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

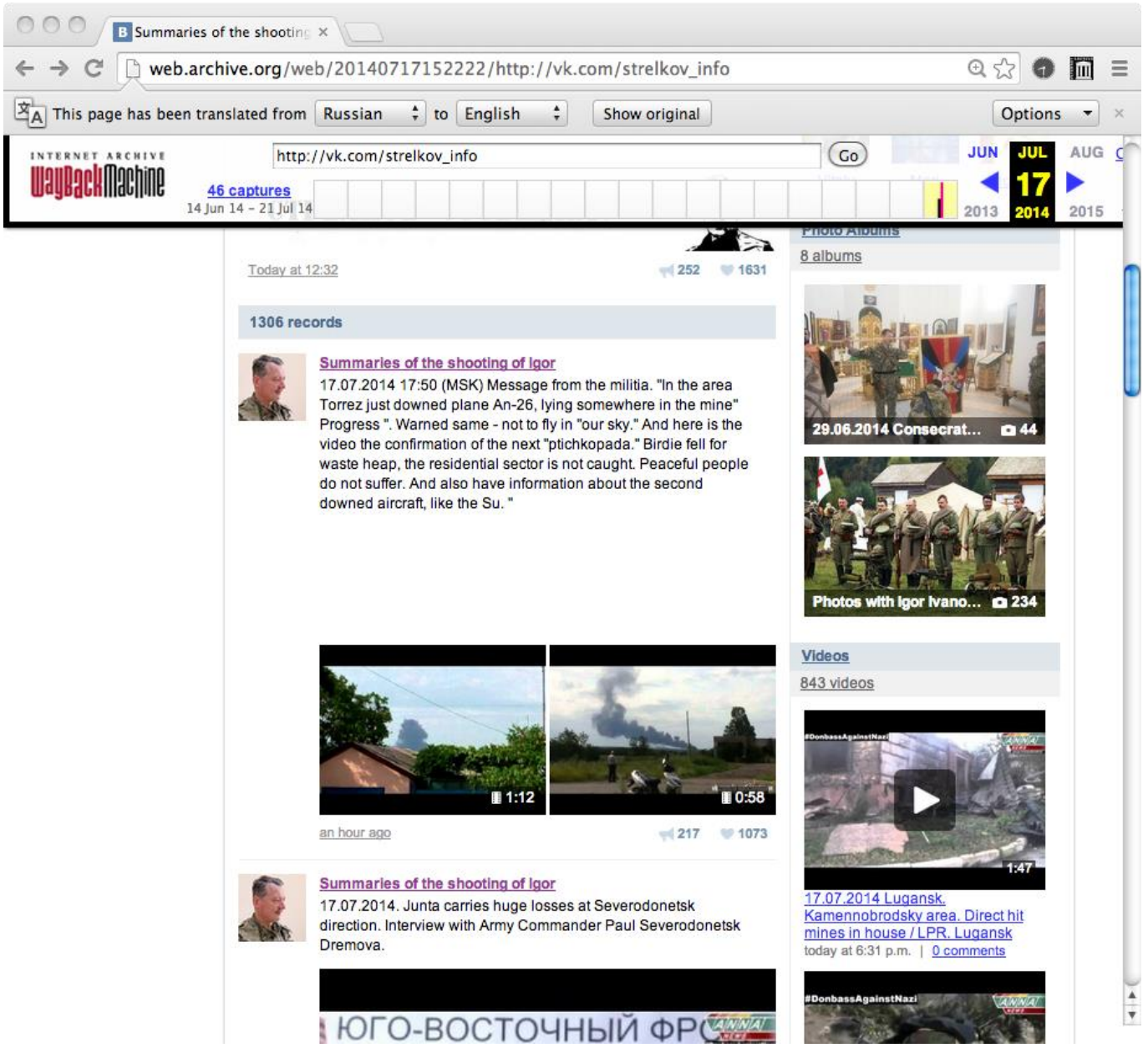
Digital Preservation, July 22-23, 2014,
Washington DC

The State of Web Archiving

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

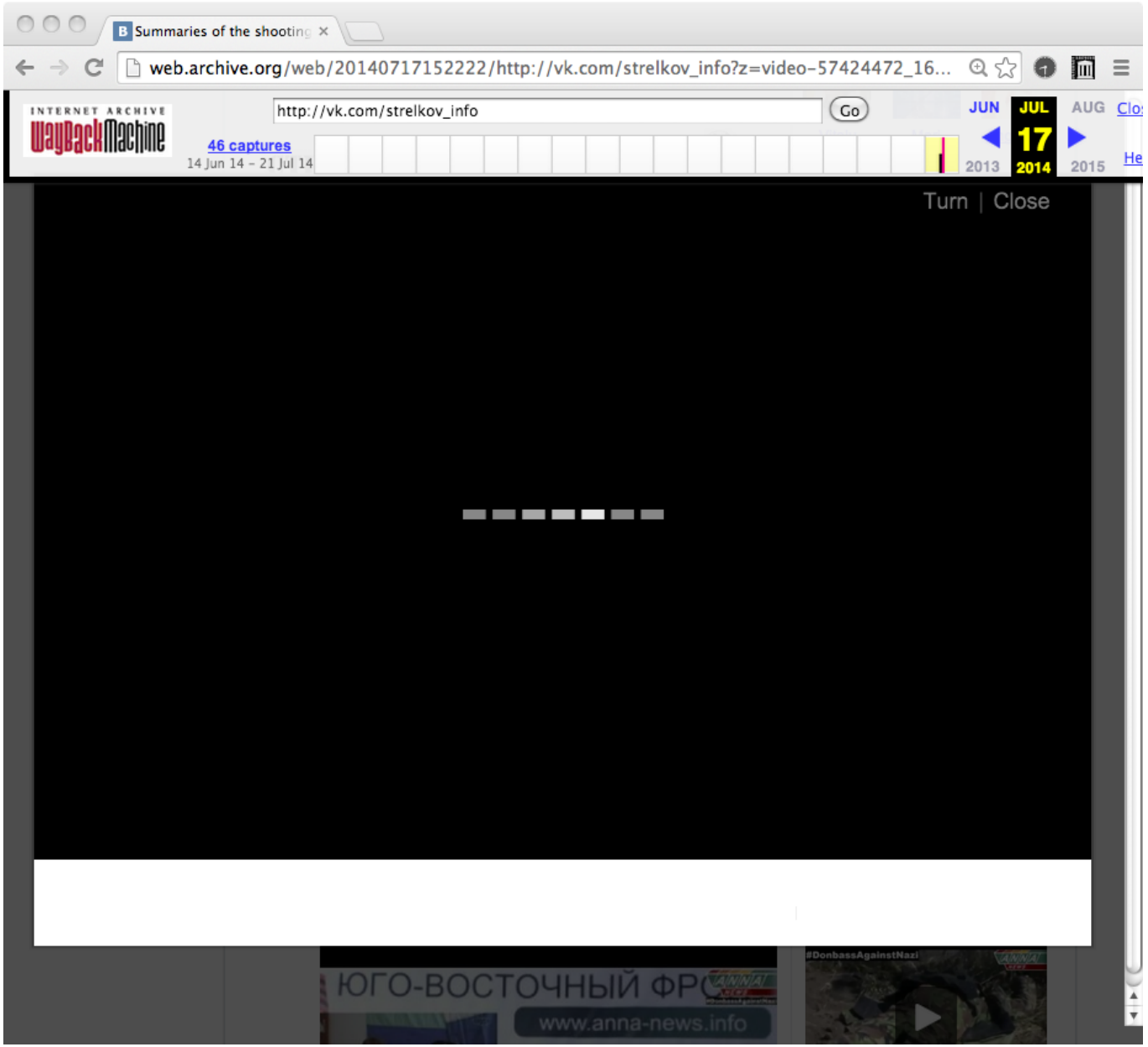
vs.

future: "How well was it archived?"



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

http://web.archive.org/web/20140717152222/http://vk.com/strelkov_info
http://www.csmonitor.com/World/Europe/2014/0717/Web-evidence-points-to-pro-Russia-rebels-in-downing-of-MH17-video



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

http://web.archive.org/web/20140717152222/http://vk.com/strelkov_info
http://www.csmonitor.com/World/Europe/2014/0717/Web-evidence-points-to-pro-Russia-rebels-in-downing-of-MH17-video

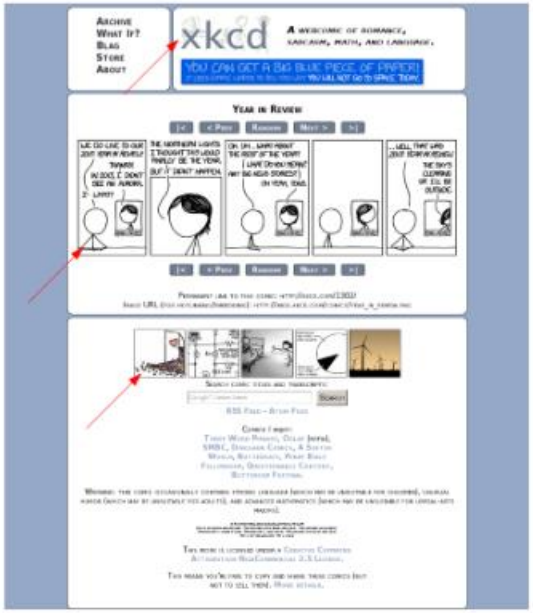
Three Ways We're Assessing Quality

- 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

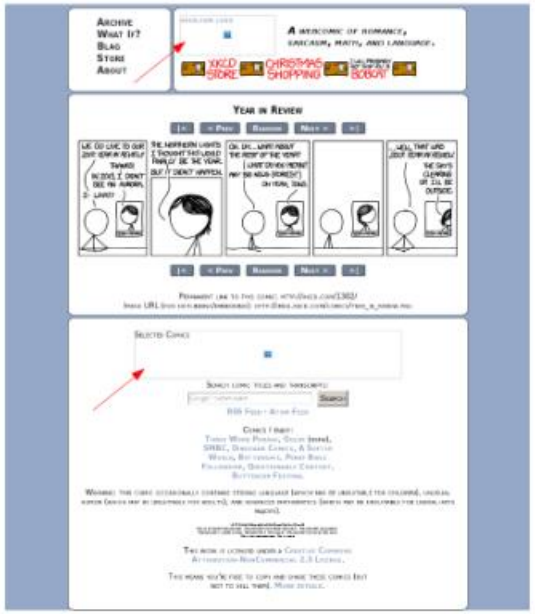
http://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf

Synthetic Damage: Removing Images From xkcd.com



M = 0.17
D = 0.09
(live web)

M = 0.24
D = 0.41
(missing main)

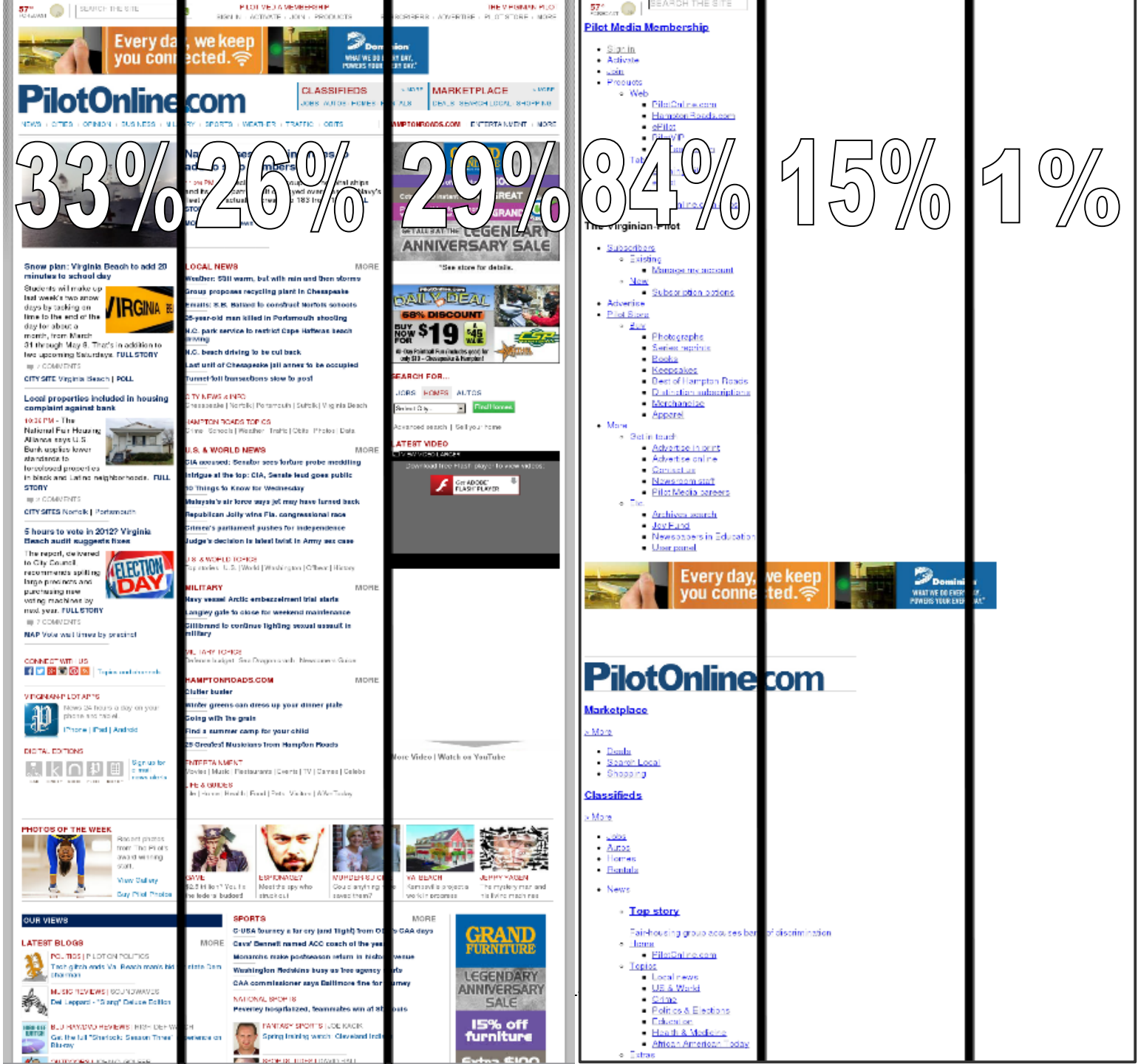


M = 0.29
D = 0.36
(missing logo + navigation)

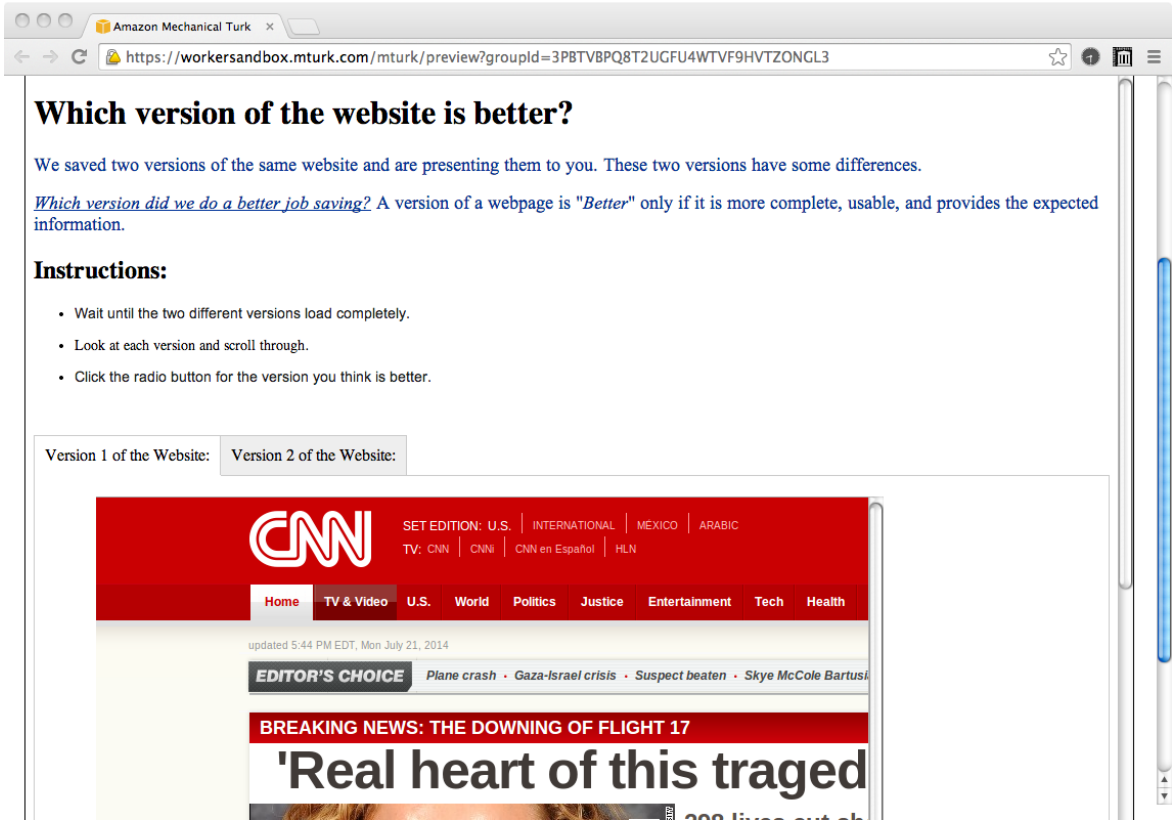
damage (D) differs from % missing (M)!

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.



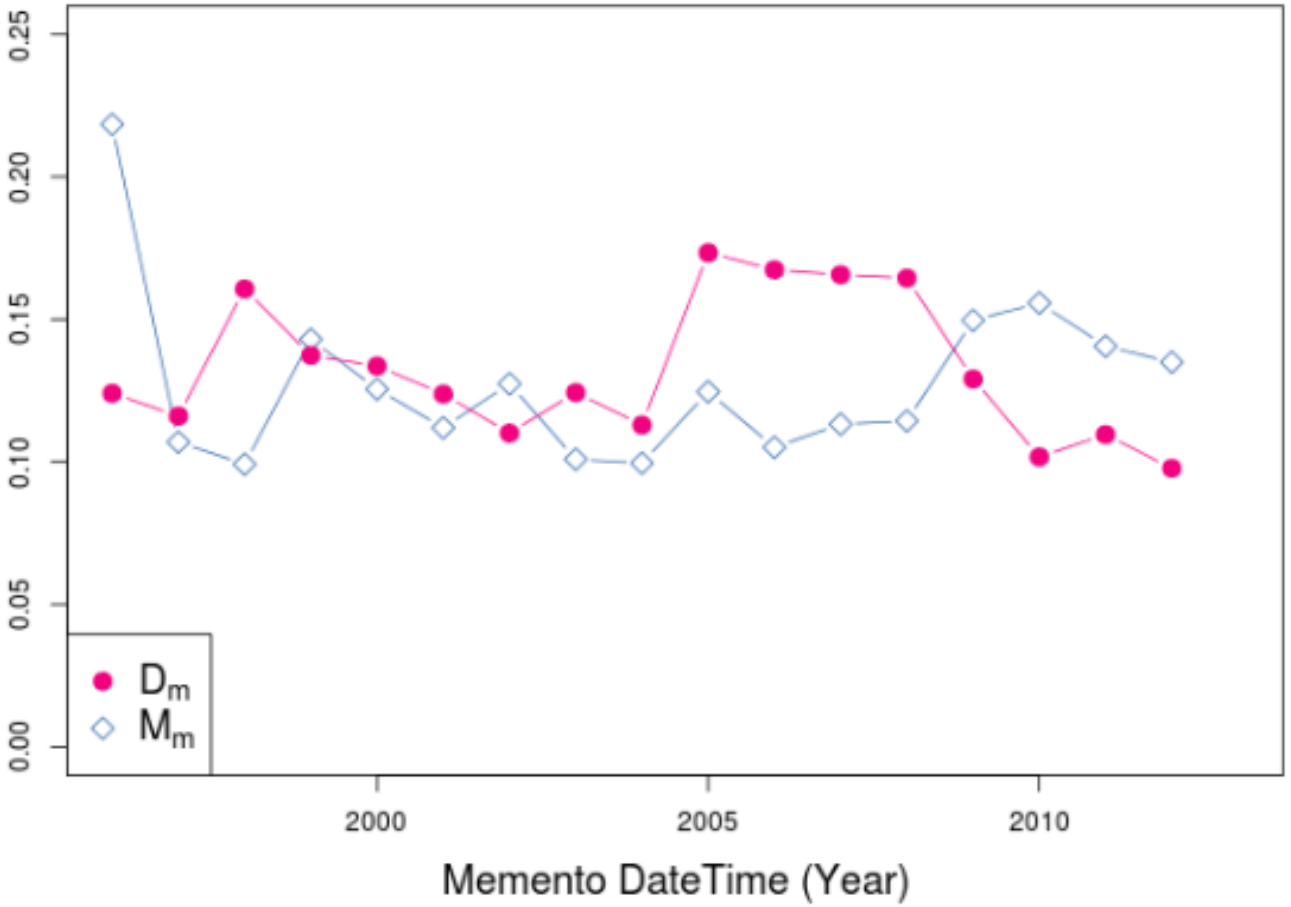
Weights from Turker Assessment of Damage



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

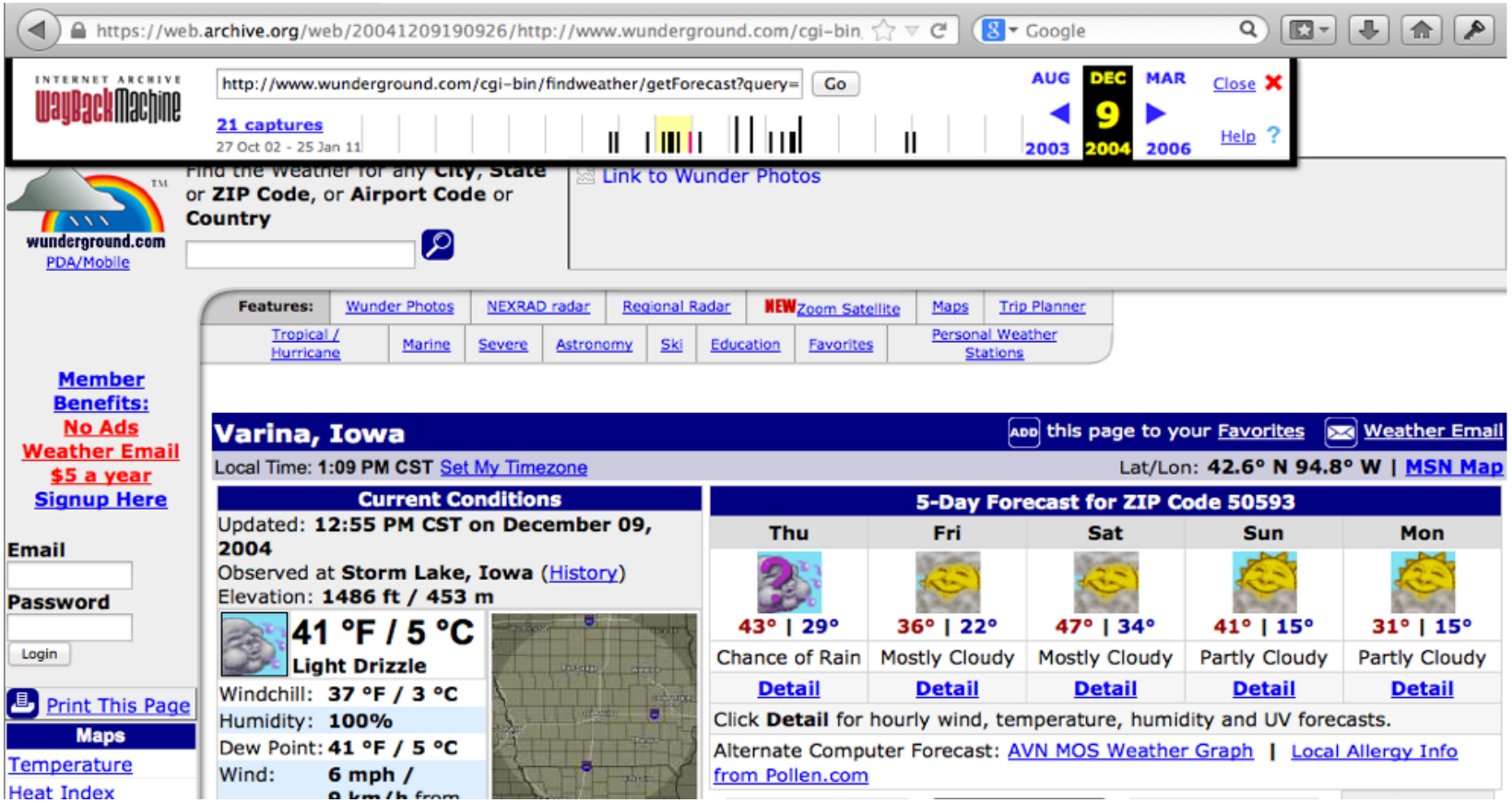
Good News:
Although M is steady/increasing, D is decreasing



A Framework for Evaluation of Composite Memento Temporal Coherence (in preparation)

http://arxiv.org/abs/1402.0928

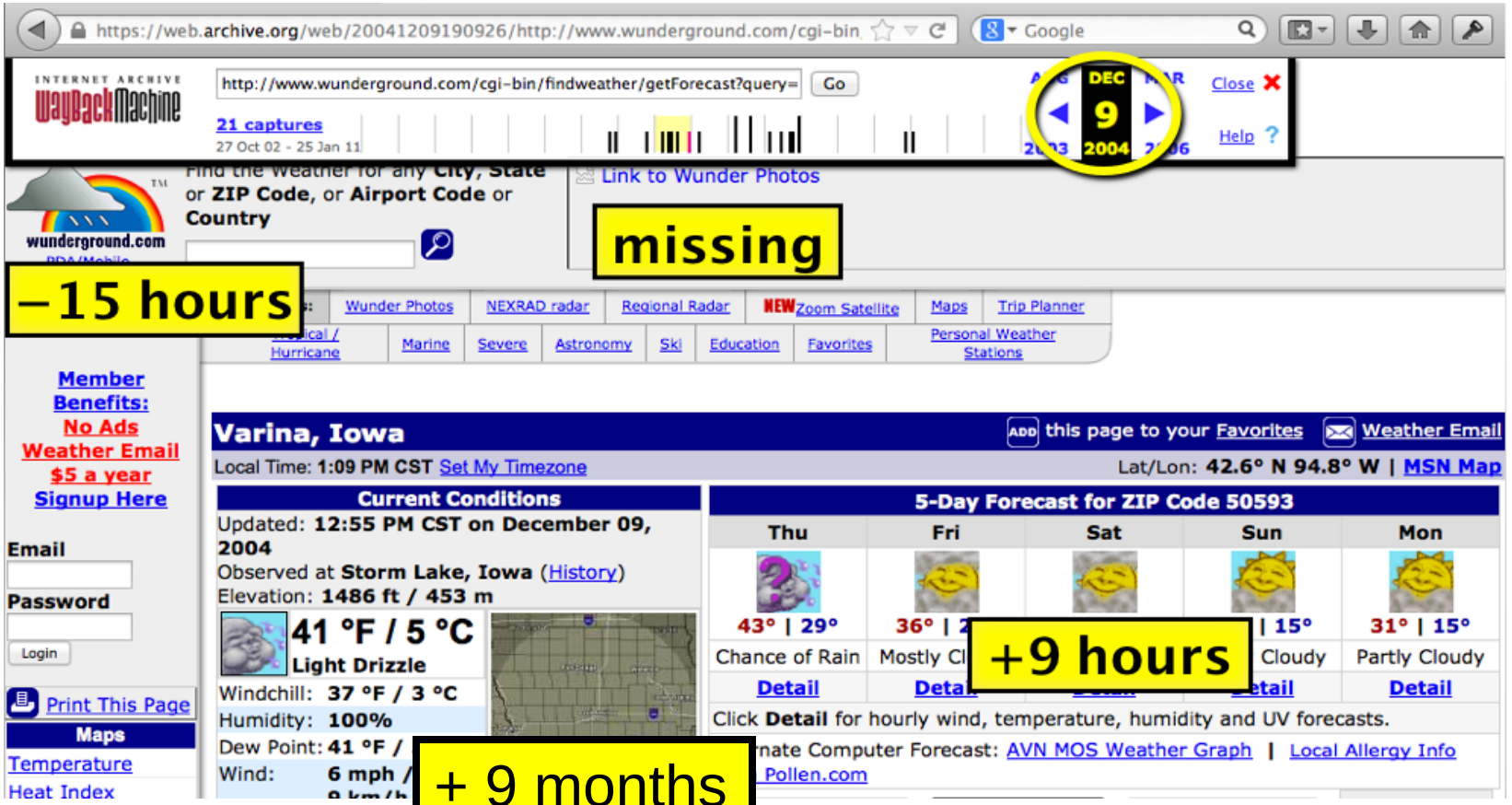
As Presented by IA



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

http://web.archive.org/web/20041209190926/http://www.wunderground.org/cgi-bin/findWeather/getForecast?query=50593 (now 404, but that's a different story…)

Not Everything Is 20041209190926

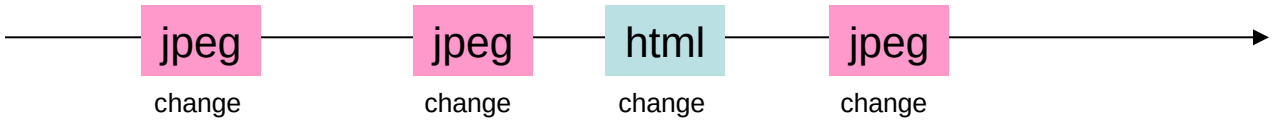


Digital Preservation, July 22-23, 2014,
Washington DC

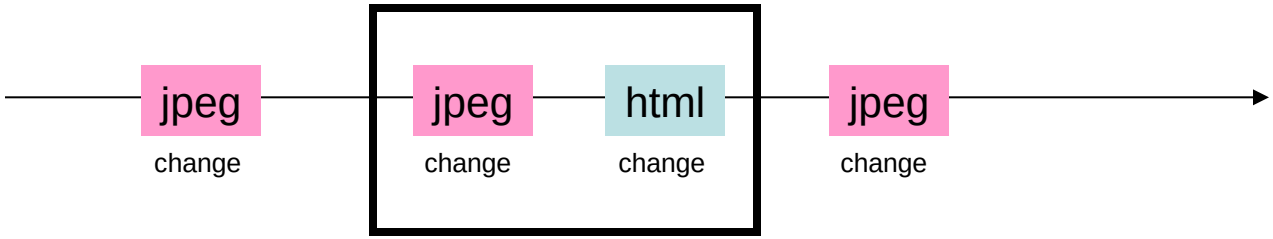
http://web.archive.org/web/20041209190926/http://www.wunderground.org/cgi-bin/findWeather/getForecast?query=50593 (now 404, but that's a different story…)

Consider:

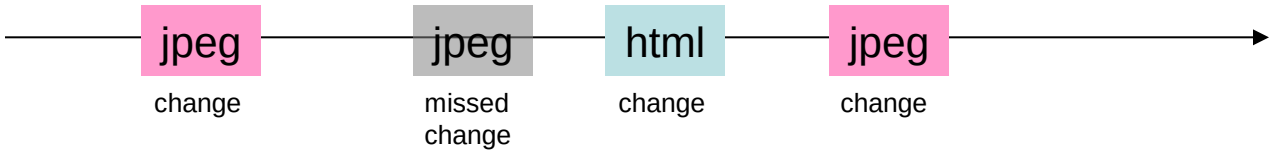
```
<html>
<img src="foo.jpeg">
</html>
```



Correct Archival Rendering

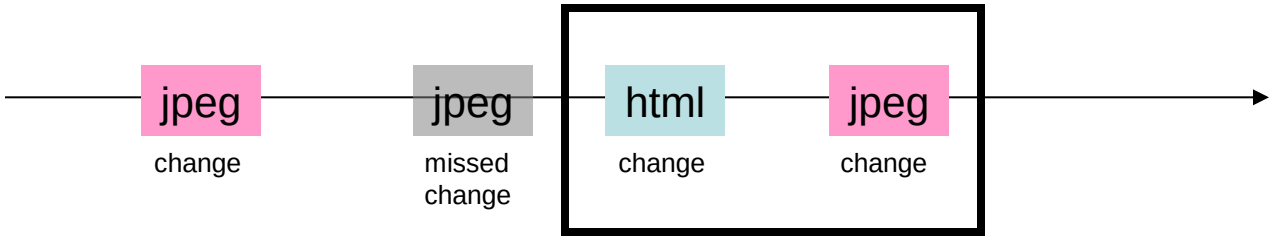


But Archives Miss Updates…

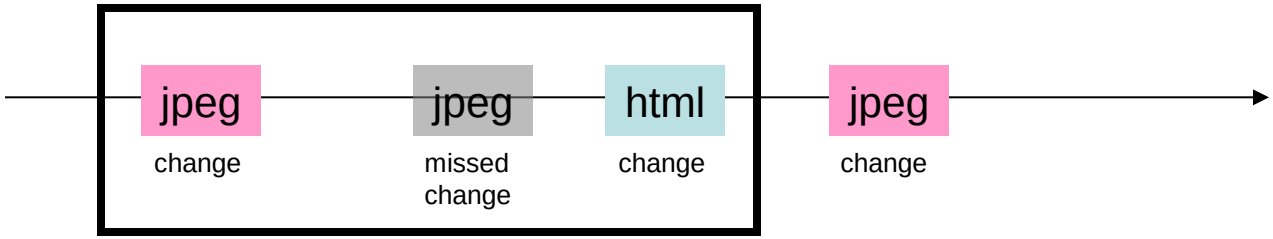


You Can Choose the Closest

(closest is the current policy of most archives)

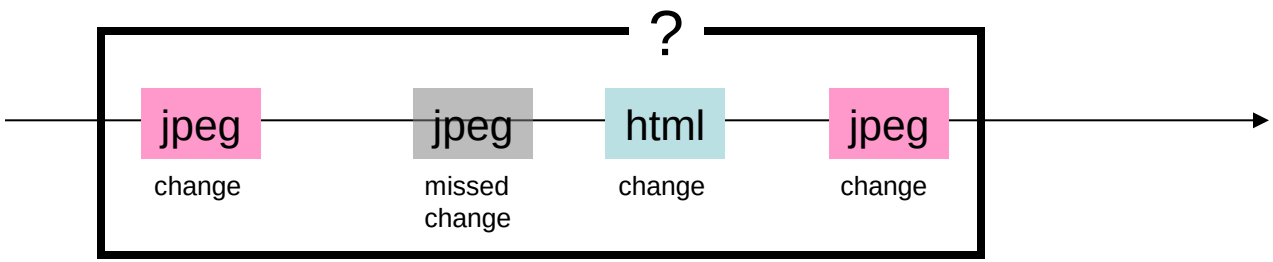


You Can Choose the Past



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

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

At least 5% of pages can be shown to be temporal violations

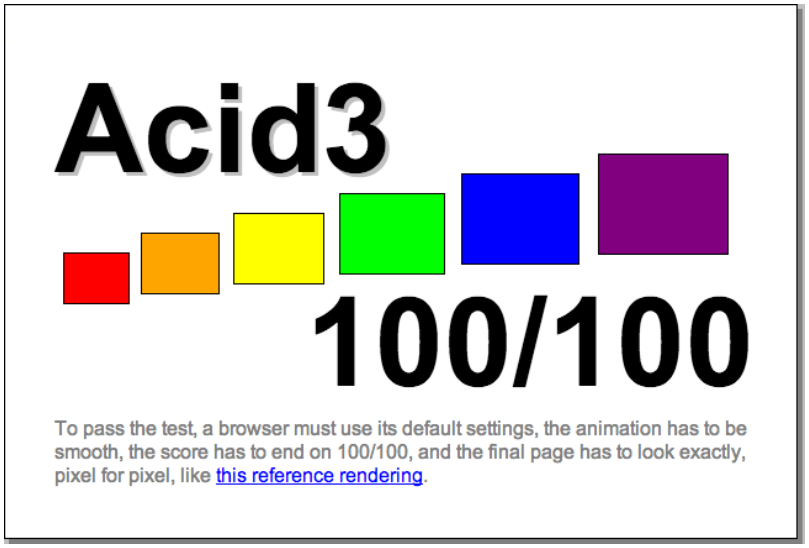
The Archival Acid Test: Evaluating Archive Performance on Advanced HTML and JavaScript JCDL 2014

http://ws-dl.blogspot.com/2014/07/2014-07-14-archival-acid-test.html

http://acid.matkelly.com/

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

Inspired by the Acid3 Test for Browsers



chrome

http://acid3.acidtests.org/
http://en.wikipedia.org/wiki/Acid3

The Archival Acid Test

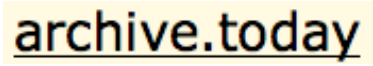
Archiving Tools

Heritrix

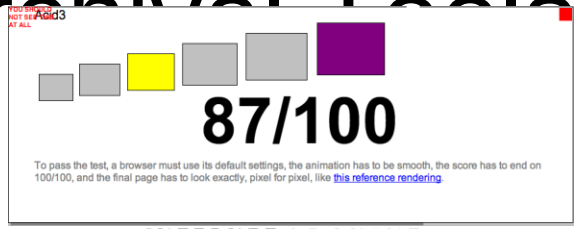
GNU Wget

WARCreate

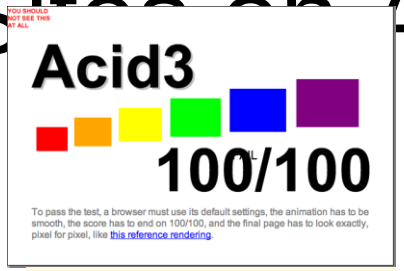
Archives



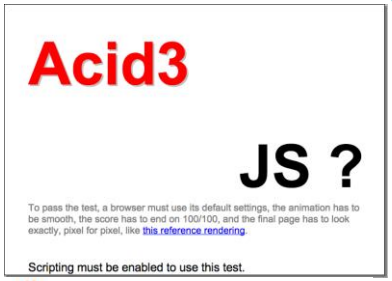
Archival Tools & Sites on Acid3



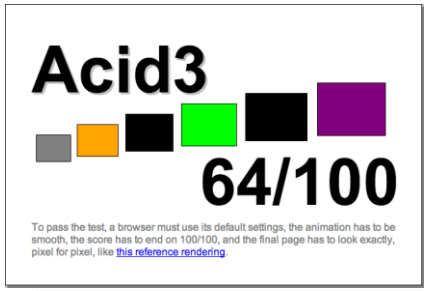
INTERNET ARCHIVE
WayBackMachine
Save Page Now



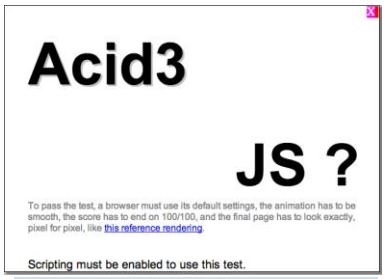
archive.today



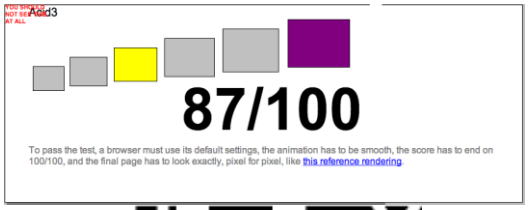
MUMMIFY



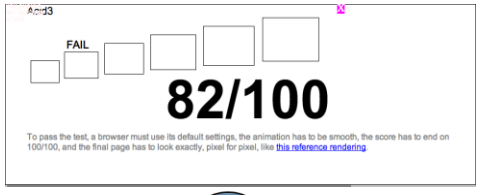
perma.cc∞

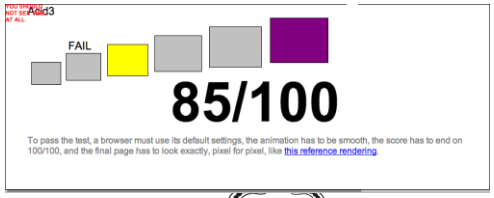


WebCite



HERITRIX





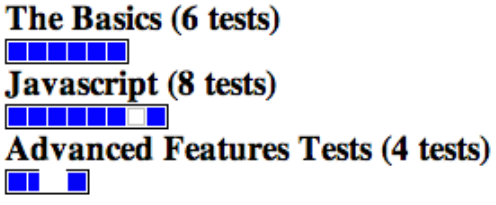
Archival Acid Tests

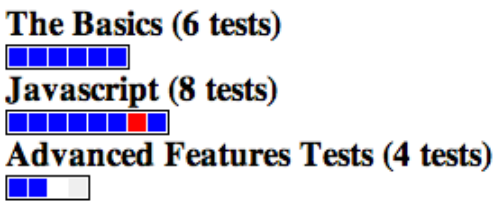
The Basics (6 tests)

Javascript (8 tests)

Advanced Features Tests (4 tests)

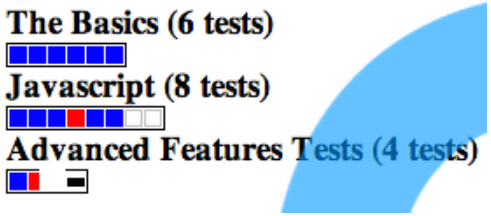
Archival Tools & Sites on AAT



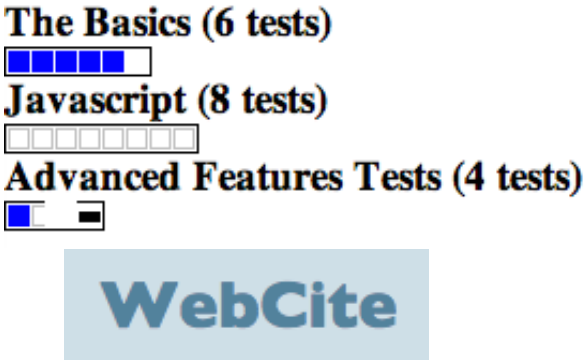


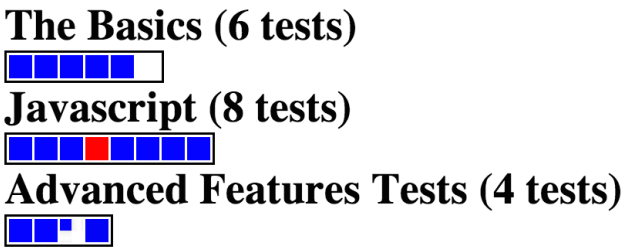
(mummify.it died in early 2014)

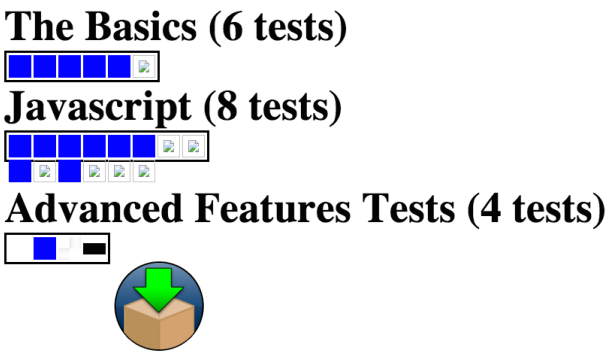
archive.today

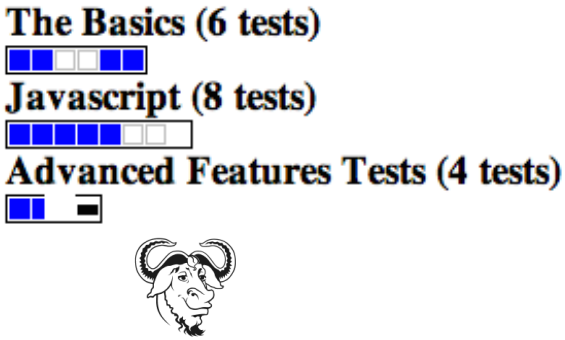


perma.cc∞









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