Archival Quality in Digital Preservation Repositories

Constructing a New Approach to Metrics
“I aspired to authenticity, but I never got beyond verisimilitude.”

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
IQ research establishes framework of attributes and clusters
- 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]
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 à 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 martyrser.
Archival Quality / Large-Scale Digitization

At Present: Quality Standards
- Material centered
- One size fits all
- Vertical integration
- Process control
- Compromise is failure

Trends Forward: Acceptable Loss
- User centered
- Fitness for use
- Third party creators
- Acceptance testing
- 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**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Thick text</td>
<td>scanning</td>
<td>189 0.57%</td>
<td>70 0.19%</td>
<td>19 0.06%</td>
<td>144 0.81%</td>
<td>422</td>
</tr>
<tr>
<td>Broken text</td>
<td>scanning</td>
<td>518 1.57%</td>
<td>121 0.33%</td>
<td>76 0.26%</td>
<td>64 0.36%</td>
<td>779</td>
</tr>
<tr>
<td>Blurred text</td>
<td>scanning</td>
<td>252 0.76%</td>
<td>40 0.11%</td>
<td>10 0.03%</td>
<td>54 0.30%</td>
<td>356</td>
</tr>
<tr>
<td>Obscured text</td>
<td>source</td>
<td>57 0.17%</td>
<td>35 0.09%</td>
<td>21 0.07%</td>
<td>8 0.04%</td>
<td>121</td>
</tr>
<tr>
<td>Warped page</td>
<td>post-scan</td>
<td>47 0.14%</td>
<td>37 0.10%</td>
<td>14 0.05%</td>
<td>22 0.12%</td>
<td>120</td>
</tr>
<tr>
<td>Cropped text block</td>
<td>post-scan</td>
<td>424 1.28%</td>
<td>246 0.67%</td>
<td>100 0.34%</td>
<td>67 0.38%</td>
<td>837</td>
</tr>
<tr>
<td>Cleaning</td>
<td>post-scan</td>
<td>208 0.63%</td>
<td>214 0.58%</td>
<td>1256 4.23%</td>
<td>439 2.46%</td>
<td>2117</td>
</tr>
<tr>
<td>Colorization</td>
<td>post-scan</td>
<td>3250 9.83%</td>
<td>272 0.74%</td>
<td>35 0.12%</td>
<td>19 0.11%</td>
<td>3576</td>
</tr>
<tr>
<td>Volumes ingested</td>
<td></td>
<td>288,044</td>
<td>460,620</td>
<td>2,523,049</td>
<td>1,665,167</td>
<td>4,936,880</td>
</tr>
<tr>
<td>Volumes reviewed (20 pages/vol.)</td>
<td></td>
<td>33,047</td>
<td>36,981</td>
<td>29,677</td>
<td>17,850</td>
<td>117,555</td>
</tr>
<tr>
<td>Ingested/Received</td>
<td></td>
<td>11.47%</td>
<td>8.03%</td>
<td>1.18%</td>
<td>1.07%</td>
<td>2.38%</td>
</tr>
</tbody>
</table>

10 August 2010

Archival Quality in Digital Preservation
Two Examples [“... flattening & thickening of meaning...”]

Heather MacNeil

Warped Page

Thick Text
Errors in Source or Scanning

Source Crop

Scan Crop

The text is printed incorrectly and runs off the page. This is clearly a publisher crop and not a human error.
Errors in Source or Scanning

Source Blur

Scanning Blur
Fingers in Manual Scanning

Traces of human error

Traces digitally cleaned

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 [moire 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

- 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

Metrics

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)


References (2)

Thank you for your attention!

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