# Only One Out of Five Archived Web Pages Existed as Presented

Scott G. Ainsworth, Michael L. Nelson, Herbert Van de Sompel (Proceedings of Hypertext 2015)

CS895 - Web Archiving Forensics, Fall 2020
Old Dominion University
Department of Computer Science
Kritika Garg, kgarg001@odu.edu

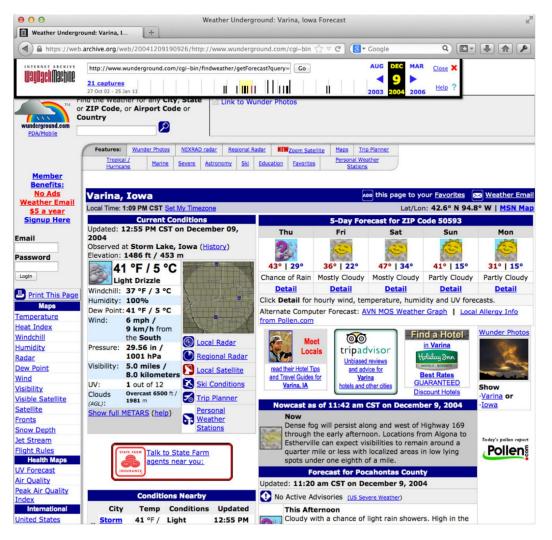


Figure 1: Wunderground
Composite Memento,
MementoDatetime 2004-12-09
19:09:26
https://citeseerx.ist.psu.ed
u/viewdoc/download?doi=
10.1.1.711.77&rep=rep1&t
vpe=pdf

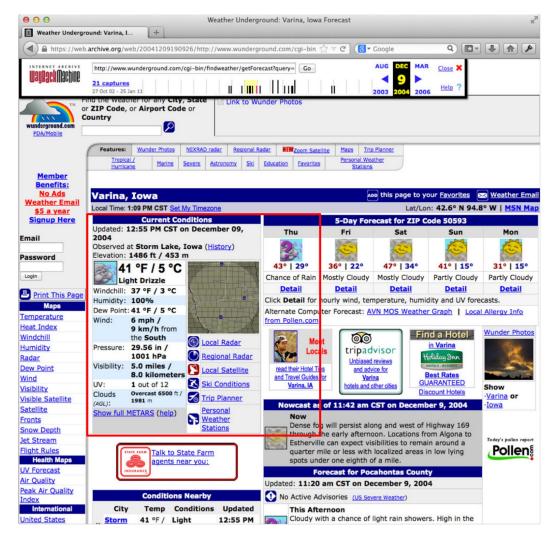
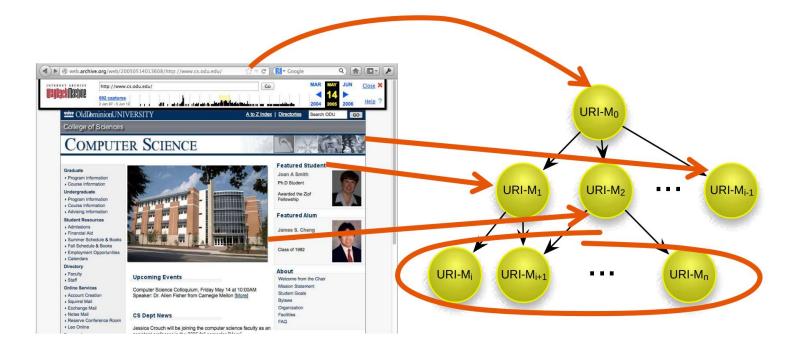


Figure 1: Wunderground
Composite Memento,
MementoDatetime 2004-12-09
19:09:26
<a href="https://citeseerx.ist.psu.ed/">https://citeseerx.ist.psu.ed/</a>
u/viewdoc/download?doi=
10.1.1.711.77&rep=rep1&t
vpe=pdf

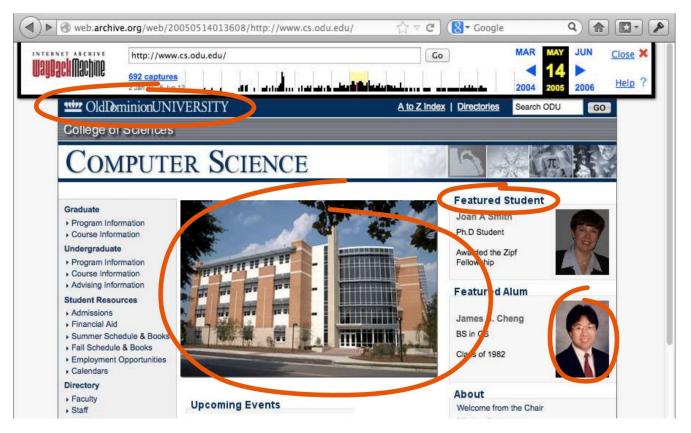


Figure 2: Wunderground
Composite Memento,
MementoDatetime 2004-12-09
19:09:26
https://citeseerx.ist.psu.ed
u/viewdoc/download?doi=
10.1.1.711.77&rep=rep1&t
ype=pdf

#### **Composite Memento**

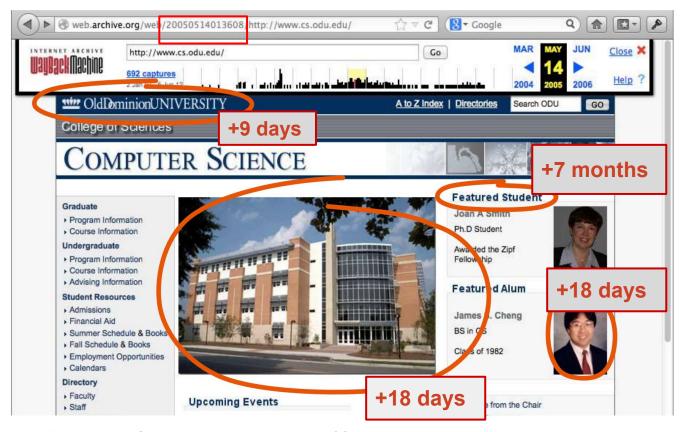


## **Temporal Coherence**



Slide 20, <a href="https://www.slideshare.net/ScottAinsworth/only-one-out-of-five-archived-web-pages-existed-as-presented">https://www.slideshare.net/ScottAinsworth/only-one-out-of-five-archived-web-pages-existed-as-presented</a>

## **Temporal Coherence**



Slide 20, <a href="https://www.slideshare.net/ScottAinsworth/only-one-out-of-five-archived-web-pages-existed-as-presented">https://www.slideshare.net/ScottAinsworth/only-one-out-of-five-archived-web-pages-existed-as-presented</a>

## Completeness

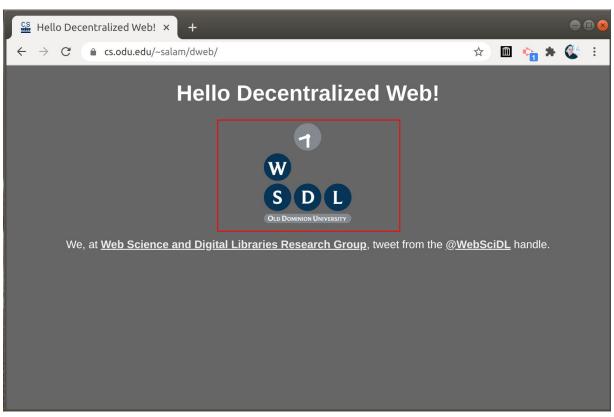




Figure 2: Wunderground
Composite Memento,
MementoDatetime 2004-12-09
19:09:26
https://citeseerx.ist.psu.ed
u/viewdoc/download?doi=
10.1.1.711.77&rep=rep1&t
ype=pdf

**Table 3: Embedded Memento Capture Datetimes (15 of 128)** 

URI	Memento-Datetime	Delta
http://ads.wunderground.com/ads/images/wu-9.jpg	2004-10-16 03:40:53Z	-1.8 months
http://icons.wunderground.com/graphics/smalllogo2.gif	2004-11-22 05:46:03Z	-18 days
http://ads.wunderground.com/ads/images/Statefarm-sfcom001_120x30.gif	2004-11-22 05:46:32Z	-18 days
http://icons.wunderground.com/ads/images/Davi-00009-vp2_125x125.gif	2004-12-08 07:01:08Z	-2 days
http://icons.wunderground.com/graphics/smash/wunderTransparent.gif	2004-12-09 04:36:14Z	-15 hours
http://www.wunderground.com/cgi-bin/findweather/getForecast?query=50593	2004-12-09 19:29:26Z	- root
http://icons.wunderground.com/graphics/conds/cloudy.GIF	2004-12-10 04:48:55Z	+21 minutes
http://icons.wunderground.com/graphics/conds/mostlycloudy.GIF	2004-12-10 04:48:55Z	+10 hours
http://icons.wunderground.com/graphics/conds/rait.GIF	2004-12-12 14:54:01Z	+3 days
http://icons.wunderground.com/ads/images/TripAdvisor-Blinky.gif	2005-01-27 02:58:30Z	+1.6 months
http://banners.wunderground.com/cgi-bin/statefarmbanner?zip=50593&width=150	2006-03-26 03:29:00Z	+1.3 years
http://www.valueclick.com/system/files/coupon-mountain-slider.png?1310511429	2011-07-13 09:08:00Z	+6.6 years
http://z1.adserver.com/system/files/logo.gif	2013-01-13 06:09:14Z	+8.1 years
http://icons.wunderground.com/data/wximagenew/d/d70dave/0-thumb.jpg	Not Archived	
http://pagead2.googlesyndication.com/pagead/show_ads.js	Missing Memento	

**Temporal Spread** = 8.3 years Mean delta = 1.8 years standard deviation = 2.9 years

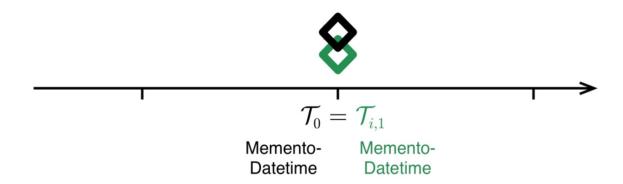
- Embedded resource captured within the seconds of root, may still not be temporally coherent
- Embedded resource captured much after than root, may not be incoherent.

## **Temporal Coherence States**

- Prima Facie Coherent
- Possibly Coherent
- Probably Violative
- Prima Facie Violative

#### **Prima Facie Coherent**

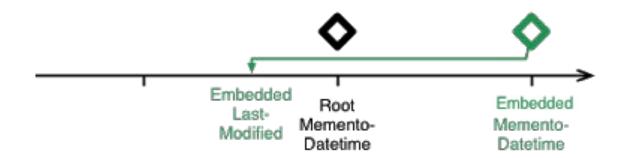
An embedded memento existed in its archived state at the time the root memento was captured.



```
<html>
<img src="foo.jpeg">!
</html>
                       HTTP/1.1 200 OK
                       Server: Tengine/2.0.3
                       Date: Mon, 27 Apr 2015 22:03:32 GMT
                       Content-Type: image/jpeg
                       Content-Length: 15632
                       Connection: keep-alive
                       Memento-Datetime: Tue, 07 Feb 2006 00:58:23 GMT
                       Link: <Memento links deleted...>
                       X-Archive-Orig-server: Apache/1.3.26 (Unix)
                       ApacheJServ/1.1.2 PHP/4.3.4
                       X-Archive-Orig-etag: "4978-3d10-3e4d822e"
                       X-Archive-Orig-content-length: 15632
                       X-Archive-Orig-accept-ranges: bytes
                       X-Archive-Orig-date: Tue, 07 Feb 2006 00:58:20 GMT
                       X-Archive-Orig-content-type: image/jpeg
                       X-Archive-Orig-last-modified: \varphi
                             Fri, 14 Feb 2003 23:56:30 GMT
                       X-Archive-Orig-connection: close
                       <other headers deleted>
```

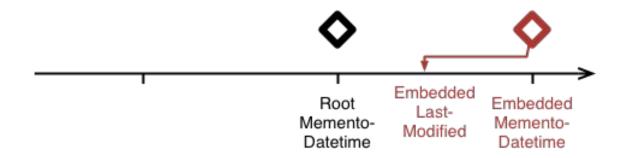
#### **Prima Facie Coherent**

Bracket Pattern



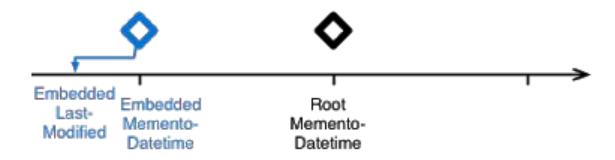
#### **Prima Facie Violative**

An embedded memento that did not exist in its archived state at the time the root memento was captured.



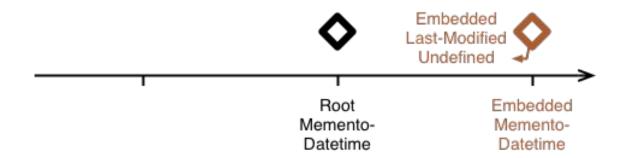
## **Possibly Coherent**

An embedded memento might have been in its archived state at the time the root memento was captured.



## **Probably Violative**

An embedded memento might not have existed in its archived state at the time the root memento was captured.



# **Representing Coherence**

Table 1: Sample Mementos	from http:	//www.rpgdreamers.com/
		/ ······

URI-R	Last-Modified	Memento-Datetime	Delta	Root?	Bracket?	Coherence	Color
http://www.rpgdreamers.com/	N/A	2006-11-10 04:37:51	5.00	Y		C	Black
/aslmain.js	2006-04-25 14:03:39	2006-11-10 23:44:02	_	N	Y	C	Green
/home_exp_leader_728x90_468x60.js	2006-11-11 00:30:18	2006-11-11 01:02:19	19.9h	N	N	V	Red
http://www.rpgdreamers.com/	N/A	2006-12-06 08:04:42	1	Y	-	C	Black
/aslmain.js	2006-04-25 14:03:39	2006-12-07 01:07:32	_	N	Y	C	Green
/home_exp_leader_728x90_468x60.js	2006-12-05 17:54:31	2006-12-05 18:30:00	13.6h	N	N	PC	Blue
/show_ads.js	N/A	2006-12-06 12:00:00	3.9h	N	N	PV	Yellow
http://www.rpgdreamers.com/	N/A	2007-01-01 21:31:16	===	Y	_	С	Black
/aslmain.js	2006-04-25 14:03:39	2007-01-02 16:23:01	-	N	Y	C	Green
/home_exp_leader_728x90_468x60.js	2006-12-09 14:41:27	2006-12-09 15:04:45	23.3d	N	N	PC	Blue
/show_ads.js	N/A	2007-01-05 12:00:00	86.5d	N	N	PV	Yellow

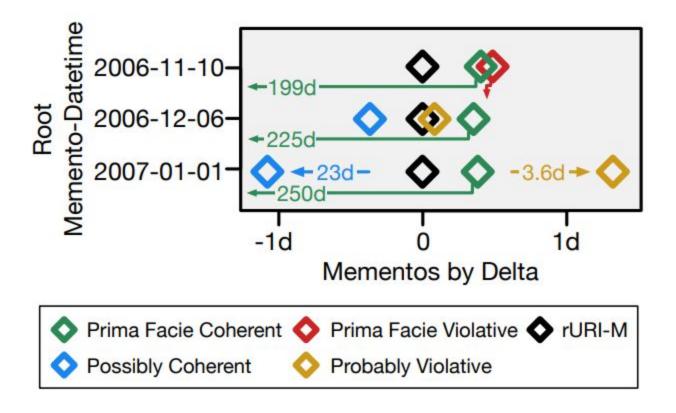


Figure 5: Scatter Plot Sample

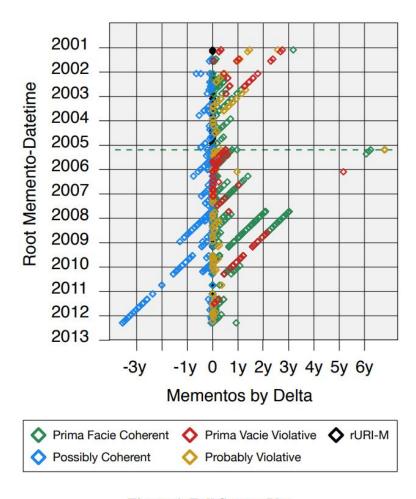


Figure 6: Full Scatter Plot <a href="https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.711.77&rep=rep1&type=pdf">https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.711.77&rep=rep1&type=pdf</a>

#### **Dataset**

**Table 3: Archival Rates** 

_	WOLU OF THE CIT	TI COL TECTOR	
Sample	This study	Jan'13 [1]	Jan'11 [3]
DMOZ	97.0%	95.2%	79%
Search Engine	55.2%	26.4%	19%
Bitly	28.3%	23.5%	16%
Delicious	95.0%	91.9%	68%
Aggregate	68.9%	59.4%	46%

- Single and Multiple Archives
- > Heuristics
  - Mindst: Selects the nearest capture of embedded URI
  - Bracket: (Last modified, Memento datetime)

#### Selection of rURIM's

- A. Retrieve timemaps for all 4,000 sample rURI-Rs from 15 archives
- B. Retrieve a single root memento per month

**Table 4: rURI-R Timemap Statistics** 

Desc.	<b>DMOZ</b>	S.Eng.	Bitly	Deli	Total
URI-Rs	1,000	1,000	1,000	1,000	4,000
<b>URI-Rs</b> Archived	971	552	282	951	2,756
in Multiple Archives	14.5%	11.1%	23.0%	57.5%	29.5%
Timemaps	1,154	682	404	2,167	4,252
per URI-R	1.19	1.14	1.43	2.28	1.58
Mementos	120,765	11,508	36,816	174,214	343,303
per URI-R	124.37	20.85	130.55	183.19	124.57

## Recomposition of composite mementos

- 1. Retrieve 88,054 rURIM-M using curl.
- 2. Extract eURI-Rs for the root mementos that received 200 response.
- 3. Retrieve the eURI-R timemap using curl.
- 4. Select eURI-M for eURI-R using heuristics.
- 5. Download embedded memento using curl
- 6. Steps 2–5 were repeated for other resources.

Table 6: eURI-M Retrieval Statistics							
	Mindist	Mindist	Bracket	Bracket			
Description	Single	Multi	Single	Multi			
<b>Memento Counts</b>							
eURI-Rs	1,620,597	1,623,354	1,619,805	1,623,127			
per rURI-M	19.7	19.7	19.7	19.7			
eURI-Ms available	1,250,641	1,330,858	1,252,125	1,332,993			
per rURI-M	14.2	15.1	14.2	15.1			
eURI-Ms not found	369,956	292,494	367,680	290,134			
<b>Not-Found Reasons</b>							
No timemaps	395,545	312,843	395,065	312,641			
404	41,588	46,330	40,428	44,852			
403	6,186	6,194	6,105	6,116			
503	6,092	5,806	5,697	5,442			
Other	2,970	3,746	2,810	3,508			

#### Results

**Table 7: Completeness and Temporal Coherence** 

	Mindist	Mindist	Bracket	Bracket
Description	Single	Multi	Single	Multi
Completeness				**
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%	28.7%	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%

#### Results

- > 38.7 % of web pages are temporally coherent
- > 17.9% (~1 in 5) temporally coherent mementos can be fully recomposed.

**Table 9: Composite Memento Coherence** Mindist Single Mindist Multi **Bracket Single** Bracket Multi Description Quantity Percent Quantity Percent Quantity Percent Quantity Percent All Root Mementos in Timemaps 88,054 100.0% 88,054 100.0% 88,054 100.0% 88,054 100.0% All Composite Mementos (complete and incomplete) All Coherence States 82,425 93.6% 82,425 93.6% 82,425 93.6% 82,425 93.6% Probably Violative through Prima Facie Coherent 62,749 71.3% 62,603 71.1% 60.525 68.7% 60,525 68.7% Possibly Coherent & Prima Facie Coherent 54,120 61.5% 48,289 54.8% 52,309 59,4% 52,309 59,4% Prima Facie Coherent only 23.147 26.3% 20,262 23.0% 34.074 38.7% 34.074 38.7% **Complete Composite Mementos Only** All Coherence States 29,200 33,2% 27,077 30.8% 30.8% 26,969 30.6% 27.077 Probably Violative through Prima Facie Coherent 23,284 26,4% 25,180 28,6% 22,678 25.8% 22,678 25.8% Possibly Coherent & Prima Facie Coherent 21.576 24.5% 21.976 25.0% 21.010 23.9% 21,010 23.9% Prima Facie Coherent only 10.944 12.4% 10,901 12.4% 15.750 17.9% 15,750 17.9%

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.711.77&rep=rep1&type=pdf

#### **Future Work**

- Develop a heuristic to locate reasonable substitute in timemap.
- Use Lexical signatures and Simhash to find similar resource.
- Interface indicating the temporal coherence of the composite mementos.