and confidence

- Logical consistency from user's perspective
  - Generalized error models
  - Few, but fatal, errors
  - Personalization of error perception

Use Cases

- Reading online
  - Digital page images
  - Text legibility; illustration interpretability; graphic accuracy
- Reading volumes printed on demand
  - Whole or substantial parts of volumes
  - Accuracy, completeness, consistency
- Processing full-text data
  - Underlying text content
  - Accuracy thresholds, readiness for analysis; “non-consumptive”
- Managing print collection
  - Surrogacy of the whole
  - Low cumulative error; non-critical errors; completeness; redundancy

Research Workflow



Outline of Presentation

- Concepts of quality
- Research design for measuring quality
- Implications for a theory of archival quality

Implications for Preservation/DL Practice

- Tools and techniques for measuring quality
- Expose content quality as part of certification process
- Limitations of use case scenarios
  - Fruitless pursuit of complete user satisfaction
- Need for automated quality validation routines
  - Error models as first steps toward machine processing
  - Distinguishing errors that matter from those that don't

- Proposition: Certification of trustworthy repositories must encompass the content within.

Implications for Archival Theory

- Digital "archiving" through preservation is theoretically defensible
- Establish the archival nature of digitized surrogates
- Establish preservation value of digital surrogates
- Reaffirm relationship of provenance and reliability
- Archival quality defined through use

- Proposition: In the digital world, archival quality is the absence of error relative to expected uses.

Acknowledgements

- Planning: Andrew W. Mellon Foundation
- Research: Institute of Museum and Library Services
- Support: HathiTrust, John Wilkin, Exec. Director

- Planning team (Mellon)
  - Jeremy York (HathiTrust)
  - Emily Campbell (Mlibrary)
  - Nikki Calderone (School of Information)
  - Devan Donaldson (School of Information)
  - Sarah Shreeves (University of Illinois)
  - Robin Dale (Lyrasis)
- Research team (IMLS)
  - Ed Rothman, co-PI
  - Jackie Bronicki, coordination
  - Ken Guire, statistician
  - Ryan Rotter, system design
  - Jeremy York, liaison
  - John Butler, U. Minnesota
  - Advisory Board

References (1)

- Alexander IF & Maiden NAM, eds. (2004). Scenarios, stories and use cases. John Wiley, New York.
- 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.
- Bovee M, Srivastava R, Mak B (2003) A conceptual framework and belief-function approach to assessing overall information quality. International Journal of Intelligent Systems. 18 (1): 51-74.
- Carroll J (2000) Making use: Scenario-based design of human-computer interactions. Cambridge, MA: The MIT Press.
- Center for Research Libraries (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. Accessed 27 September 2010.
- Conway P (1989) "Archival preservation: Definitions for improving education and training." Restaurator 10 (1): 47-60.
- 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.
- Duranti, L. (1995). "Reliability and authenticity: The concepts and their implications." *Archivaria* 39: 5-10.

References (2)

- Fishbein MH (1970) "A Viewpoint on appraisal of national records." American Archivist 33 (2): 175-187.
- Garrison C (1939) The relation of historical manuscripts to archival materials. American Archivist 2 (2): 97-105.
- Harley, Diane et al. (2006). *Use and Users of Digital Resources: a Focus on Undergraduate Education in the Humanities and Social Sciences* Center for Studies in Higher Education, UC Berkeley.
- Hedstrom, M. and Lee, C. (2002). "Significant properties of digital objects: Definitions, applications, implications." Paper presented at DLM-Forum 2002, Barcelona, Spain, http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf
- 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, DC. See supplemental online report and data by Gevinson A (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 27 September 2010.
- Jones, H. G. (1961). "North Carolina's Local Records Program," *American Archivist* 24 (1): 25-41.

References (3)

- Kenney AR & Rieger OY (2000) Moving theory into practice: Digital imaging for libraries and archives. Research Libraries Group, Mountain View, CA.
- Knight S (2008) User Perceptions of Information Quality in World Wide Web Information Retrieval Behaviour. (PhD Dissertation) Perth, Australia: Edith Cowan University.
- Le Bourgeois F, Trinh E 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.
- 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.
- O'Toole, JM (1989) On the idea of permanence. American Archivist 52 (1): 10-25.
- Pisciotta, H. et al. (2005). Penn State's Visual Image User Study," portal: Libraries and the Academy 5 (January): 33-58.
- Poole FG (1977) "Some aspects of the conservation problem in archives," American Archivist 40 (2): 163-171.
- Rieger O (2008) Preservation in the age of large-scale digitization: A white paper. Council on Library and Information Resources, Washington, DC.
- 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 September 2007).

References (4)

- Saracevic, T. (2004). Tefko Saracevic, "How were digital libraries evaluated?" Paper first presented at the DELOS WP7 Workshop on the Evaluation of Digital Libraries.
- Saracevic, T. (2007). "Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance." *Journal of the American Society for Information Science and Technology* 58 (13): 2126-2144.
- 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.
- Taylor HA (1979) "Documentary art and the role of the archivist," American Archivist 42 (4): 417-428.
- Thomassen T (2001) A first introduction to archival science. Archival Science 1: 373-385.
- Vullo G, Innocenti P, Ross S (2010) Towards policy and quality interoperability: Challenges and approaches for digital libraries, Conference Proceedings IS&T Archiving 2010, 1-4 June, The Hague, The Netherlands, Society for Imaging Science and Technology, pp. 33-38.
- Wang R and Strong D (1996) Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12 (4): 5-34.
- 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

Thank you for your attention!

Paul Conway, Associate Professor
School of Information, University of Michigan pconway@umich.edu