Archival Quality in Digital Preservation Repositories

Constructing a New Approach to Metrics

"I aspired to authenticity, but I never got beyond verisimilitude."

© *The New Yorker* 6/25/2007

Outline of Presentation

- Concepts
- Research Design
- Implications

Archival Quality – A Value Proposition

- Archival nature
  - 1939 on : distinguishing characteristics of archives
- Preservation media and procedures
  - 1961 on : technical characteristics of longevity
  - 1985 on : protection against loss
- Reliability [InterPARES]
  - 1995 on : completeness and process control
- Significant properties
  - 2001 on : migration of essential elements

Information Quality – DL Evaluation

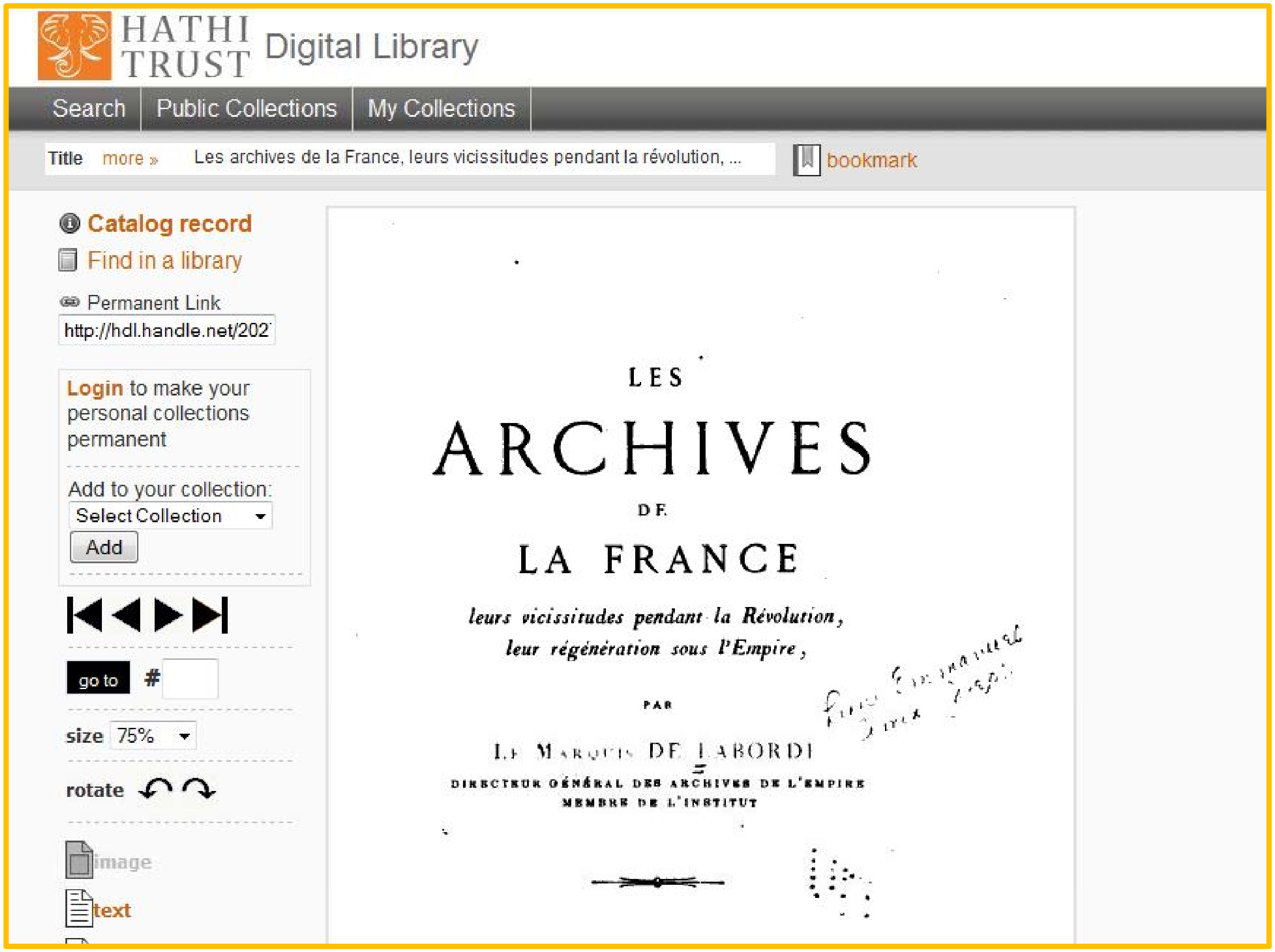
- IQ research establishes framework of attributes and clusters
  - Wang & Strong (1996); Lin (2006) – MIS
  - Bovee (2003) – Accounting
  - Stvilia (2007); Rieh (2002) – Information Science
  - Knight (2008) – IQ/DQ community
- Digital library evaluation establishes [mostly weak] end-user evaluation models and methods
  - Saracevic (2005) – retrieval effectiveness
  - Saracevic (2007) – weaknesses in relevance research
  - Harley (2004); Pisciotta (2005) – image based user studies

Research Environment

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

- HathiTrust
  - 27 partners
  - 6 million+ volumes
  - Infrastructure, business model, TRAC certification
- Google hysteria
  - Data-poor reaction to a variety of socio-political-technical phenomena

Les Archives de La France [Laborde, 1867]



Title more » Les archives de la France, leurs vicissitudes pendant la révolution, ...

Permanent Link http://hdl.handle.net/202

Login to make your personal collections permanent

Add to your collection: Select Collection

#

size 75%

LES

ARCHIVES

DE

LA FRANCE

leurs vicissitudes pendant la Révolution,
leur régénération sous l'Empire,

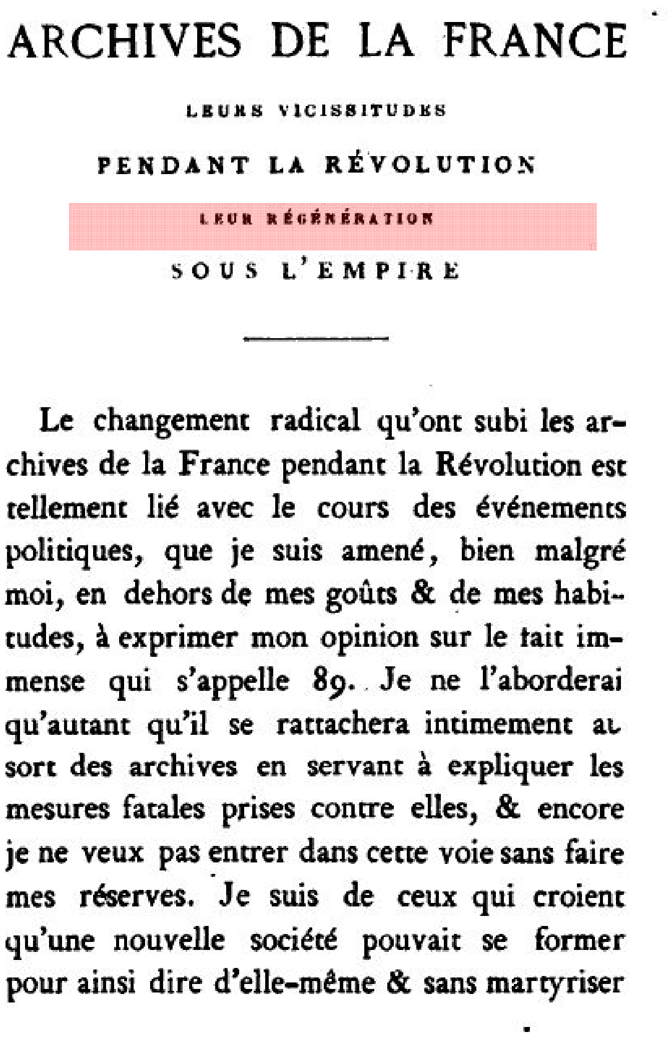
PAR

LE MARQUIS DE LABORDE

DIRECTEUR GÉNÉRAL DES ARCHIVES DE L'EMPIRE
MEMBRE DE L'INSTITUT

Pour Emmanuel [illegible]

Google Book Search: Image and Text



ARCHIVES DE LA FRANCE

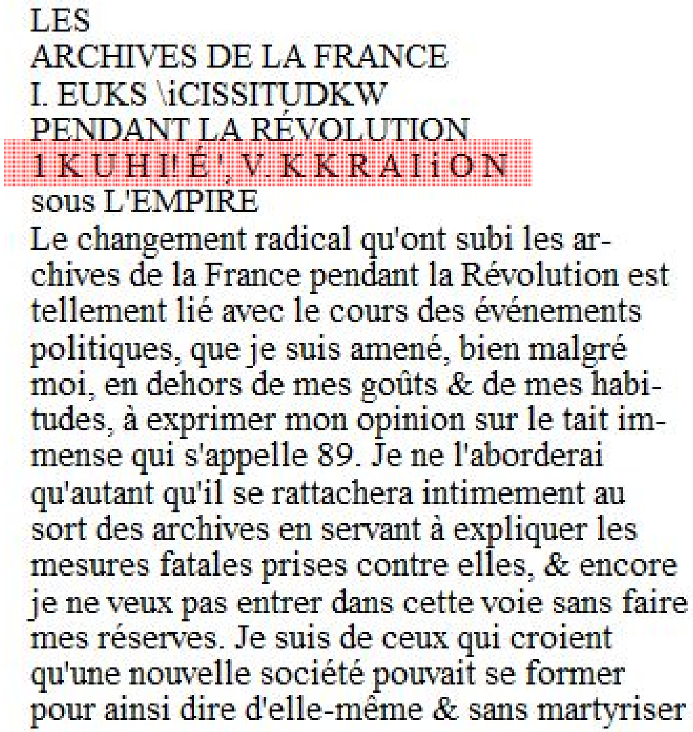
LEURS VICISSITUDES

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
1KUHI!É',V.KKRAIiON
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 tait 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

Archival Quality / Large-Scale Digitization

At Present: Quality Standards	Trends Forward: Acceptable Loss
Material centered	User centered
One size fits all	Fitness for use
Vertical integration	Third party creators
Process control	Acceptance testing
Compromise is failure	Good enough is a value

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

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

Outline of Presentation

- Concepts
- Research Design
- Implications

Research Question 1

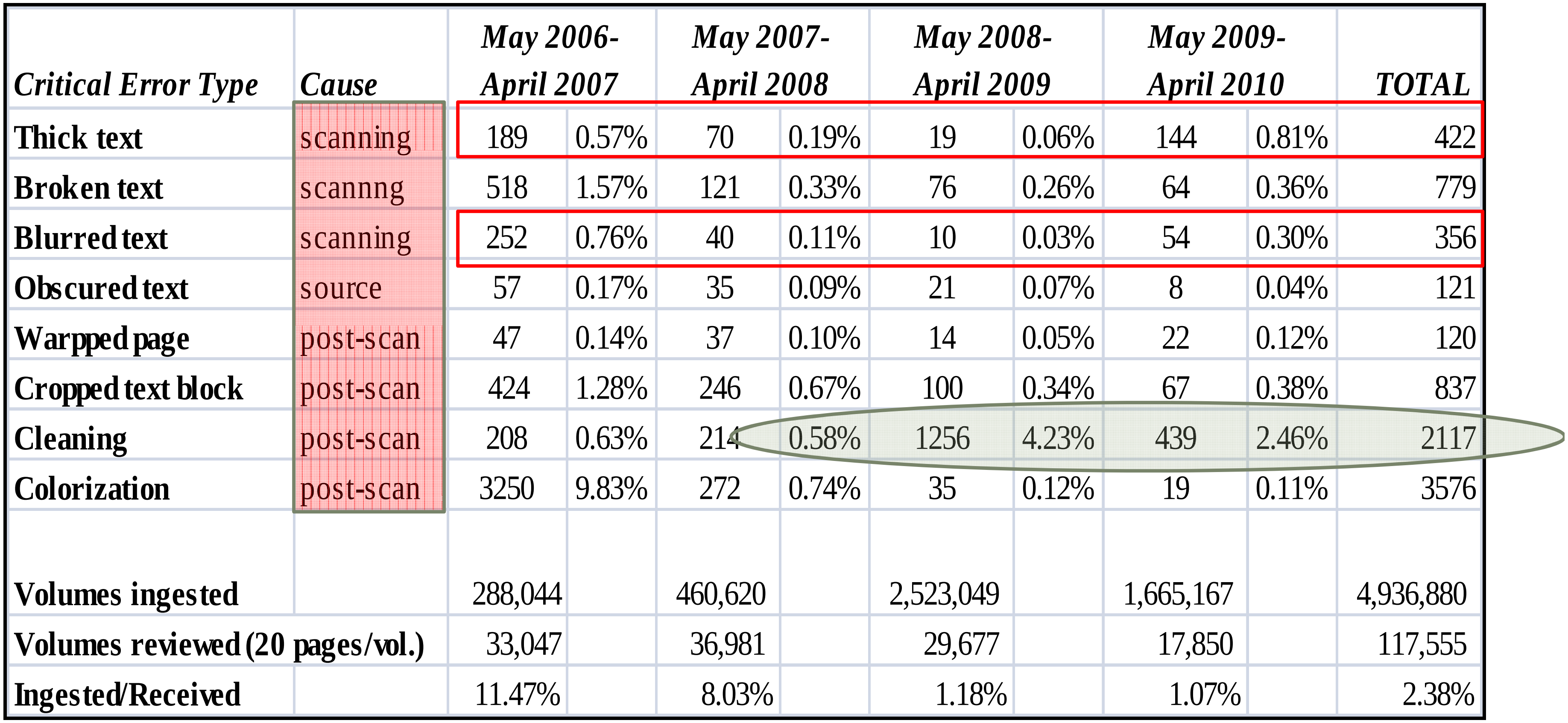
What is "intrinsic quality" within the context of digitized books and serials? [or anything bound]

- 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

Incidence of Critical Error in HathiTrust

University of Michigan Quality Review, 2006-10



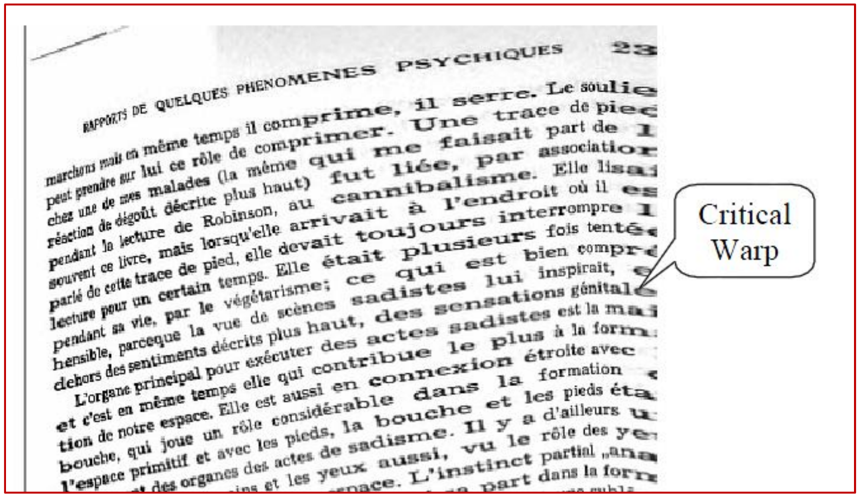
Critical Error Type	*Cause*	*May 2006-April 2007*		*May 2007-April 2008*		*May 2008-April 2009*		*May 2009-April 2010*		*TOTAL*
Thick text	scanning	189	0.57%	70	0.19%	19	0.06%	144	0.81%	422
Broken text	scannng	518	1.57%	121	0.33%	76	0.26%	64	0.36%	779
Blurred text	scanning	252	0.76%	40	0.11%	10	0.03%	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

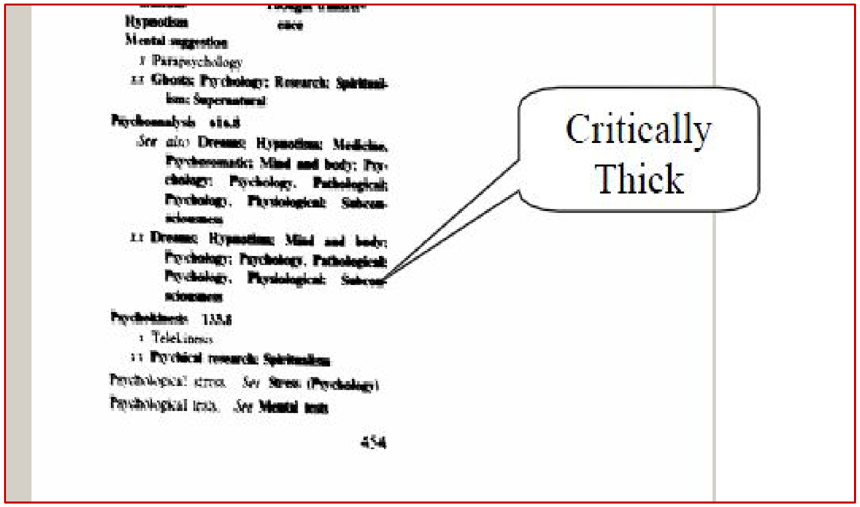
Thick Text



RAPPORTS DE QUELQUES PHENOMENES PSYCHIQUES 23

marchons mais en même temps il comprime, il serre. Le soulie
peut prendre sur lui ce rôle de comprimer. Une trace de piec
chez une de mes malades (la même qui me faisait part de l
réaction de dégoût décrite plus haut) fut liée, par associatio
pendant la lecture de Robinson, au cannibalisme. Elle lisai
souvent ce livre, mais lorsqu'elle arrivait à l'endroit où il es
parlé de cette trace de pied, elle devait toujours interrompre l
lecture pour un certain temps. Elle était plusieurs fois tenté
pendant sa vie, par le végétarisme; ce qui est bien compr
hensible, parceque la vue de scènes sadistes lui inspirait, e
dehors des sentiments décrits plus haut, des sensations génital
L'organe principal pour exécuter des actes sadistes est la mai
et c'est en même temps elle qui contribue le plus à la form
tion de notre espace. Elle est aussi en connexion étroite avec
bouche, qui joue un rôle considérable dans la formation
l'espace primitif et avec les pieds, la bouche et les pieds éta
t des organes des actes de sadisme. Il y a d'ailleurs u
ins et les yeux aussi, vu le rôle des ye
espace. L'instinct partial „ana
a part dans la form

Critical Warp



Hypnotism ence
Mental suggestion
x Parapsychology
xx Ghosts; Psychology; Research; Spiritualism; Supernatural

Psychoanalysis 616.8
See also Dreams; Hypnotism; Medicine, Psychosomatic; Mind and body; Psychology; Psychology, Pathological; Psychology, Physiological; Subconsciousness
xx Dreams; Hypnotism; Mind and body; Psychology; Psychology, Pathological; Psychology, Physiological; Subconsciousness

Psychokinesis 133.8
x Telekinesis
xx Psychical research; Spiritualism

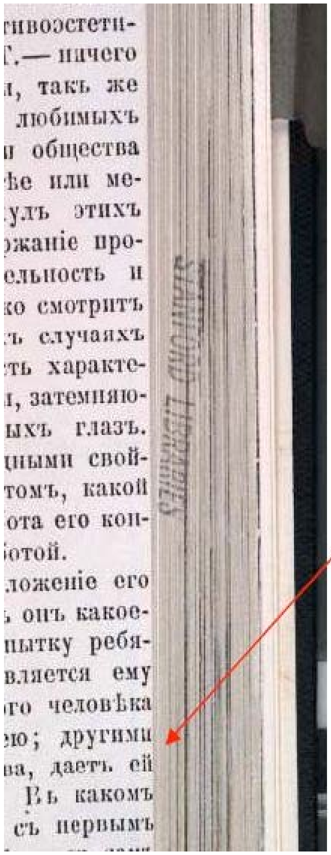
Psychological stress. See Stress (Psychology)

Psychological tests. See Mental tests

454

Errors in Source or Scanning

Source Crop

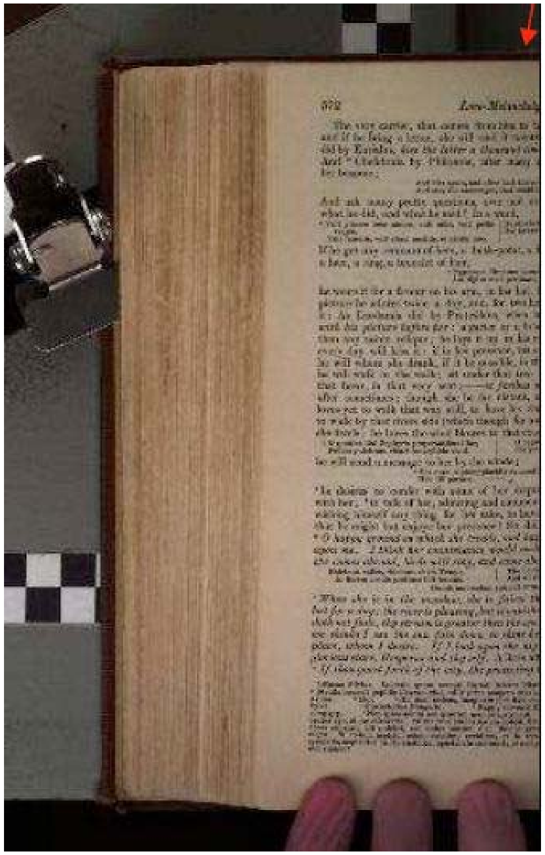


тивоэстети-
Г.— ничего
л, такъ же
любимыхъ
и общества
ѣе или ме-
уль этихъ
ржаніе про-
ельность и
ко смотрить
ъ случаяхъ
ть характе-
л, затемняю-
ыхъ глазъ.
цными свой-
томъ, какой
ота его кон-
ботой.
ложеніе его
, онъ какое-
пытку ребя-
вляется ему
го человѣка
ю; другими
на, даетъ ей
Въ какомъ
съ первымъ

STANFORD LIBRARIES

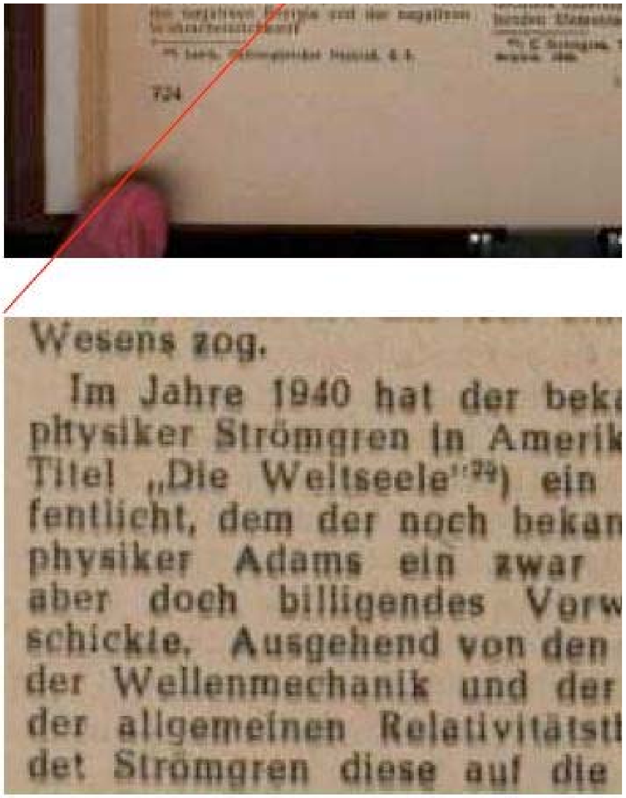
The text is printed incorrectly and runs off the page. This is clearly a publisher crop and not a human error.

Scan Crop



Errors in Source or Scanning

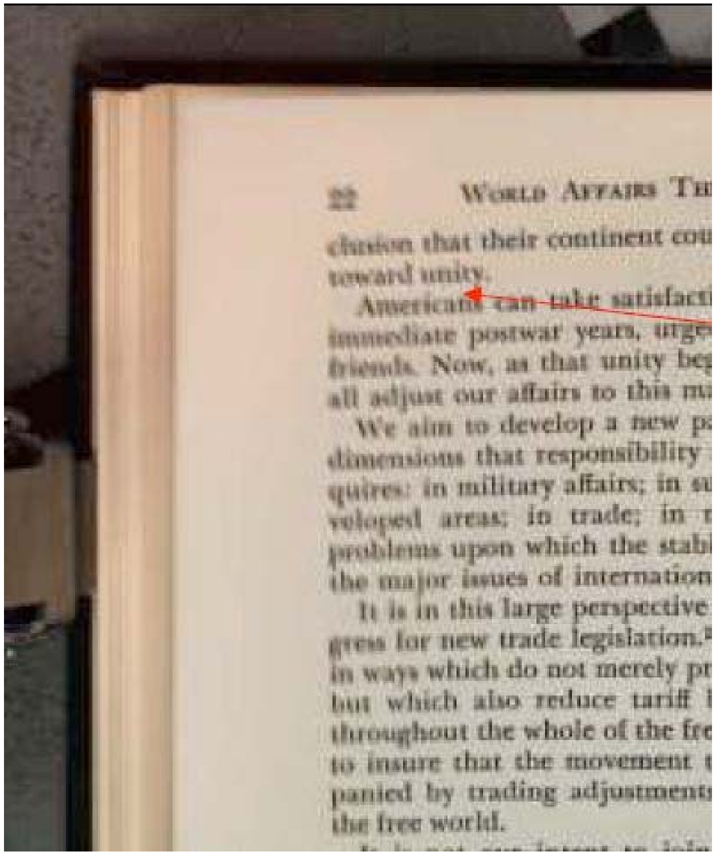
Source Blur



724

Wesens zog.
Im Jahre 1940 hat der beka
pltysiker Strömgren in Amerik
Titel „Die Weltseele''²²) ein
fentlicht, dem der noch bekan
physiker Adams ein zwar
aber doch billigendes Vorw
schickte. Ausgehend von den
der Wellenmechanik und der
der allgemeinen Relativitätstl
det Strömgren diese auf die

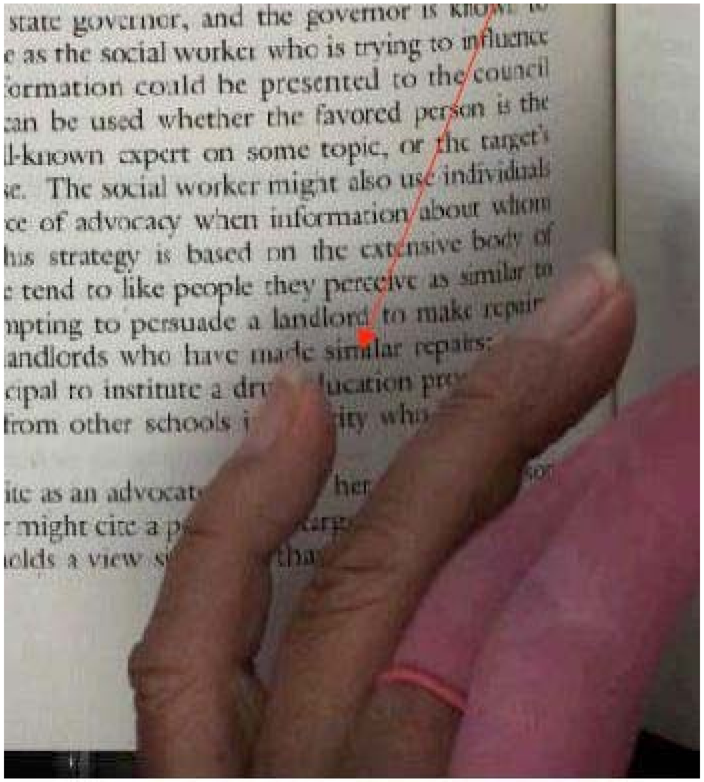
Scanning Blur



22 WORLD AFFAIRS TH
clusion that their continent cou
toward unity.
Americans can take satisfacti
immediate postwar years, urge
friends. Now, as that unity beg
all adjust our affairs to this ne
We aim to develop a new p
dimensions that responsibility
quires: in military affairs; in su
veloped areas; in trade; in t
problems upon which the stabi
the major issues of internation
It is in this large perspective
gress for new trade legislation.²
in ways which do not merely p
but which also reduce tariff b
throughout the whole of the fre
to insure that the movement
panied by trading adjustments
the free world.

Fingers in Manual Scanning

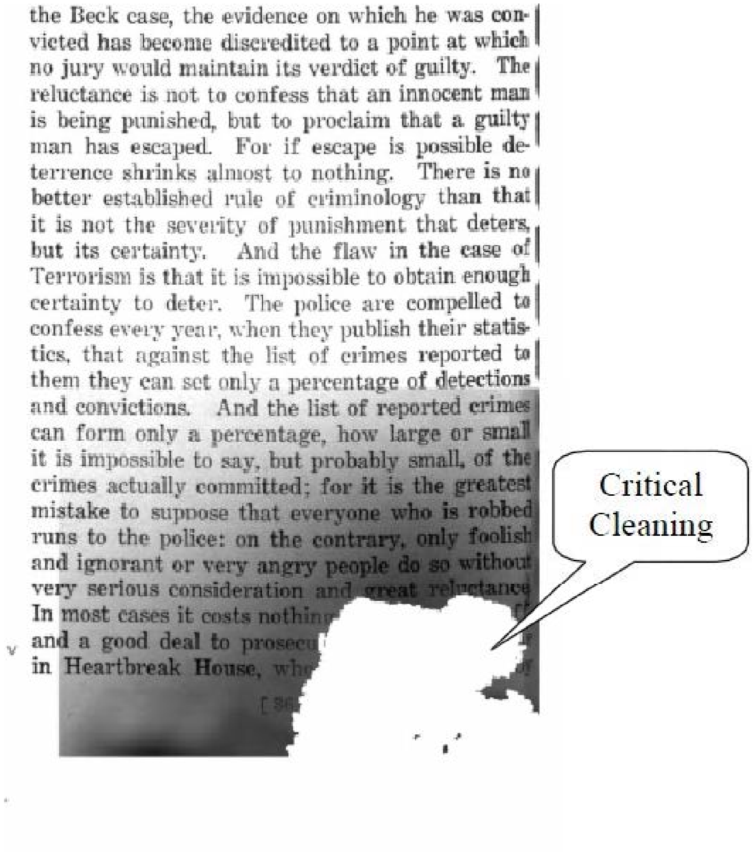
Traces of human error



state governor, and the governor is known to
e as the social worker who is trying to influence
formation could be presented to the council
can be used whether the favored person is the
ll-known expert on some topic, or the target's
se. The social worker might also use individuals
ce of advocacy when information about whom
his strategy is based on the extensive body of
e tend to like people they perceive as similar to
mpting to persuade a landlord to make repair
landlords who have made similar repairs;
cipal to institute a dru[illegible]ucation pro
from other schools i[illegible]ity who

ite as an advocat[illegible] her[illegible] sor
might cite a p[illegible]arge
olds a view s[illegible] tha

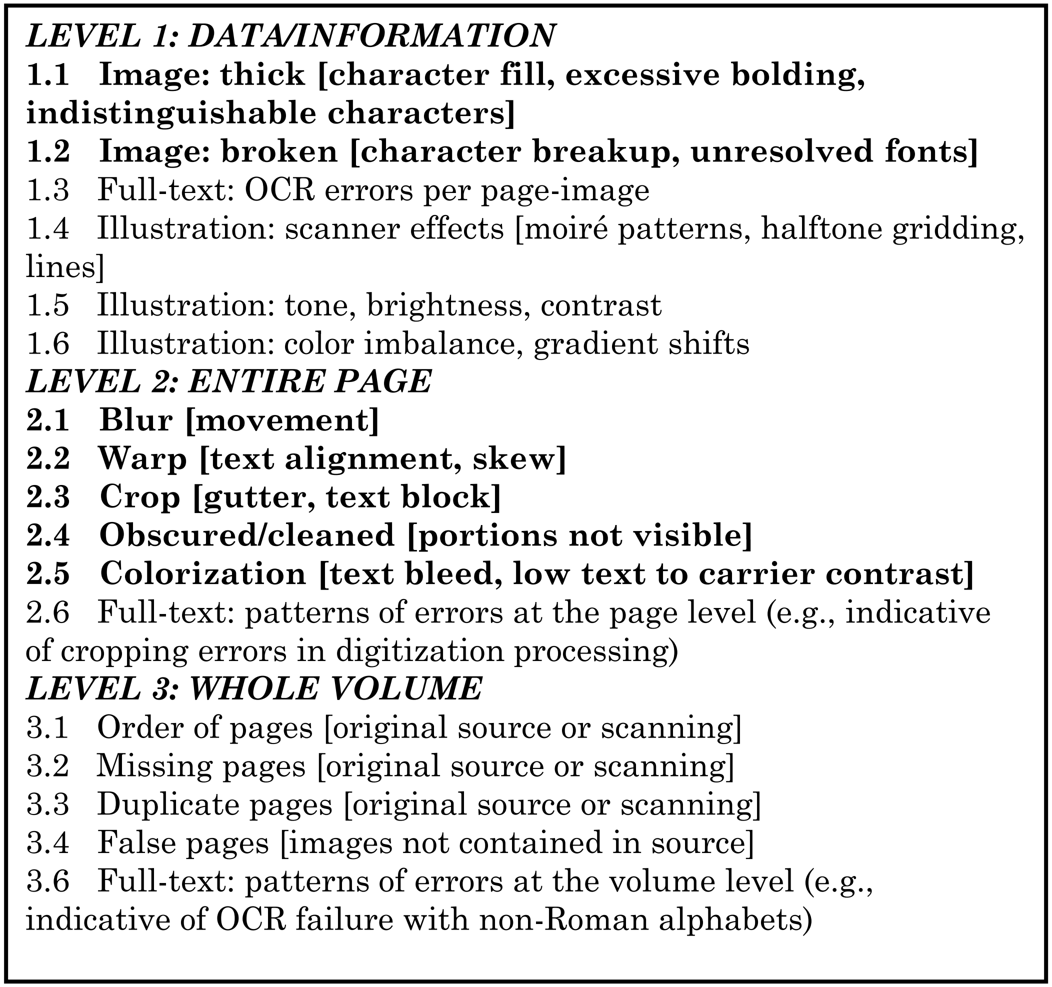
Traces digitally cleaned



the Beck case, the evidence on which he was convicted has become discredited to a point at which no jury would maintain its verdict of guilty. The reluctance is not to confess that an innocent man is being punished, but to proclaim that a guilty man has escaped. For if escape is possible deterrence shrinks almost to nothing. There is no better established rule of criminology than that it is not the severity of punishment that deters, but its certainty. And the flaw in the case of Terrorism is that it is impossible to obtain enough certainty to deter. The police are compelled to confess every year, when they publish their statistics, that against the list of crimes reported to them they can set only a percentage of detections and convictions. And the list of reported crimes can form only a percentage, how large or small it is impossible to say, but probably small, of the crimes actually committed; for it is the greatest mistake to suppose that everyone who is robbed runs to the police: on the contrary, only foolish and ignorant or very angry people do so without very serious consideration and great reluctance. In most cases it costs nothin
and a good deal to prosecu
in Heartbreak House, wh

Critical Cleaning

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)

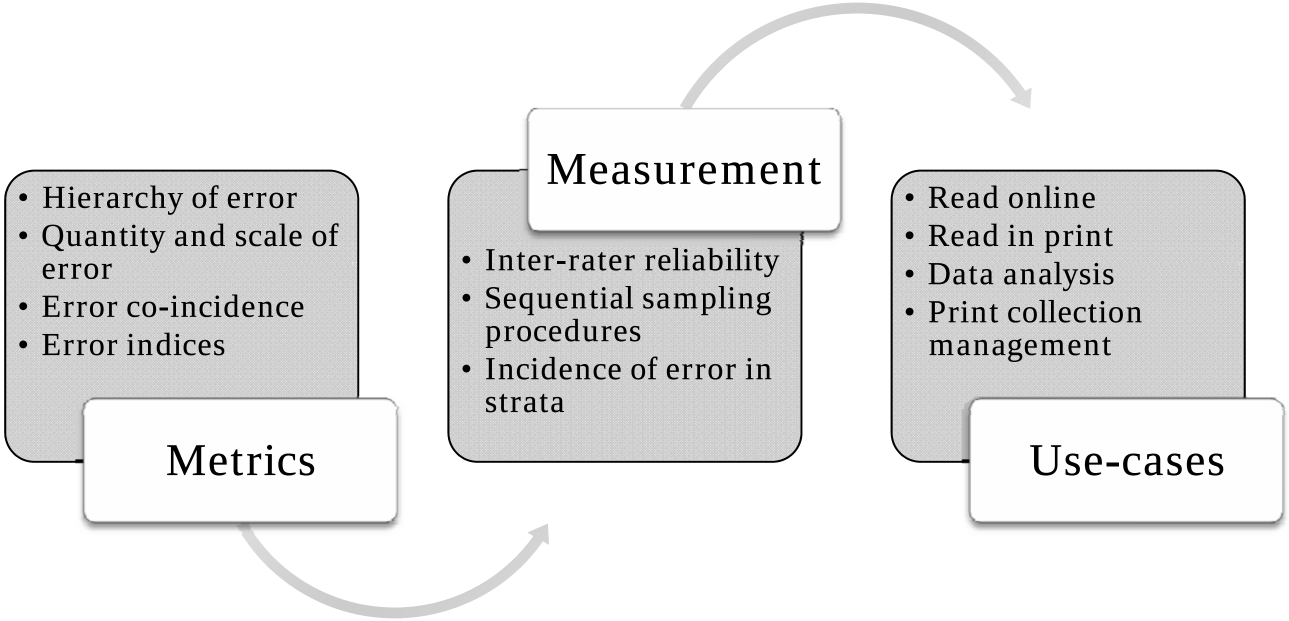
Research Question 2

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

Research Workflow



Metrics

- Hierarchy of error
- Quantity and scale of error
- Error co-incidence
- Error indices

Measurement

- Inter-rater reliability
- Sequential sampling procedures
- Incidence of error in strata

Use-cases

- Read online
- Read in print
- Data analysis
- Print collection management

Outline of Presentation

- Concepts
- Research Design
- Implications

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

- Question: To what extent can or should a fundamental archival principle be measured?

Acknowledgements

- Planning support from Andrew W. Mellon Foundation

- John Wilkin, Exec. Director, HathiTrust

- Planning team
  - Jeremy York (HathiTrust)
  - Emily Campbell (Mlibrary)
  - Nikki Calderone (School of Information)
  - Devan Donaldson (School of Information)
  - Sarah Shreeves (University of Illinois)
  - Robin Dale (Lyrasis)

References (1)

- Bovee, M, Srivastava, R and Mak, B (2003). A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. *International Journal of Intelligent Systems*. 18 (1): 51-74.
- Cockburn, A (2000). *Writing Effective Use Cases*. Boston: Addison-Wesley.
- Harley, D et al. (2006). Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences. Berkeley: Center for Studies in Higher Education.
- Knight, S (2008). *User Perceptions of Information Quality in World Wide Web Information Retrieval Behaviour*. (PhD Dissertation) Perth, Australia: Edith Cowan University.
- 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.
- Pisciotta, H et al. (2005). Penn State’s Visual Image User Study,” portal: Libraries and the Academy 5 (January): 33-58.

References (2)

- Rieh, S (2002). Judgment of Information Quality and Cognitive Authority in the Web. *Journal of the American Society for Information Science and Technology* 53 (2): 145-161.
- Saracevic, T. (2000). “Digital Library Evaluation: Toward an Evolution of Concepts.” Library Trends 49 (2): 349-369.
- Saracevic, T. (2004). *How Were Digital Libraries Evaluated?* Paper first presented at the DELOS WP7 Workshop on the Evaluation of Digital Libraries, p. 9.
- 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.
- Tanner, S, Munoz, T., and Ros, P. (2009). Measuring Mass Text Digitization Quality and Usefulness. *D-Lib Magazine* 15 (July/August): 209.
- 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 (2009). This Library Never Forgets: Preservation, Cooperation, and the Making of HathiTrust Digital Library. *Proc. IS&T Archiving* 2009, Arlington, VA, pp. 5-10.

Thank you for your attention!

Paul Conway, Associate Professor

School of Information, University of Michigan pconway@umich.edu