Assessing the Quality of Web Archives

Michael L. Nelson  
Old Dominion University, mnelson@odu.edu

Scott G. Ainsworth  
Old Dominion University

Justin F. Brunelle  
Old Dominion University

Mat Kelly  
Old Dominion University

Hany SalahEldeen  
Old Dominion University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons

Recommended Citation
Nelson, Michael L.; Ainsworth, Scott G.; Brunelle, Justin F.; Kelly, Mat; SalahEldeen, Hany; and Weigle, Michele C., "Assessing the Quality of Web Archives" (2014). Computer Science Presentations. 8.
https://digitalcommons.odu.edu/computerscience_presentations/8

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
Assessing the Quality of Web Archives

Michael L. Nelson
Scott G. Ainsworth, Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle

Old Dominion University
Web Science & Digital Libraries Research Group
ws-dl.cs.odu.edu
@WebSciDL
The State of Web Archiving

current:  "Hooray! It's in the archive!"

vs.

future:  "How well was it archived?"
Summaries of the shooting of Igor
17.07.2014 17:50 (MSK) Message from the militia: “In the area Toroz just downed plane An-26, lying somewhere in the mine” Progress”. Warned same - not to fly in “our sky.” And here is the video, the confirmation of the next “picklepads.” Birdie fail for waste heap, the residential sector is not caught. Peaceful people do not suffer. And also have information about the second downed aircraft, like the Su. “

Summaries of the shooting of Igor

Digital Preservation, July 22-23, 2014,
Washington DC

Three Ways We're Assessing Quality

- Weighting the "importance" of missing embedded resources
  - "damage" measure for comparing archived pages
- Detecting "temporal violations"
  - some rendered pages never existed
- Defining an archival tool benchmark
  - "Archive Acid Test"
Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources
JCDL 2014

Synthetic Damage: Removing Images From xkcd.com

\[ M = 0.17 \]
\[ D = 0.09 \]
(live web)

\[ M = 0.24 \]
\[ D = 0.41 \]
(missing main)

\[ M = 0.29 \]
\[ D = 0.36 \]
(missing logo + navigation)

*damage (D) differs from % missing (M)!*
Was missing resource important? `<img>` and `<embed>` can leave hints about size and centrality.

For CSS, we look at the distribution of background color in page divided into vertical thirds.
Weights from Turker Assessment of Damage

first: establish that Turkers can determine damaged vs. undamaged pages (81% of the time)

second: find weights that match Turker's rankings of (real) differently damaged versions of the same page
Good News:
Although M is steady/increasing, D is decreasing
A Framework for Evaluation of Composite Memento Temporal Coherence (in preparation)

http://arxiv.org/abs/1402.0928
As Presented by IA


(now 404, but that's a different story...)

Not Everything Is
200412091900926

+ 9 months

- 15 hours

+9 hours

Digital Preservation, July 22-23, 2014,
Washington DC

(now 404, but that's a different story...)

http://www.wunderground.com/cgi-bin/findWeather/getForecast?query=50593
Consider:

```html
<html>
<img src="foo.jpeg">
</html>
```
Correct Archival Rendering
But Archives Miss Updates…
You Can Choose the Closest

(closest is the current policy of most archives)
You Can Choose the Past

[Diagram showing changes in formats: jpeg, jpeg, html, jpeg]
Or You Can "Bracket" the HTML

(when possible, brackets can be made via HTTP metadata or content comparison)

In this case, there is no right answer.

*Either choice will result in a temporal violation.*
## Completeness vs. Coherence

<table>
<thead>
<tr>
<th>Description</th>
<th>Closest Single Archive</th>
<th>Closest Multi-Archive</th>
<th>Bracket Single Archive</th>
<th>Bracket Multi-Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean complete</td>
<td>76.1%</td>
<td>80.2%</td>
<td>76.2%</td>
<td>80.3%</td>
</tr>
<tr>
<td>Mean missing</td>
<td>23.9%</td>
<td>19.8%</td>
<td>23.8%</td>
<td>19.7%</td>
</tr>
<tr>
<td>Temporal Coherence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean prima facie coherent</td>
<td>41.0%</td>
<td>40.9%</td>
<td>54.7%</td>
<td>54.6%</td>
</tr>
<tr>
<td>Mean possibly coherent</td>
<td>27.3%</td>
<td>27.3%</td>
<td>12.8%</td>
<td>14.2%</td>
</tr>
<tr>
<td>Mean probably violative</td>
<td>2.5%</td>
<td>5.3%</td>
<td>2.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Mean prima facie violative</td>
<td>5.3%</td>
<td>5.3%</td>
<td>6.2%</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

At least 5% of pages can be shown to be temporal violations
The Archival Acid Test:
Evaluating Archive Performance on Advanced HTML and JavaScript
JCDL 2014


http://acid.matkelly.com/
Inspired by the Acid3 Test for Browsers

Acid3

100/100

To pass the test, a browser must use its default settings, the animation has to be smooth, the score has to end on 100/100, and the final page has to look exactly, pixel for pixel, like this reference rendering.

http://acid3.acidtests.org/
http://en.wikipedia.org/wiki/Acid3
The Archival Acid Test

Archiving Tools

Heritrix

GNU Wget

WARCreate

Archives

Internet Archive

WayBack Machine

WebCite

perma.cc

archive.today
Archival Tools & Sites on Acid3

To pass the test, a browser must use its default settings, the animation has to be smooth, the score has to end on 100/100, and the final page has to look exactly, pixel for pixel, like the reference rendering.

Acid3

JS ?

Acid3

64/100

Acid3

JS ?

Acid3

87/100

Acid3

82/100

Acid3

85/100

Archival Acid Tests

The Basics (6 tests)

Javascript (8 tests)

Advanced Features Tests (4 tests)

chrome
Archival Tools & Sites on AAT

(interrupted by a note: mummify.it died in early 2014)

The Basics (6 tests)
Javascript (8 tests)
Advanced Features Tests (4 tests)

The Basics (6 tests)
Javascript (8 tests)
Advanced Features Tests (4 tests)

The Basics (6 tests)
Javascript (8 tests)
Advanced Features Tests (4 tests)

The Basics (6 tests)
Javascript (8 tests)
Advanced Features Tests (4 tests)

Future of Web Archiving: Increasing Quantitative Analysis

• Measure "damage" instead of completeness of archived pages
  – enables large-scale comparison of archives
• Even if an embedded resource is present, it doesn't mean it's right
  – ~5% of archived pages have temporal violations
• To improve the quality of the archives, we need to be able to benchmark archival tools
  – Archival Acid Test is an easy to use benchmark