

# 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](http://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?"

Today at 12:32 252 1631

1306 records



**Summaries of the shooting of Igor**  
 17.07.2014 17:50 (MSK) Message from the militia. "In the area Torrez 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 "ptichkopada." Birdie fell 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."



an hour ago 217 1073



**Summaries of the shooting of Igor**  
 17.07.2014. Junta carries huge losses at Severodonetsk direction. Interview with Army Commander Paul Severodonetsk Dremova.

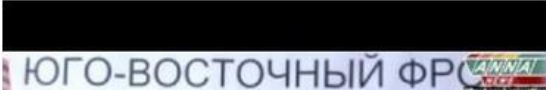


Photo Albums

8 albums

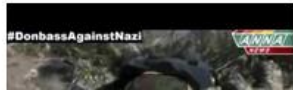


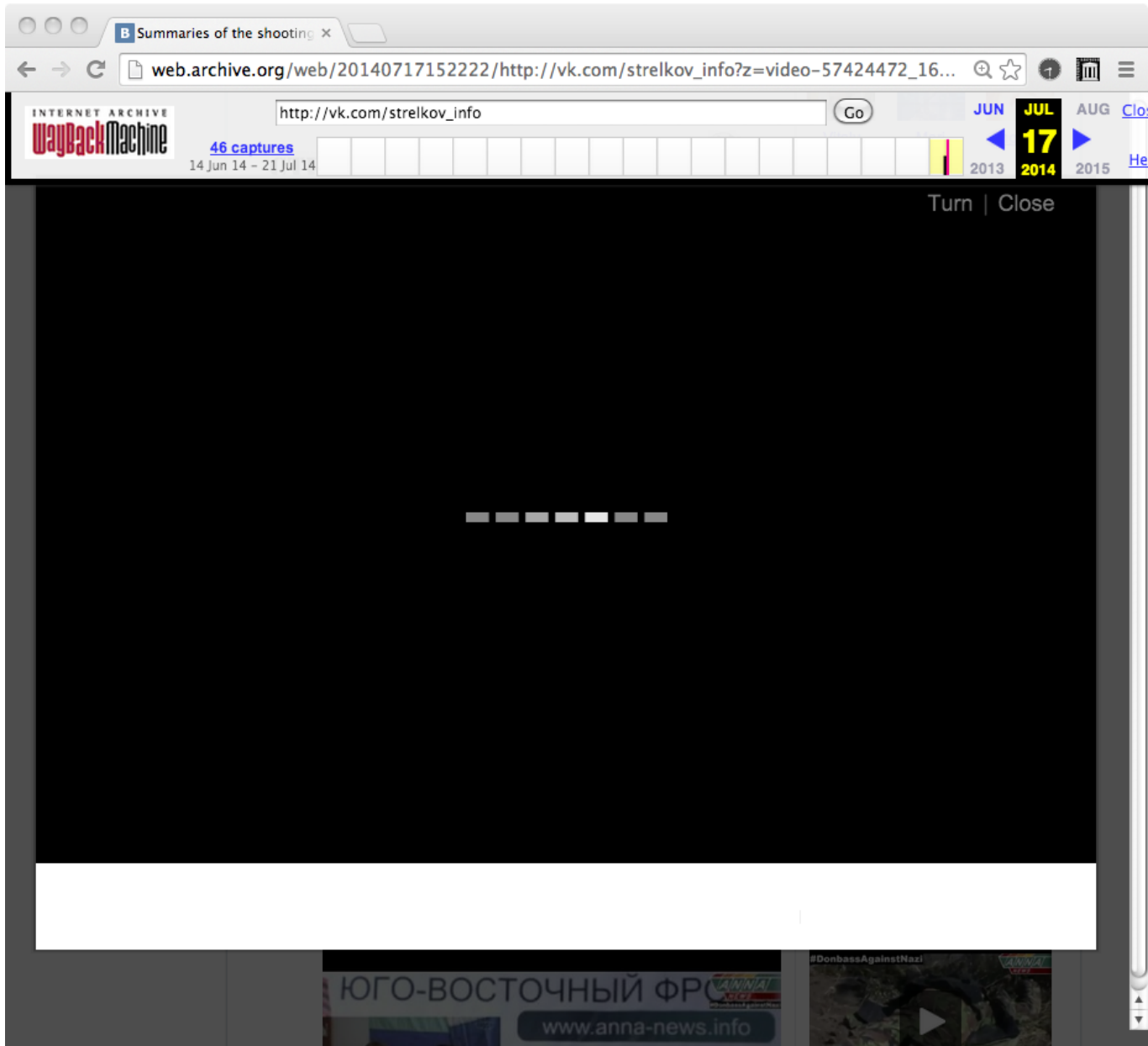
Videos

843 videos



17.07.2014 Lugansk. Kamennobrodsky area. Direct hit mines in house / LPR, Lugansk today at 6:31 p.m. | 0 comments





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

[http://web.archive.org/web/20140717152222/http://vk.com/strelkov\\_info](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>

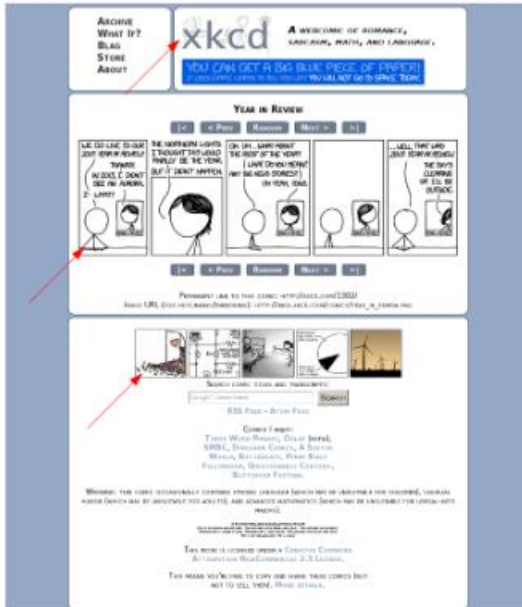
# Three Ways We're Assessing Quality

- 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>

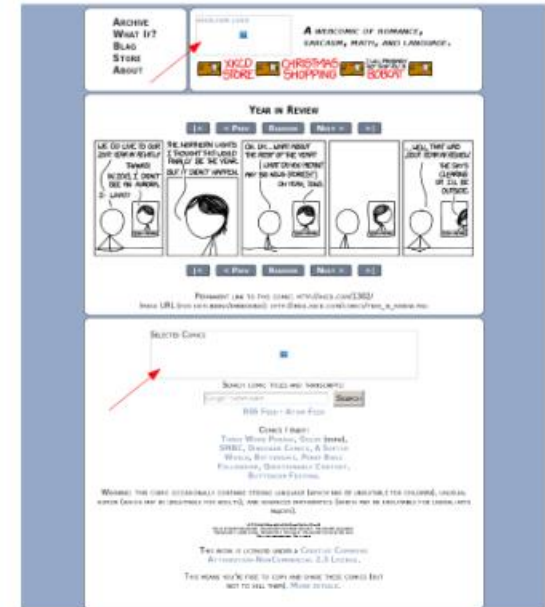
# Synthetic Damage: Removing Images From xkcd.com



$M = 0.17$   
 $D = 0.09$   
 (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.

33% 26% 29%

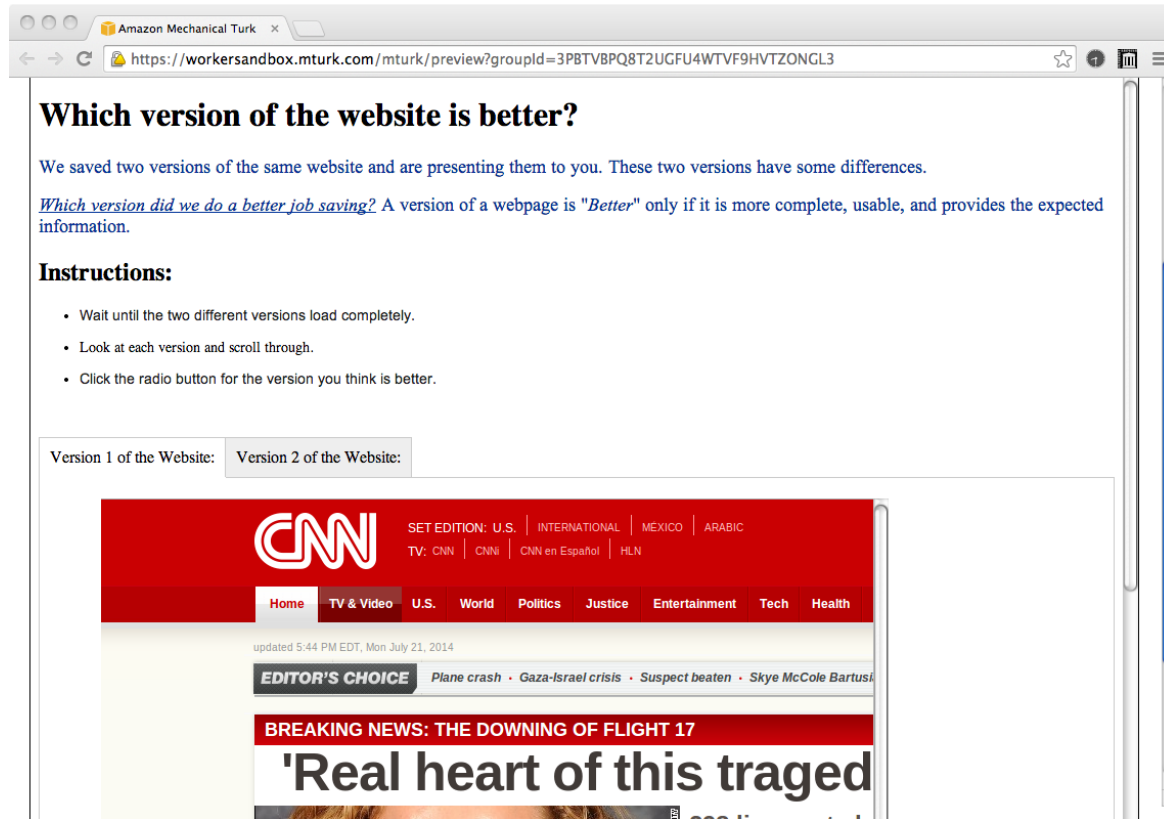
The screenshot shows the top portion of the PilotOnline.com website. The top navigation bar includes links for 'Pilot Media Membership', 'SEARCH THE SITE', and 'PilotMedia.com'. Below this is a banner for 'Every day, we keep you connected.' with a 'Domain' logo. The main content area is divided into several sections: 'LOCAL NEWS' (Snow plan: Virginia Beach to add 20 minutes to school day), 'CLASSIFIEDS', 'MARKETPLACE', and 'ENTERTAINMENT'. Large white percentage overlays are placed over the top third of the page content.

84% 15% 1%

The screenshot shows the bottom portion of the PilotOnline.com website. It features a 'PilotMedia Membership' section with a list of links including 'Local news', 'US & World', 'Politics & Elections', 'Calendar', 'Health & Medicine', and 'Black & American Today'. Below this is a 'Marketplace' section with links for 'Books', 'Home & Local', and 'Shopping'. A 'Classifieds' section is also visible. Large white percentage overlays are placed over the bottom third of the page content.



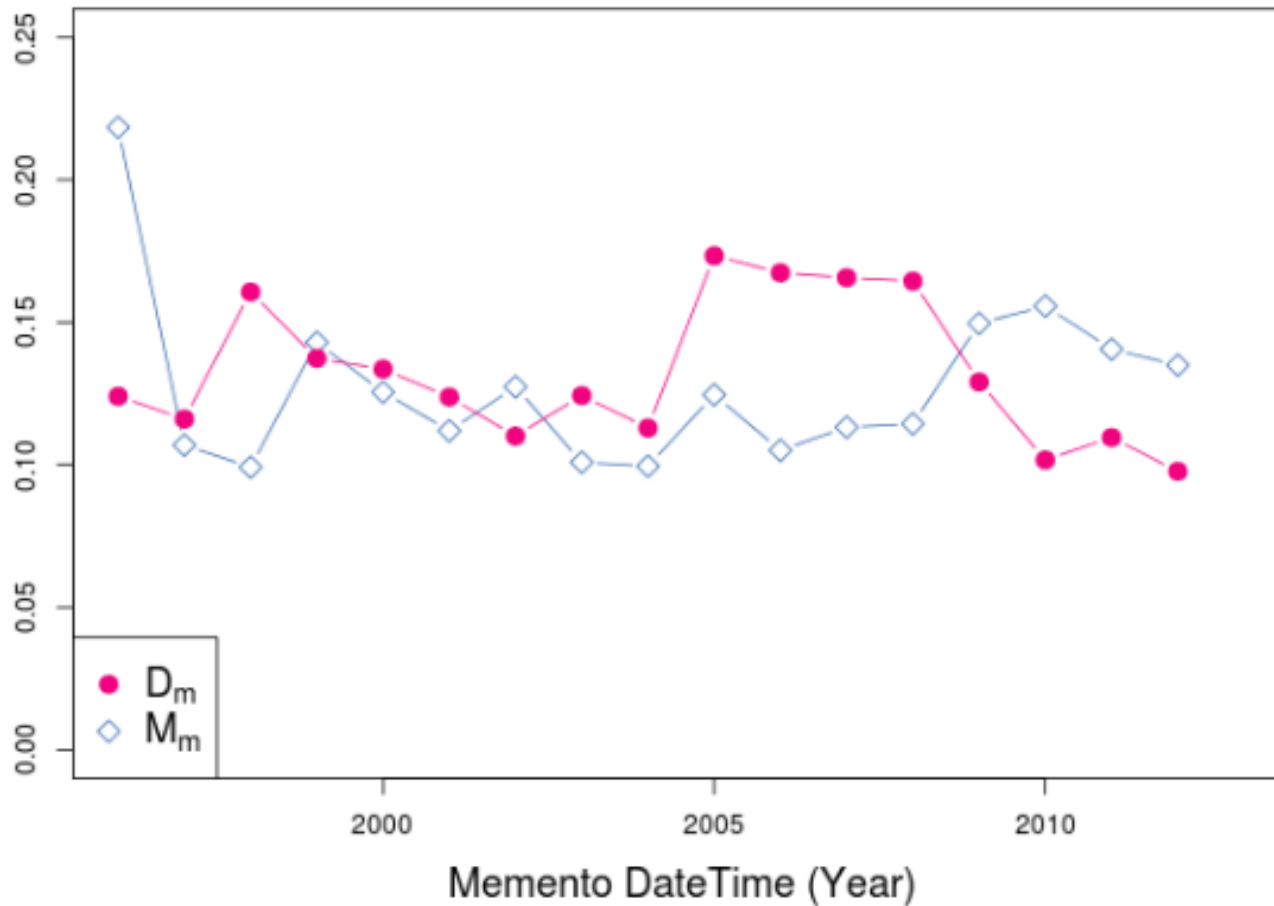
# Weights from Turker Assessment of Damage



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

Internet Archive WayBack Machine

https://web.archive.org/web/20041209190926/http://www.wunderground.com/cgi-bin/... Google

http://www.wunderground.com/cgi-bin/findweather/getForecast?query= Go

21 captures 27 Oct 02 - 25 Jan 11

AUG DEC MAR Close X  
2003 2004 2006 Help ?

Find the weather for any City, State or ZIP Code, or Airport Code or Country

Link to Wunder Photos

wunderground.com PDA/Mobile

Features: Wunder Photos NEXRAD radar Regional Radar NEW Zoom Satellite Maps Trip Planner

Tropical / Hurricane Marine Severe Astronomy Ski Education Favorites Personal Weather Stations

**Member Benefits:**  
No Ads  
Weather Email  
\$5 a year  
Signup Here

Email  
Password  
Login

Print This Page  
Maps  
Temperature  
Heat Index

## Varina, Iowa


Local Time: 1:09 PM CST Set My Timezone Lat/Lon: 42.6° N 94.8° W | MSN Map

**Current Conditions**






Updated: 12:55 PM CST on December 09, 2004  
Observed at Storm Lake, Iowa (History)  
Elevation: 1486 ft / 453 m

**41 °F / 5 °C**  
Light Drizzle

Windchill: 37 °F / 3 °C  
Humidity: 100%  
Dew Point: 41 °F / 5 °C  
Wind: 6 mph / 9 km/h from



**5-Day Forecast for ZIP Code 50593**

Thu	Fri	Sat	Sun	Mon
 43°   29°	 36°   22°	 47°   34°	 41°   15°	 31°   15°
Chance of Rain	Mostly Cloudy	Mostly Cloudy	Partly Cloudy	Partly Cloudy
<a href="#">Detail</a>	<a href="#">Detail</a>	<a href="#">Detail</a>	<a href="#">Detail</a>	<a href="#">Detail</a>

Click **Detail** for hourly wind, temperature, humidity and UV forecasts.  
Alternate Computer Forecast: [AVN MOS Weather Graph](#) | [Local Allergy Info from Pollen.com](#)

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

# Not Everything Is 200412091900926

Internet Archive Wayback Machine  
21 captures  
27 Oct 02 - 25 Jan 11

Find the weather for any City, State or ZIP Code, or Airport Code or Country

**missing**

**-15 hours**

Member Benefits:  
No Ads  
Weather Email  
\$5 a year  
Signup Here

Email  
Password  
Login

Print This Page  
Maps  
Temperature  
Heat Index

**Varina, Iowa**  
Local Time: 1:09 PM CST Set My Timezone  
Lat/Lon: 42.6° N 94.8° W | MSN Map

**Current Conditions**  
Updated: 12:55 PM CST on December 09, 2004  
Observed at Storm Lake, Iowa (History)  
Elevation: 1486 ft / 453 m  
41°F / 5°C  
Light Drizzle  
Windchill: 37°F / 3°C  
Humidity: 100%  
Dew Point: 41°F / 5°C  
Wind: 6 mph / 9 km/h

**5-Day Forecast for ZIP Code 50593**

Thu	Fri	Sat	Sun	Mon
43°   29°	36°   21°	36°   21°	15°	31°   15°
Chance of Rain	Mostly Cloudy	Mostly Cloudy	Cloudy	Partly Cloudy
<a href="#">Detail</a>	<a href="#">Detail</a>	<a href="#">Detail</a>	<a href="#">Detail</a>	<a href="#">Detail</a>

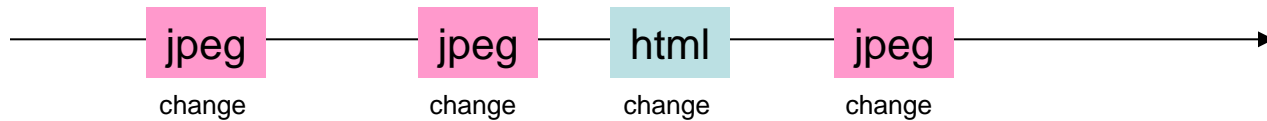
**+9 hours**

**+9 months**

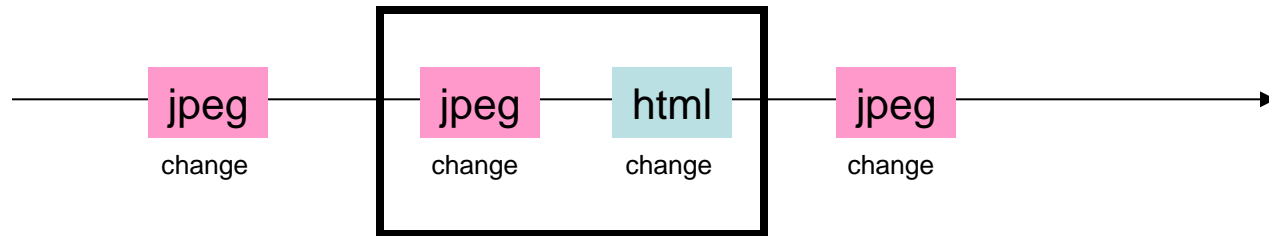
Digital Preservation, July 22-23, 2014,  
Washington DC

# Consider:

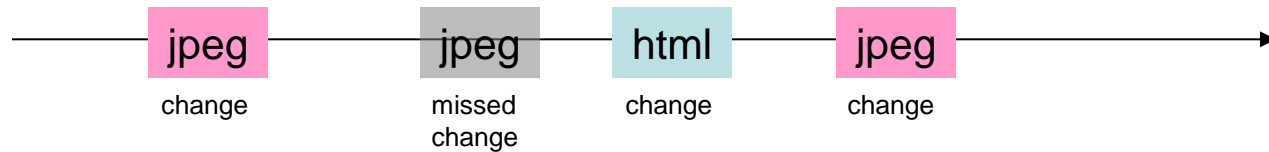
```
<html>  
  
</html>
```



# Correct Archival Rendering



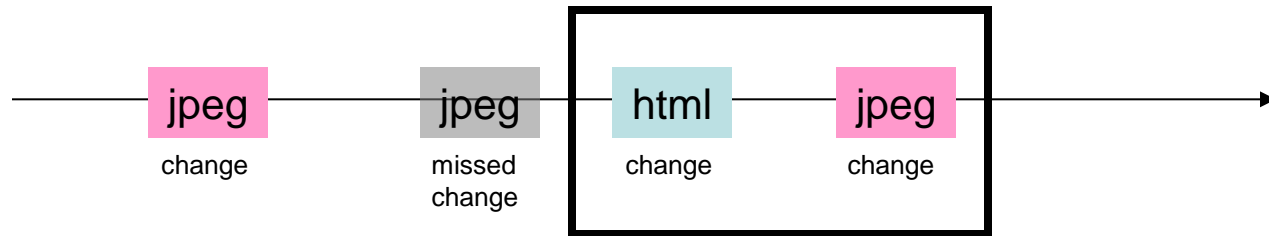
# But Archives Miss Updates...



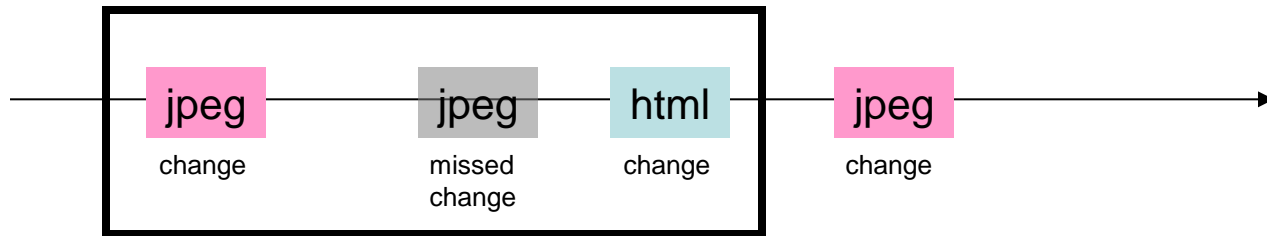


# You Can Choose the Closest

(closest is the current policy of most archives)

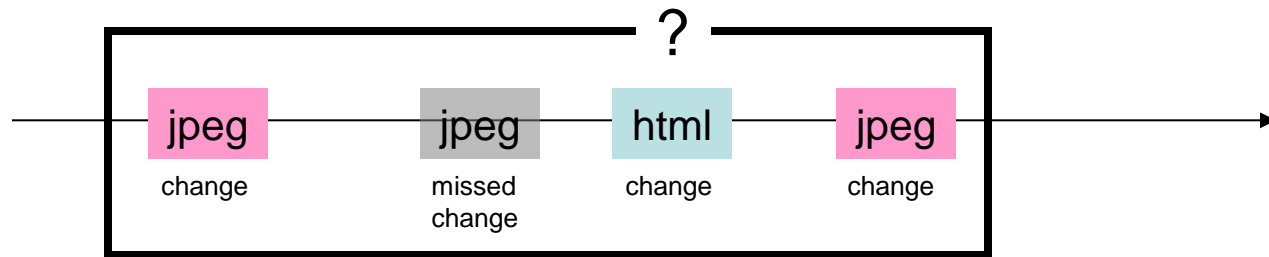


# You Can Choose the Past



# 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
<b>Completeness</b>				
Mean complete	76.1%	80.2%	76.2%	80.3%
Mean missing	23.9%	19.8%	23.8%	19.7%
<b>Temporal Coherence</b>				
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%

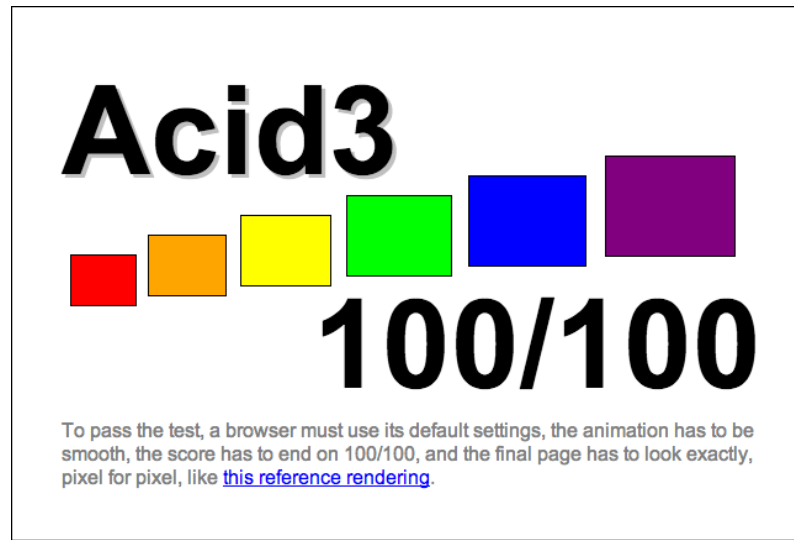
*At least 5% of pages can be shown to be temporal violations*

# 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/>

# Inspired by the Acid3 Test for Browsers



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

# The Archival Acid Test

## *Archiving Tools*

Heritrix



GNU Wget



WARCreate



## *Archives*



WebCite




perma.cc∞

archive.today

# Archival Tools & Sites on Acid3

YOU SHOULD NOT SEE THIS AT ALL




**87/100**

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](#).

INTERNET ARCHIVE



YOU SHOULD NOT SEE THIS AT ALL



**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](#).

archive.today

**Acid3**


**JS ?**

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](#).

Scripting must be enabled to use this test.



**Acid3**



**64/100**

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](#).



**Acid3**


**JS ?**

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](#).

Scripting must be enabled to use this test.



YOU SHOULD NOT SEE THIS AT ALL




**87/100**

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](#).



Acid3

FAIL




**82/100**

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](#).



YOU SHOULD NOT SEE THIS AT ALL

FAIL



**85/100**

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](#).





# Archival Acid Tests

**The Basics (6 tests)**



**Javascript (8 tests)**



**Advanced Features Tests (4 tests)**



# Archival Tools & Sites on AAT

The Basics (6 tests)



Javascript (8 tests)



Advanced Features Tests (4 tests)



INTERNET ARCHIVE



The Basics (6 tests)



Javascript (8 tests)



Advanced Features Tests (4 tests)



archive.today

(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)



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