Validating the Quality of Digitized Books for Long-Term Preservation

Findings on Digitization Error and Archival Quality
Outline of Presentation

- Archival quality in an archival context
- Selective findings (more with poster and on website)
- Implications for practice and theory

Project Website: http://hathitrust-quality.projects.si.umich.edu/
45 People Are Involved

- Planning: Andrew W. Mellon Foundation [$49,000]
- Research: Institute of Museum and Library Services [$674,722]

**Research Team (Michigan)**
- Paul Conway, School of Information, PI
- Ed Rothman, LSA Statistics, co-PI
- John Wilkin, HathiTrust (5%)
- Leigann Ayers, oversight and space (5%)
- Jeremy York, liaison (10%)
- Jackie Bronicki, MLibrary, coordination (100%)
- Ken Guire, CSCAR, statistician (10%)
- Ryan Rotter, MLibrary, system design (75%)
- Stacy Maat, SI, physical review coordinator (25%)
- Melissa Chalmers, SI, user research (50%)
- Eugene Malin & Nina Elias, SI, website (10%)
- Sarah Jones, qualitative coder
- Jennifer Wright, qualitative coder
- Cherie Edmonds, qualitative coder
- Sarah Helm, qualitative coder
- Jenny Vainberg, qualitative coder
- Nadia Sion, qualitative coder
- Naomi Scheinemann, qualitative coder
- Siqi Wei, qualitative coder

**Research Team (Minnesota)**
- John Butler, leadership
- Jason Roy, liaison
- Ahnna Mahoney, coordination
- Kelly Frosch, qualitative coder
- Megan Scherer, qualitative coder
- Rui Zhao, qualitative coder
- Christina Graber, qualitative coder

**Advisory Board**
- Roy Tennant, OCLC
- Oya Rieger, Cornell University
- Ed Van Gemert, Wisconsin
- Robin Dale, LYRASIS
- Besiki Stvilia, Florida State Univ.

**13 School of Information Volunteers**
- Julia Corrin, Natalie Bond, Graham Hukill, Jacqueline DiOrio, Janice Wong, Laura Brubacher, Brendan Coates, Laura Andrews, Holly Little, Monique Lowe, Molly Des Jardin, Amelia Lowry, David Fulmer
Archival Quality - A Value Proposition

- Archival nature
  - 1939 on: distinguishing characteristics of archives
  - 2000 on: significant properties of digital objects

- Preservation
  - 1961 on: media longevity [e.g., microfilm and acid-free paper]
  - 1985 on: processes to protect against loss [archival processes]
  - 1990 on: digitization image quality [archival master]

- Reliability [InterPARES]
  - 1995 on: completeness and process control

“... degree of completeness and degree of control of the procedure of creation are the only two factors that determine reliability of records.” [Duranti 1995, p. 6]
Archival Quality in Archival Theory

- Seamus Ross: digital libraries are digital archives

  “But if we think more carefully about digital libraries we easily observe that they may be libraries by name, but they are archives by nature.” [Ross 2007, p. 8]

- Terry Cook: context of creation

  “This new paradigm for [has] a renewed focus on the context, purpose, intent, interrelationships, functionality, and accountability of the record, its creator, and its creation processes, wherever these occur.” [Cook 1997, p. 48]

- Geoffrey Yeo: boundary objects & persistent representation

  “Records are “persistent representations of activities... or other occurrents... created by participants or observers of those occurrents or by their authorized proxies...” [Yeo 2008, p. 136]
Archival Quality is Archival Science [Thomassen 2001]

- **Object of Archival Science:**
  - defining the nature of “process-bound information,” which is “information itself and the processes that have generated and structured that information.”

- **Aims of Archival Science:**
  - “the establishment and maintenance of archival quality, that is to say: of the optimal visibility and durability of the records, the generating work processes, and their mutual bond.”

- **Methodology of Archival Science:**
  - “maintaining the formal quality of process-bound information, by ensuring its availability, readability, completeness, relevance, representativeness, topicality, authenticity and reliability.”

Research Environment - Digitization

- New preservation reality: from vertical integration to distributed management
  - Preservation programs used to exercise end-to-end control of reformatting
  - Now: preserving digitized content: “take what we can get”
  - End-user trust turns on validating “fitness-for-use”

- Two Research Questions for our project
  - What is quality? [definition, measurement, distribution]
  - What difference does lack of quality make for users? [barriers, acceptance testing]

Testbed: HathiTrust Digital Library

- [http://www.hathitrust.org/]

Andrew W. Mellon Foundation [planning]
IMLS NLG [research, reporting]
Rethinking **Quality** for Preservation and Access

Research Workflow [2011-13]

1. Metrics 2011
   - Error model
   - Quantity and severity of error
   - Inter-rater reliability
   - Data-gathering system

2. Measurement 2011-12
   - Population sampling
   - Sequential sampling procedures
   - Distribution of error
   - Co-occurrence of error
   - Correlation with traits

3. Use studies 2012-13
   - Read online
   - Collection management
   - File to print transform
   - Text-Image prediction

Podium Talk!
Poster!
Next Year!
## Phase 1 [2011] - Metrics of Digitization Error

<table>
<thead>
<tr>
<th>Level of Abstraction</th>
<th>Possible Cause of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LEVEL 1: DATA/INFORMATION</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 Text: thick text [fill, excessive]</td>
<td>Source or post-processing</td>
</tr>
<tr>
<td>1.2 Text: broken text [character breakup]</td>
<td>Source or post-processing</td>
</tr>
<tr>
<td>1.3 Illustration: scanner effects [moire, gridding]</td>
<td>Scanning or post-processing</td>
</tr>
<tr>
<td>1.4 Illustration: tone, brightness, contrast</td>
<td>Scanning, post-processing or source</td>
</tr>
<tr>
<td>1.5 Illustration: color imbalance, gradient shifts</td>
<td>Scanning, post-processing or source</td>
</tr>
<tr>
<td><strong>LEVEL 2: ENTIRE PAGE</strong></td>
<td></td>
</tr>
<tr>
<td>2.1 Blur [distortion]</td>
<td>Scanning or source</td>
</tr>
<tr>
<td>2.2 Warp [text alignment]</td>
<td>Post-processing</td>
</tr>
<tr>
<td>2.3 Skew [page alignment]</td>
<td>Scanning, source or post-processing</td>
</tr>
<tr>
<td>2.4 Crop [gutter, text block]</td>
<td>Source or post-processing</td>
</tr>
<tr>
<td>2.5 Obscured/cleaned [portions not visible]</td>
<td>Scanning or post-processing</td>
</tr>
<tr>
<td>2.6 Colorization [text bleed, low contrast]</td>
<td>Source or post-processing</td>
</tr>
<tr>
<td><strong>LEVEL 3: WHOLE VOLUME</strong></td>
<td></td>
</tr>
<tr>
<td>3.1 Fully obscured [foldouts or objects]</td>
<td>Scanning</td>
</tr>
<tr>
<td>3.2 Missing pages [one or more]</td>
<td>Source or scanning</td>
</tr>
<tr>
<td>3.3 Duplicate pages [one or more]</td>
<td>Source or scanning</td>
</tr>
<tr>
<td>3.4 Order of pages</td>
<td>Source or scanning</td>
</tr>
<tr>
<td>3.5 False pages [not part of Original Content]</td>
<td>Scanning or post-processing</td>
</tr>
</tbody>
</table>
Phase 1 [2011] - Error Severity Scale

0 - Error is undetectable on the page.

1 - Error has negligible affect on Original Content.

2 - Error alters appearance of Original Content.

3 - Error has affects readability of Original Content.

4 - Error requires significant inference to read Original Content.

5 - Error renders Original Content undecipherable.
Error Type - Warp

the only 1st class passenger on board. She ought to have carried 60, at least.

January 19th. Still blowing hard. It being sun-
day, the crews of the foreign men-of-war have been
on shore, and the amount of drunkenness, and noisy
in the streets, could not have been surpassed by
Portsmouth itself in the good old days, when it
seemed to be Jack's first duty to let the inhabitants
realize that he was on shore.

January 20th. Coaled and watered. There is no
watering requires more attention. The prices must
go your full weight of coal or full quantity of water,
be fixed beforehand, of course. Then, if after all you
will be lucky. The Consul-General can give
some good advice about this. He advises you to enter into an arrangement with any of
the vessels who are on board to do this for your account. The vessel you send ashore, if through a
mistake I have left at the Consul-General's office, it is
racking in both cases.

Thursday 21st, I called on His Highness Kamil
efendi (late Grand Vizier). He was a very interesting
man, and next day the most fluent English. On January 22nd,
and again, he returned the call at the Consul's house. His view of the Armenian difficulty
was most evident, to publish, but I may say that it
may or may not be innocuous, except when there
sharp alert, even treasons physically, and so incapable

Error Type – Crop

Clinical

Functions of the
Graves—Barr (Prus i). Knowledge of the function
extent obtained by medical
he has recently seen in
thought and speech, with
blood-pressure, that all
the effects of a usual
ally due to thyroid is
large value of thyroid.
and Barr himself has to
treatment in the aged.
B hypothyroidism, and Barr is
due to masticity, dyspneic,
and thyroid. He has a
valuable in cases of lung
the calcium metabolism,
its value in arteriosclerosis
holds that the beneficial
to its stimulating effect.
The well-known on
the present day since it is
an useful and restorative.
An to hyperthyro-
marked cases the
without on.
extension, active hyper-
rapid action of the
mental excitement, the
ceased; and slight or
degree of hypothyroidism
of albinism, especially of
may or may not be in-
except when there
sharp alert, even treasons
physically, and so incapables.
Phase 2 [2011-12] - Error Detection, Coding, Analysis

- Random Samples of Digital Volumes
  - Populations: pre-1923 \( N = 1.3 \) million  post-1923 \( N = 6.5 \) million
  - Samples: 1,000 volumes per study
    1. Google (pre-1923, English)
    2. Google (post-1923, English, no serials)
    3. Internet Archive (pre-1923, English)
    4. non-Roman scripts (250 volumes in four alphabets)
  - Systematic sampling strategy within each volume
    - Up to 100 pages per volume, evenly distributed front to back
    - Up to 25% of a volume, evenly distributed

- Coding: page-level, whole volume, physical inspection
  - Coding of 456,217 page-images [for 11 errors]
    - Double coding of 10% of each sample [ca. 45,000 page-images]
  - Coding of 2,000 whole volumes [for volume level errors]
  - Coding of 1,490 physical volumes for book/bib. characteristics
Findings: Comparison of Most Frequent Errors

Total coding of 178,297 page-images digitized by Google

<table>
<thead>
<tr>
<th>Severity = 0</th>
<th>Severity = 1</th>
<th>Severity = 4</th>
<th>Severity = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;&lt;&lt;&lt; 1923 &gt;&gt;&gt;&gt;</td>
<td>&lt;&lt;&lt;&lt; 1923 &gt;&gt;&gt;&gt;</td>
<td>&lt;&lt;&lt;&lt; 1923 &gt;&gt;&gt;&gt;</td>
</tr>
<tr>
<td>Text</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thick</td>
<td>62.04%</td>
<td>67.52%</td>
<td>25.66%</td>
</tr>
<tr>
<td>Broken</td>
<td>61.00%</td>
<td>73.37%</td>
<td>29.96%</td>
</tr>
<tr>
<td>Page</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crop</td>
<td>99.37%</td>
<td>98.85%</td>
<td>0.27%</td>
</tr>
<tr>
<td>Warp</td>
<td>29.22%</td>
<td>45.78%</td>
<td>60.18%</td>
</tr>
<tr>
<td>Obscure</td>
<td>16.88%</td>
<td>56.83%</td>
<td>78.05%</td>
</tr>
</tbody>
</table>

|               | 182,205      | 490          | 972          |              |
| Portion of Total Error (pre-1923) | 96.9% | 82.5% | 87.9% | |
|               | 113,682      | 795          | 1,077        | |
| Portion of Total Error (post-1923) | 90.5% | 87.9% | 86.5% | |
## Findings: Comparison of Most Frequent Errors

Total coding of **85,535** page-images digitized by Internet Archive

<table>
<thead>
<tr>
<th>Severity = 0</th>
<th>Severity = 1</th>
<th>Severity = 4</th>
<th>Severity = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;&lt;&lt;&lt; 1923</td>
<td>&lt;&lt;&lt;&lt; 1923</td>
<td>&lt;&lt;&lt;&lt; 1923</td>
<td>&lt;&lt;&lt;&lt; 1923</td>
</tr>
<tr>
<td><strong>Text</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Thick</em></td>
<td>93.23%</td>
<td>4.12%</td>
<td>0.01%</td>
</tr>
<tr>
<td><em>Broken</em></td>
<td>81.71%</td>
<td>11.77%</td>
<td>0.19%</td>
</tr>
<tr>
<td><strong>Illustration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Tone</em></td>
<td>69.00%</td>
<td>24.65%</td>
<td>0.17%</td>
</tr>
<tr>
<td><strong>Page</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Crop</em></td>
<td>99.61%</td>
<td>0.22%</td>
<td>0.02%</td>
</tr>
<tr>
<td><em>Warp</em></td>
<td>41.13%</td>
<td>57.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td><em>Obscure</em></td>
<td>56.93%</td>
<td>39.61%</td>
<td>0.02%</td>
</tr>
<tr>
<td><em>Colorization</em></td>
<td>47.56%</td>
<td>45.22%</td>
<td>0.00%</td>
</tr>
<tr>
<td><em>Skew</em></td>
<td>90.24%</td>
<td>9.33%</td>
<td>-</td>
</tr>
<tr>
<td><em>Blur</em></td>
<td>94.22%</td>
<td>4.28%</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>95,293</th>
<th>199</th>
<th>206</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portion of &quot;Big Five&quot; Errors</td>
<td>57.40%</td>
<td>52.00%</td>
<td>77.40%</td>
</tr>
</tbody>
</table>
Findings: Distribution of Severe Error

- Proportion of volumes with severe error
- Level 4 or 5 severity in any error type on any page-image

<table>
<thead>
<tr>
<th>Pages w/ Severe Error</th>
<th>Number of Volumes Google</th>
<th>Number of Volumes Internet Archive</th>
<th>Cumulative Percent Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>555</td>
<td>876</td>
<td>93.19%</td>
</tr>
<tr>
<td>1</td>
<td>167</td>
<td>43</td>
<td>97.99%</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>7</td>
<td>98.51%</td>
</tr>
<tr>
<td>3</td>
<td>39</td>
<td>2</td>
<td>98.72%</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>0</td>
<td>98.72%</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>3</td>
<td>99.04%</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>3</td>
<td>99.36%</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>0</td>
<td>99.36%</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>2</td>
<td>99.57%</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>99.57%</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0</td>
<td>99.57%</td>
</tr>
<tr>
<td>11 to 21</td>
<td>20</td>
<td>1</td>
<td>99.68%</td>
</tr>
<tr>
<td>11 to 28</td>
<td>28</td>
<td>1</td>
<td>99.68%</td>
</tr>
<tr>
<td>22 to 68</td>
<td>8</td>
<td>3</td>
<td>100.00%</td>
</tr>
<tr>
<td>38 to 168</td>
<td>10</td>
<td>3</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>932</td>
<td>921</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
## Phase 1 [2011] - Metrics of Digitization Error

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

- Average number of pages per volume with whole book error.
- Certainty of loss of Original Content on some part of a page.

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages per Volume</td>
<td>997</td>
<td>397.49</td>
<td>272.75</td>
<td>8</td>
<td>1628</td>
</tr>
<tr>
<td>Whole Book Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obscured Content</td>
<td>997</td>
<td>3.36</td>
<td>20.82</td>
<td>0</td>
<td>366</td>
</tr>
<tr>
<td>Missing Page(s)</td>
<td>997</td>
<td>0.67</td>
<td>6.529</td>
<td>0</td>
<td>155</td>
</tr>
<tr>
<td>Duplicate Page(s)</td>
<td>997</td>
<td>0.62</td>
<td>4.48</td>
<td>0</td>
<td>92</td>
</tr>
<tr>
<td>Out of Order</td>
<td>997</td>
<td>0.24</td>
<td>2.185</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>False Page (s)</td>
<td>997</td>
<td>0.04</td>
<td>0.343</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

Page-level Quality Errors in a Volume

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sure Loss 1/3 page</td>
<td>996</td>
<td>2.44</td>
<td>11.69</td>
<td>0</td>
<td>177</td>
</tr>
<tr>
<td>Sure Loss 2/3 page</td>
<td>996</td>
<td>0.72</td>
<td>4.22</td>
<td>0</td>
<td>68</td>
</tr>
<tr>
<td>Sure Loss all page</td>
<td>996</td>
<td>0.58</td>
<td>0.046</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>Unsure Loss 1/3 page</td>
<td>996</td>
<td>1.51</td>
<td>7.94</td>
<td>0</td>
<td>156</td>
</tr>
<tr>
<td>Unsure Loss 2/3 page</td>
<td>996</td>
<td>0.11</td>
<td>0.603</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Unsure Loss all page</td>
<td>996</td>
<td>0.206</td>
<td>1.49</td>
<td>0</td>
<td>28</td>
</tr>
</tbody>
</table>

Mean = 4.93 pages/volume

Mean = 5.56 pages/volume
### Impact of Physical Characteristics on Error

**Number of Pages per Volume with at least 1 severe error**

<table>
<thead>
<tr>
<th>Characteristic of Volume</th>
<th>1 - 3 Pages</th>
<th>≥ 7 Pages</th>
<th>Chi-Square (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Publication Year</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1860</td>
<td>70</td>
<td>19</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>1866-1899</td>
<td>70</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>≥ 1900</td>
<td>110</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td><strong>Gutter Margin</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than 1.0 cm</td>
<td>184</td>
<td>20</td>
<td>0.002</td>
</tr>
<tr>
<td>Less than 1.0 cm</td>
<td>66</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td><strong>Graphic Content</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>153</td>
<td>20</td>
<td>0.004</td>
</tr>
<tr>
<td>Yes</td>
<td>97</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td><strong>Parts of Pages Missing</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>241</td>
<td>37</td>
<td>0.06</td>
</tr>
<tr>
<td>Yes</td>
<td>9</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Binding</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fully Intact</td>
<td>190</td>
<td>27</td>
<td>0.06</td>
</tr>
<tr>
<td>Loose</td>
<td>41</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Not Intact</td>
<td>12</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Missing All/Part</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Next step in analysis:** map physical characteristics to specific error types.
More Data Gathering - 2012

- Special focus on graphics and illustrations
  - Compile examples of digitization error
  - Get diagnosis from panel of imaging scientists
  - Explore options for correcting error

- Digitization processes documentation and analysis
  - Google digitization/post-processing techniques
  - Internet Archive post-processing techniques
  - HathiTrust ingest and re-ingest routines
  - Costs and limitations of manual review of digitization error
Study Phase 3 [2012-13] - Two Major User Studies

- **Reading online – Error threshold for user rejection**
  - **Concepts:** Text legibility; illustration interpretability
  - **Scholarship:** IQ (intrinsic); Relevance clues (object); Readability
  - **Population:** Digital humanities scholars who use books as primary sources
  - **Goal:** Identify thresholds of acceptability (limbo bar)

- **Managing library print collections**
  - **Concepts:** Low cumulative error; completeness; redundancy
  - **Scholarship:** “Last copy” criteria and policy
  - **Population:** HathiTrust members: collection development and preservation librarians
  - **Goal:** Certify individual volumes as “fit for use” (high bar)
Study Phase 3 [2012] - Process Study

- Predicting error from text to/from page-image
  - Spatial mapping of error text landscapes [training set] with hOCR text file [Google, IA, JSTOR]
- Partnership with SI Professor Qiaozhu Mei’s research team

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 nothing and a good deal to prosecute.

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 nothing and a good deal to prosecute.

In Heartbreak House, wh@
[36]
Deliverables [2013]

- **Report findings on website**
  - Tables, analysis, links to data
  - [http://hathitrust-quality.projects.si.umich.edu/](http://hathitrust-quality.projects.si.umich.edu/)

- **Publish peer-reviewed articles & proceedings**
  - American Archivist, Archival Science, JASIST, IJDL, IJIQ
  - JCDL, iPres, IS&T Archiving
  - College & Research Libraries, First Monday

- **Distribute Quality Review web-application(s)**
  - Three tools that use sampling strategies
  - One tool for volume-by-volume certification
Summary

- **What is quality?**
  - Absence of **page-image error** relative to expected uses.
  - Presence of **intrinsic character** sufficient to inspire trust.
  - “Fit for purpose” – exploring the limits of “one size fits all”

- **How bad is it?**
  - Very low incidence of very severe error?
    - Likely findable with automated processing
  - High incidence of low-severity text error (Google)
  - Very low incidence of whole volume error
    - Unlikely findable with machine processing algorithms
  - Very high incidence (likely) of scanner effects on book illustrations

- **Why does error occur?**
  - Physical book characteristics have little or no impact
  - Faith of digitizers in post-scan image processing at scale
Implications for Practice

- Lowering the bar on image quality is not necessarily an ethical or professional compromise
- New tools and techniques for measuring quality will emerge from this study
- Communicating error to users is important
- 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 qualities of the content within.
Implications for Archival Theory

An archival principal [archival quality] can be described empirically.
- Scoping the “intrinsic value” of copies [Boon 2010]

Reaffirm value of digital surrogates as preservable products
- Preservation trumps access as a compelling archival rationale

Establish the archival nature of digitized surrogates
- “Archivalness” derives from creation processes [reliability]
- Provenance derives in part from digital curation
- Appraisal of value through assessment of use
References

References

Project Summary

IN LESS THAN A DECADE the large-scale digitization of books has begun transforming the way we read and learn and changing how research libraries manage and preserve their collections.

DIGITIZED BOOKS made by third-party vendors are being preserved in online repositories. In this new preservation environment, the quality of what is preserved becomes an important factor in inspiring trust that digitized books are fit for the purposes envisioned for them.

INNOVATIVE RESEARCH presented at this website is developing and testing methods for measuring the severity of detectable errors in digitized books and validating the impact of error on the end-user. Here you will find information on the project, selected findings, and links to the project’s reports, presentations, publications, and products.

HATHITRUST DIGITAL LIBRARY serves as a testbed of digitized books and serials for the project, which has three overlapping phases.

- Phase 1 (2011) - Define a model of digitization error and a severity scale for recording observed error consistently and accurately.
- Phase 2 (2011-12) - Apply the research methodology to representative samples of digitized volumes.
- Phase 3 (2012-13) – Validate the results of the error analysis for specific use-case scenarios.

THE SCHOOL OF INFORMATION at the University of Michigan is leading a multi-year collaboration with the University of Michigan Press, the HathiTrust Digital Library, and the University of Minnesota Libraries. The
Thank you for your attention!

Project Website: http://hathitrust-quality.projects.si.umich.edu/

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