#### **Old Dominion University ODU Digital Commons**

**Computer Science Presentations** 

Computer Science

11-10-2010

#### Memento: Time Travel for the Web

Herbert Van de Sompel

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

Robert Sanderson

Lyudmila Balakireva

Scott Ainsworth 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

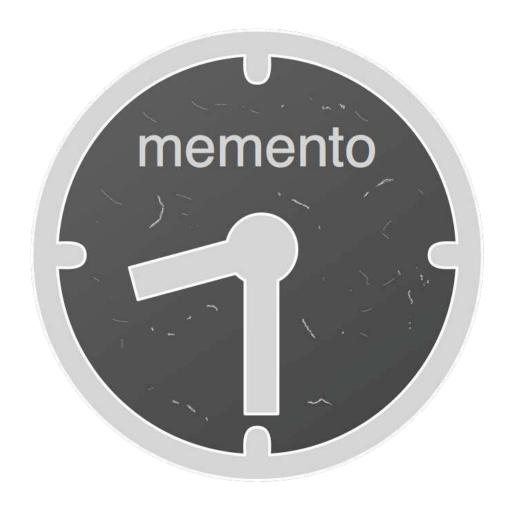
#### Recommended Citation

de Sompel, Herbert Van; Nelson, Michael L.; Sanderson, Robert; Balakireva, Lyudmila; Ainsworth, Scott; and Shankar, Harihar, "Memento: Time Travel for the Web" (2010). Computer Science Presentations. 18. https://digitalcommons.odu.edu/computerscience\_presentations/18

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.

Authors Herbert Van de Sompel, Michael L. Nelson, Robert Sanderson, Lyudmila Balakireva, Scott Ainsworth, and Harihar Shankar

#### Memento: Time Travel for the Web



The Memento Team

Michael L. Nelson
Robert Sanderson
Lyudmila Balakireva
Scott Ainsworth
Harihar Shankar

Memento is partially funded by the Library of Congress

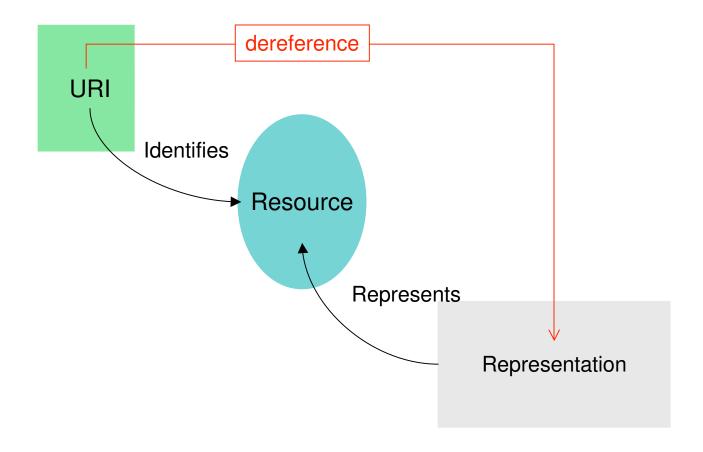








# W3C Web Architecture: Resource – URI - Representation



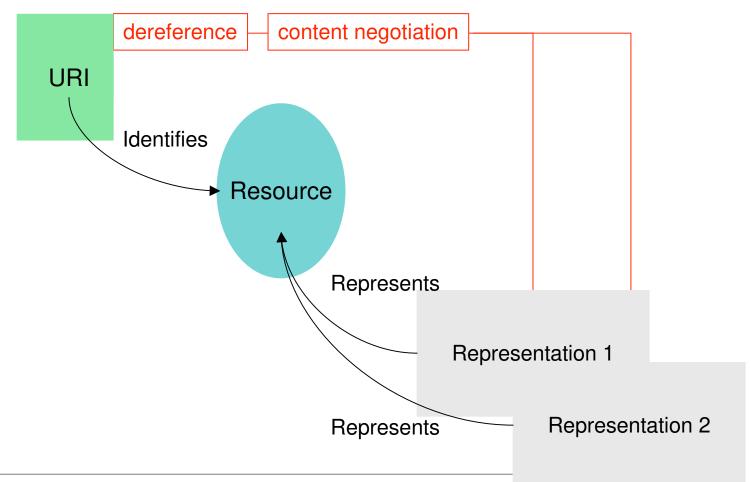








# W3C Web Architecture: Resource – URI - Representation



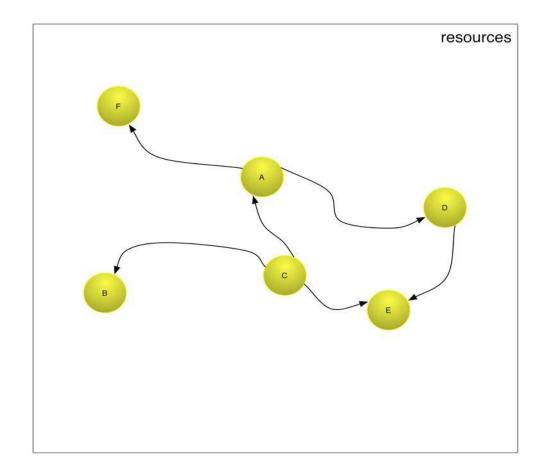








#### Resources



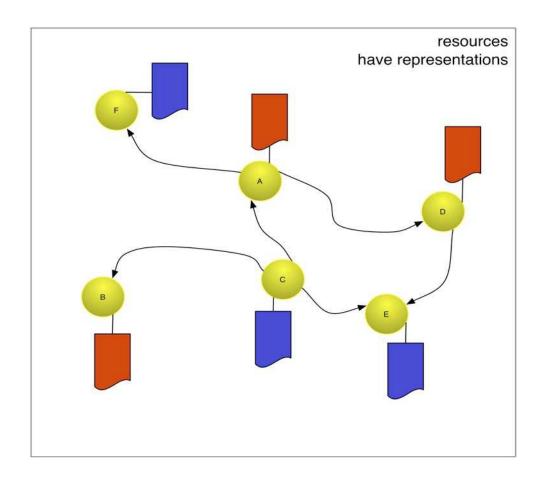








# Resources have Representations



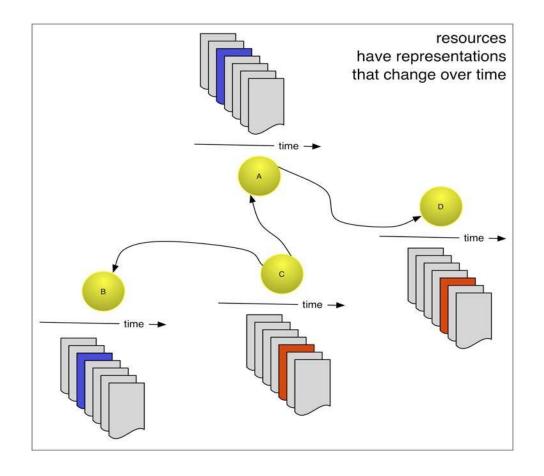








# Resources have Representations that Change over Time



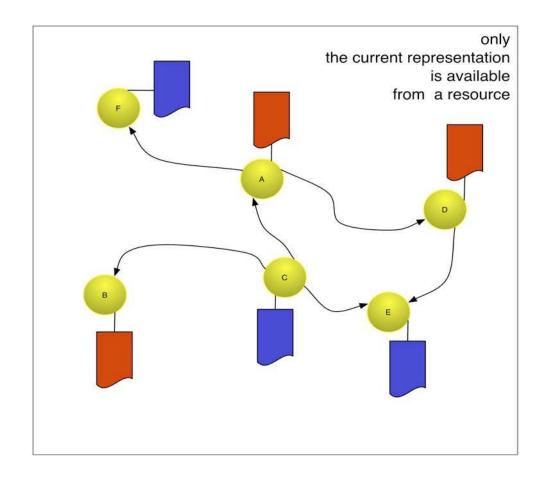








# Only the Current Representation is Available from a Resource



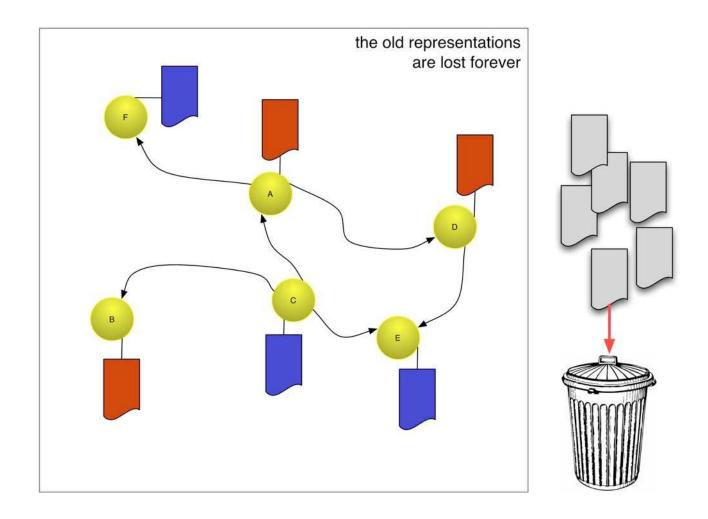








# Old Representations are Lost Forever



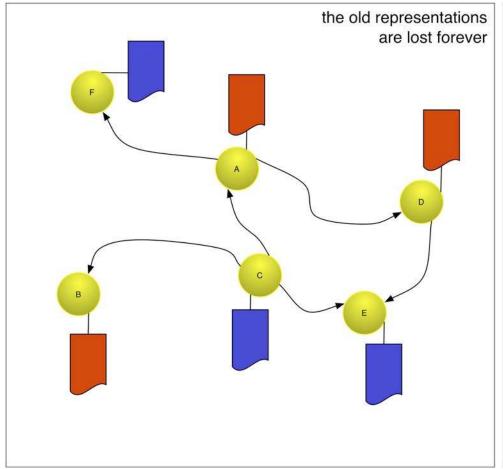


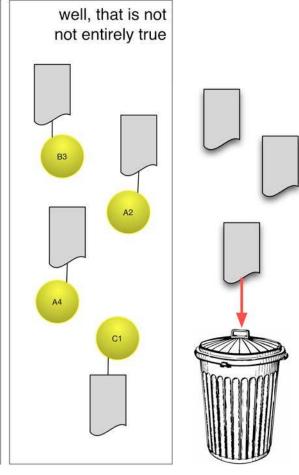






#### **Archived Resources Exist**













#### Dec 20 2001, 4:51:00 UTC

Try Beta 2 Log in / create account

#### **Archived Resources**



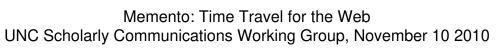
30 September 11 attacks From Wikipedia, the free encyclopedia This is an old revision of this page, as edited by The Cunctator (talk I contribs) at 04:51, 20 December 2001. It may differ significantly from the current revision WIKIPEDIA (diff) ← Previous revision | Current revision (diff) | Newer revision → (diff) navigation In Memoriam, September 11, 2001 Main page = Contents Featured content Current events On the morning of September 11, 2001, what might well be the most devastating terrorist attack in the history Bandom article of the world occurred concurrently in New York City, Washington, D.C. and near Pittsburgh. Four passenger jets were hijacked and then deliberately crashed into the World Trade Center and the Pentagon. Both towers of the World Trade Center subsequently collapsed, and part of the Pentagon was destroyed in the ensuing fire. Casualties are expected to be in the thousands: 265 on the planes: about 3000 people (early estimates Go Search ranged as high as 6500 people), including hundreds of firefighters who had rushed in, at the World Trade interaction Center; and 125 at the Pentagon. About Wikinedia Community portal Recent changes Some passengers on the doomed flights were able to make phone calls reporting on events on board. They reported that there were more than one hijacker on each plane and that they took control of the planes using Contact Wikipedia Donate to Wikipedia box-cutter knives. It appears that the passengers on the fourth jet tried to overpower the hijackers and that the = Help plane crashed in a sparsely populated area as a result, thereby missing its intended target, which may have been the White House. toolbox What links here Related changes The attack had immediate and deep global political effects and economic effects and inspired an international Upload file outpouring of memorials and services. The massive undertaking of rescue and recovery, and of providing Special pages assistance to the survivors and victims, is ongoing. There will be a great need for donations for a long time. Printable version Permanent link Cite this page Though no group has claimed responsibility, the U.S. government immediately launched a full-scale response, stating its intentions to go to war against those responsible. On October 7, a coalition led by the United States launched an attack in Afghanistan. See 2001 U.S. Attack on Afghanistan. Following the attack, the United States has been on heightened alert for new terrorist attacks. In late September, cases of anthrax started breaking out; although, no connection to the September 11 attack has been found. See 2001 anthrax attack

article discussion view source history

http://web.archive.org/web/20010911203610/http://www.cnn.com/archived resource for http://cnn.com



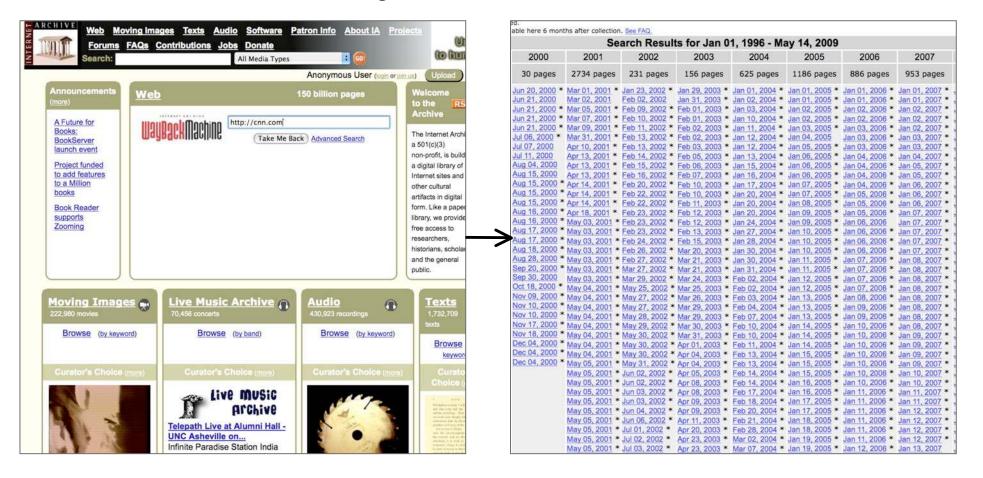








# Finding Archived Resources



Go to http://www.archive.org/ and search http://cnn.com

On http://web.archive.org/web/\*/http://cnn.com, select desired datetime

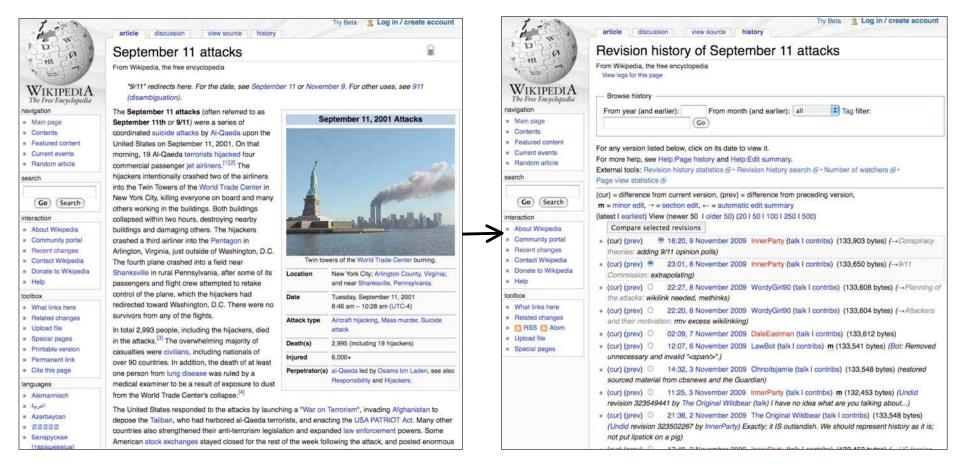








# Finding Archived Resources



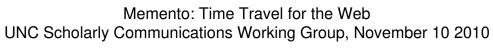
Go to

http://en.wikipedia.org/wiki/September\_11\_attacks and click History

**Browse History** 





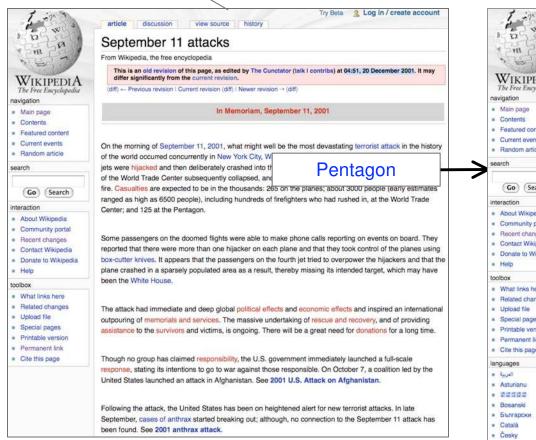




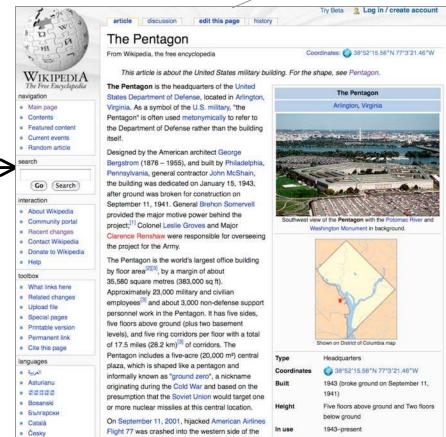


#### current

#### Navigating Archived Resources



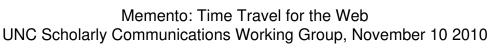
http://en.wikipedia.org/w/index.php?title=September\_1 1\_attacks&oldid=282333 archived resource for http://en.wikipedia.org/wiki/September\_11\_attacks3



http://en.wikipedia.org/wiki/The\_Pentagon



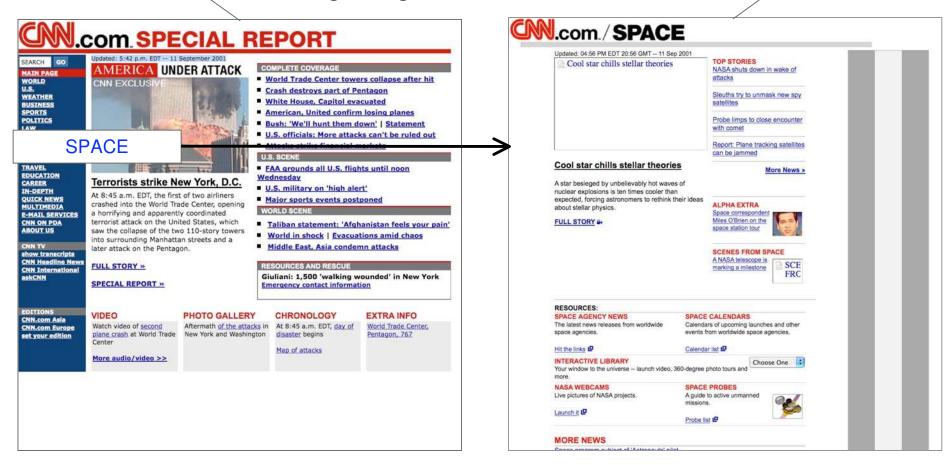








#### Navigating Archived Resources



http://web.archive.org/web/20010911203610/http://www.cnn.com/archived resource for http://cnn.com

http://web.archive.org/web/20010911213855/www.cnn .com/TECH/space/

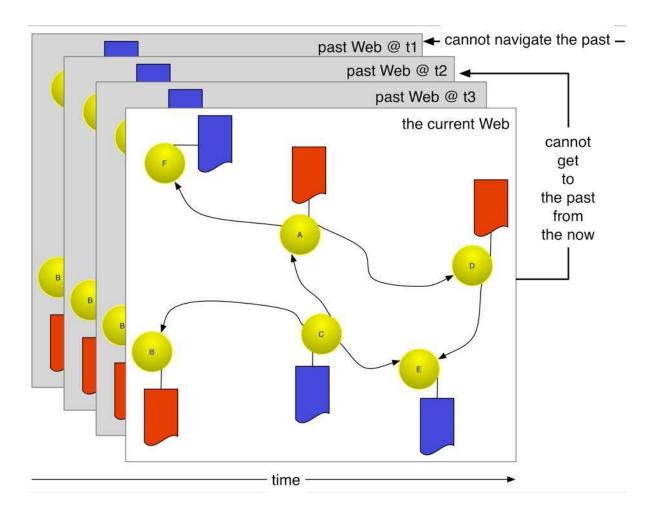








# Current and Past Web are Not Integrated



- Current and Past Web based on same technology.
- But, going from Current to Past Web is a matter of (manual) discovery.
- Memento wants to make going from Current to Past Web a (HTTP) protocol matter.
- Memento wants to integrate
   Current And Past Web.

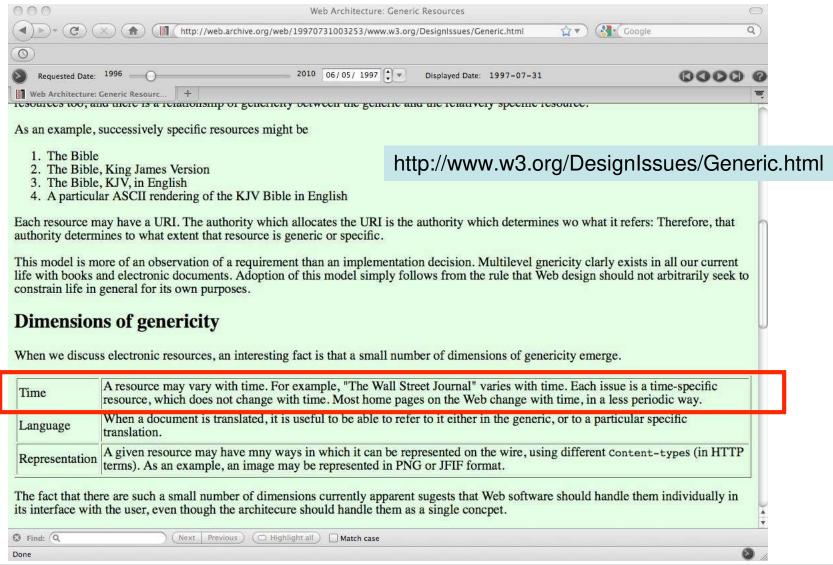








#### TBL on Generic vs. Specific Resources











#### In The Beginning... there was the inode

```
struct stat {
   dev_t st_dev; /* ID of device containing file */
   ino_t st_ino; /* inode number */
   mode_t st_mode; /* protection */
   nlink_t st_nlink; /* number of hard links */
   uid t     st uid;     /* user ID of owner */
   gid_t st_gid; /* group ID of owner */
   dev_t st_rdev; /* device ID (if special file) */
   off t st size; /* total size, in bytes */
   blksize_t st_blksize; /* blocksize for filesystem I/O */
   blkcnt_t st_blocks; /* number of blocks allocated */
   time_t st_atime; /* time of last access */
   time_t st_mtime; /* time of last modification */
   time t st ctime; /* time of last status change */
};
```









#### Limited Time Semantics...

% telnet www.digitalpreservation.gov 80
Trying 140.147.249.7...
Connected to www.digitalpreservation.gov.
Escape character is '^]'.
HEAD /images/ndiipp\_header6.jpg HTTP/1.1
Host: www.digitalpreservation.gov
Connection: close

HTTP/1.1 200 OK

Date: Mon, 19 Jul 2010 21:41:04 GMT

Server: Apache

Last-Modified: Thu, 18 Jun 2009 16:25:54 GMT

ETag: "1bc861-10935-dca24880"

Accept-Ranges: bytes Content-Length: 67893

Connection: close

Content-Type: image/jpeg

Connection closed by foreign host.









# Time Semantics Becoming Less, Not More Available

```
% telnet www.digitalpreservation.gov 80
Trying 140.147.249.7...
Connected to www.digitalpreservation.gov.
Escape character is '^]'.
HEAD / HTTP/1.1
Host: www.digitalpreservation.gov
Connection: close
HTTP/1.1 200 OK
Date: Mon, 19 Jul 2010 21:36:00 GMT
Server: Apache
Accept-Ranges: bytes
Connection: close
Content-Type: text/html
Connection closed by foreign host.
```

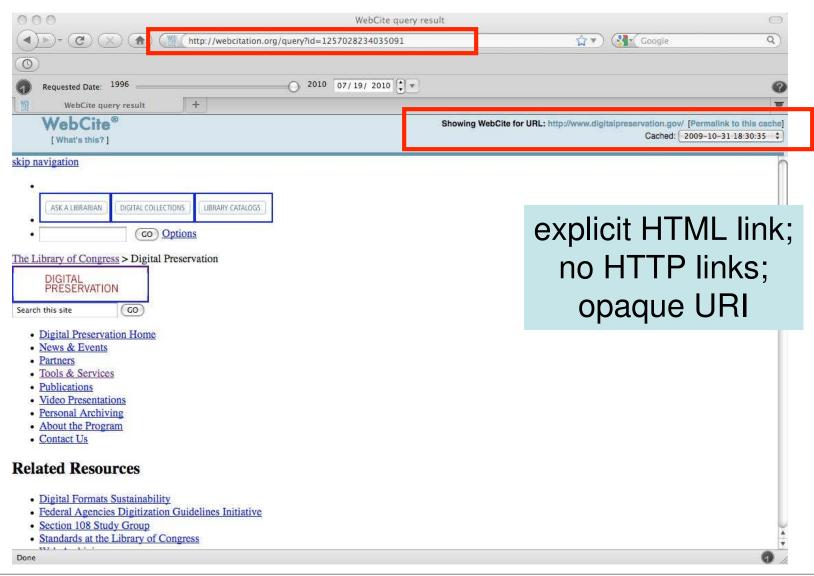








#### The Past Links to the Present...



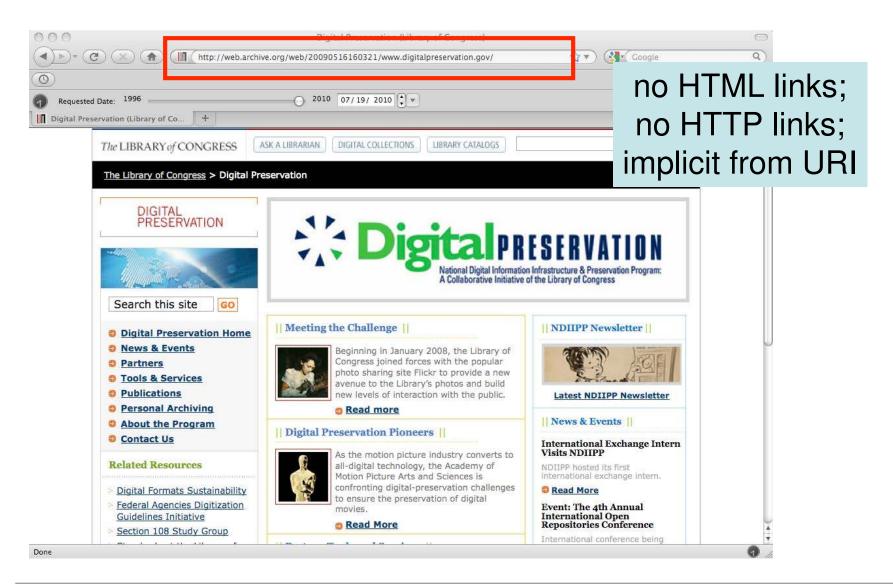








#### The Past Links to the Present...



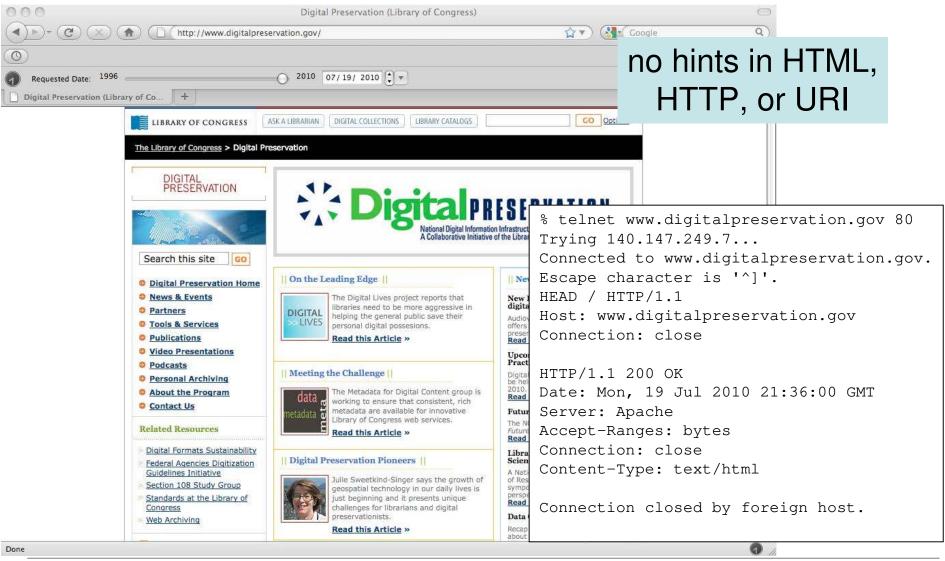








#### But the Present Does Not Link to the Past

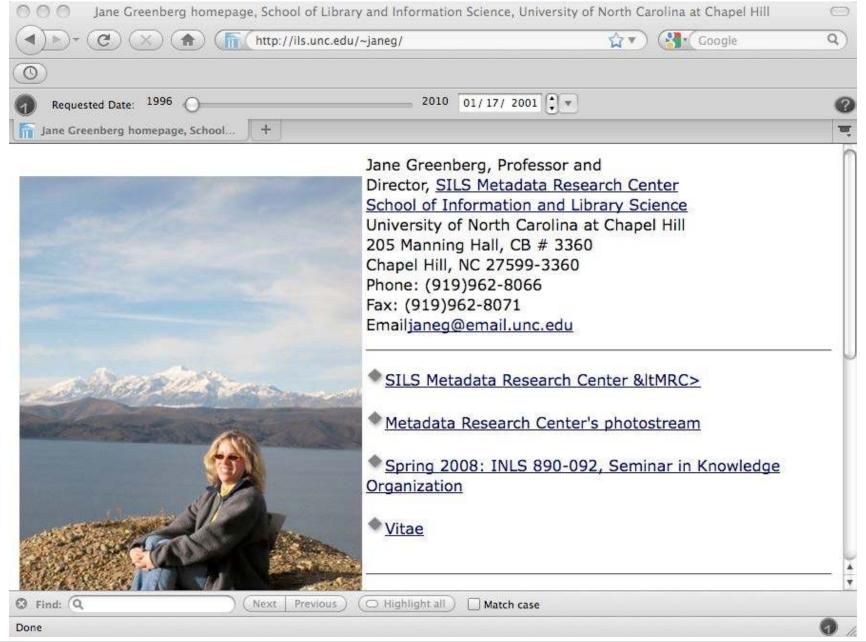










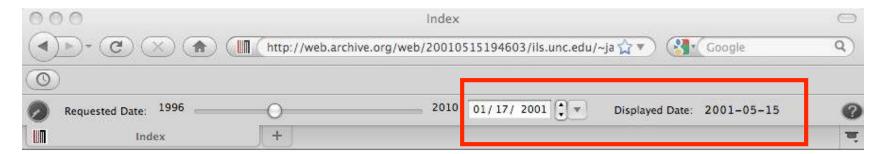














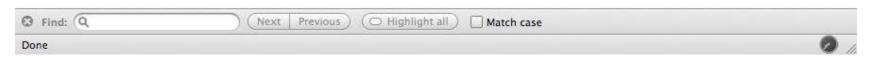
Jane Greenberg, Assistant Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360

Tel: (919)962-7024

E-mail: janeg@ils.unc.edu

Fax: (919)962-8071

• Biographical Statement • Research • Vitae • Teaching











# Linking the Past and the Present

- Codify existing methods to create linkage from the past to the present
  - easy: an archived version knows for which URI it is an archived version
- Create a linkage from the present to the past
  - hard: solve with a level of indirection from present to past

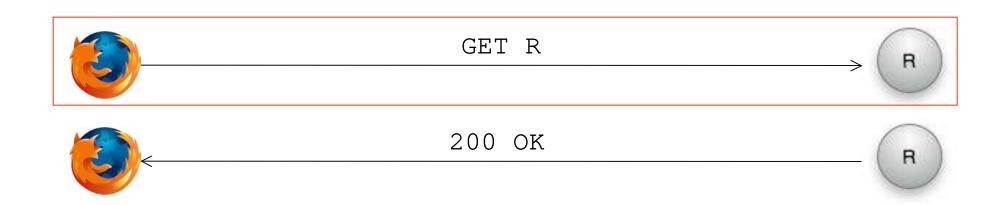








# Normal HTTP Flow



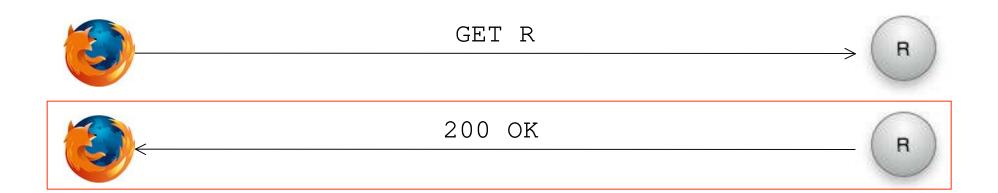








# Normal HTTP Flow











#### The Web without a Time Dimension



# Original resource

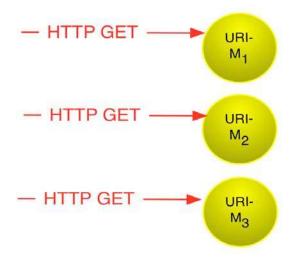








#### The Web without a Time Dimension



Archived resources

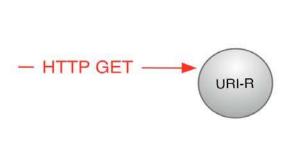


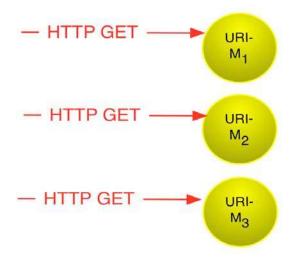






#### The Web without a Time Dimension





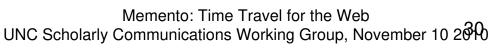
Original resource

Archived resources

Need to use a different URI to access archived versions of a resource and its current version











#### The Web with Time Dimension added by Memento



URI-M<sub>1</sub>





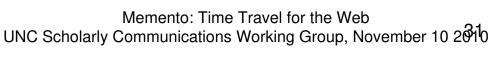
Original resource

Archived resources

In Memento: use URI of the current version to access archived versions, but qualify it with datetime



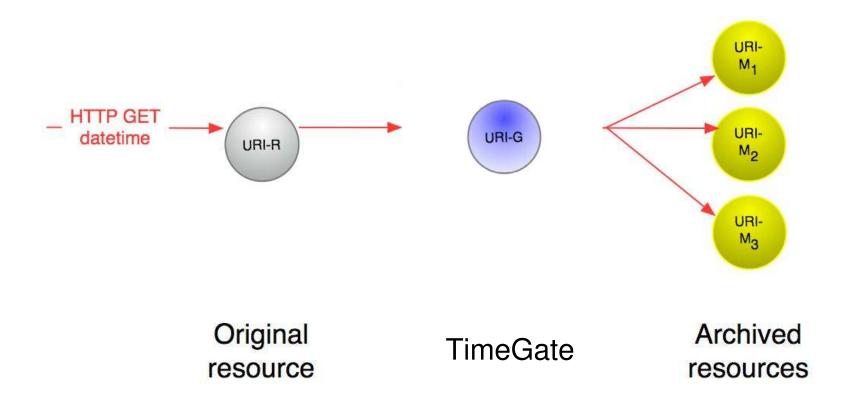








# The Web with Time Dimension added by Memento



... and arrive at an archived version via level of indirection









#### Content Negotiation in the datetime dimension

- Many systems support content negotiation for media type:
  - Your client by default asks for HTML and gets HTML;
  - But it could get PDF via the same URI.
- Memento proposes a new dimension for content negotiation time:
  - Your client by default asks for the current time, and gets it
  - But it could get an older version via the same URI
- Can be accomplished with two new HTTP headers:
  - Request header: Accept-Datetime
    - Conveys datetime of content requested by client
  - Response header: Memento-Datetime
    - Conveys datetime of content returned by server

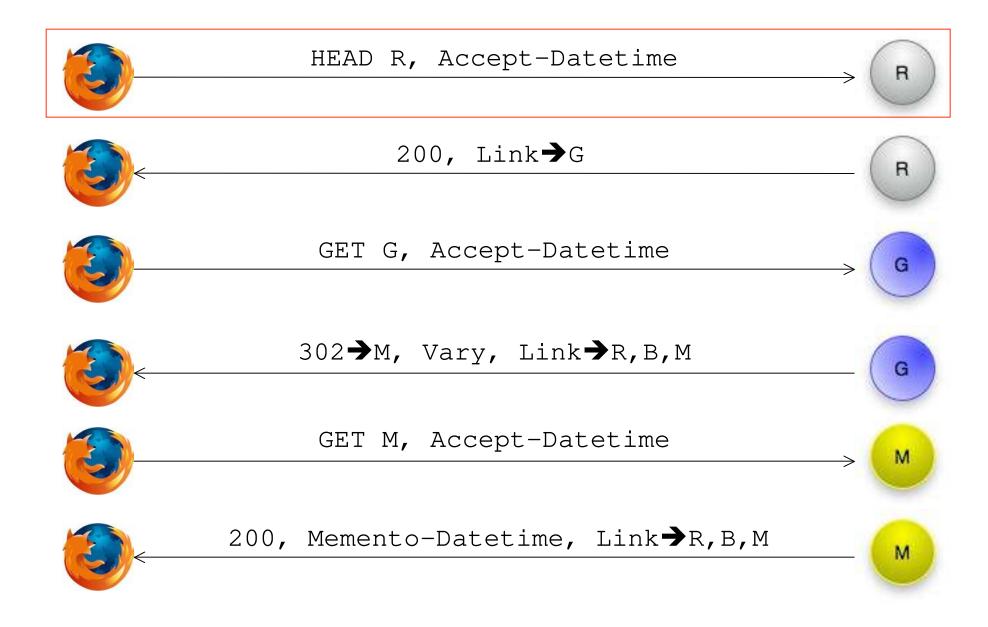


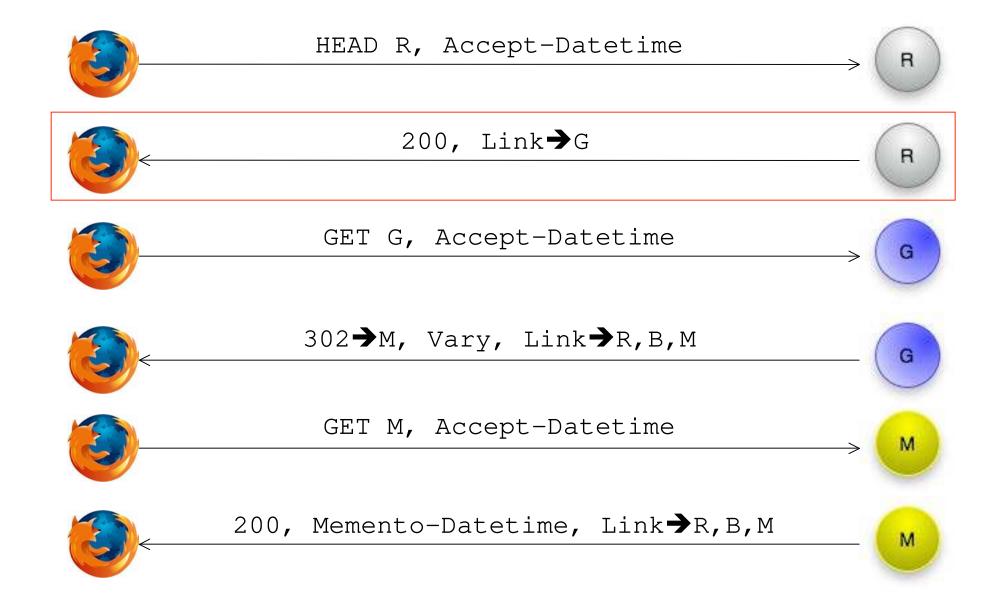


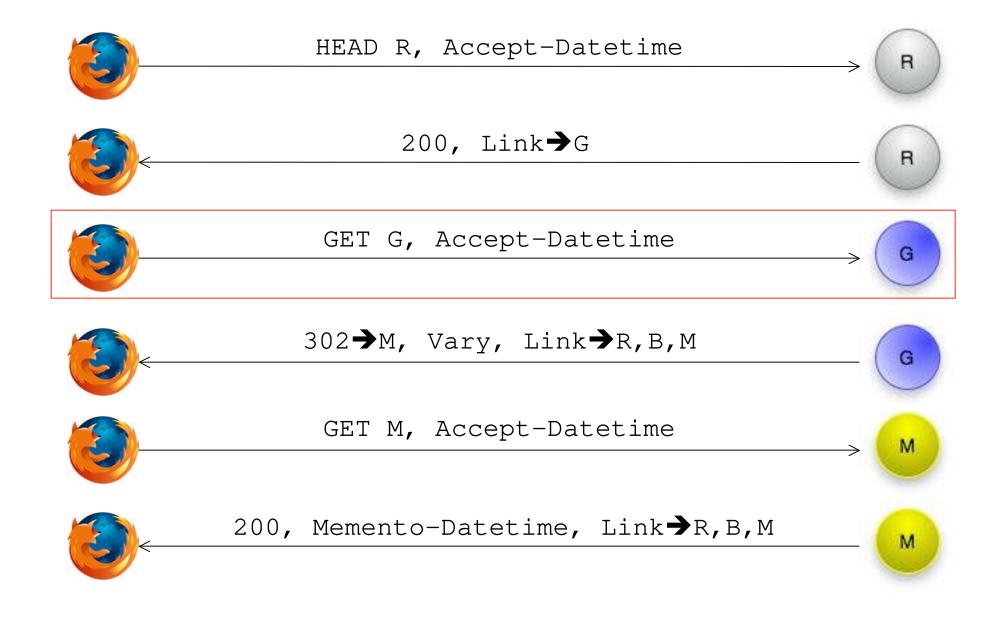


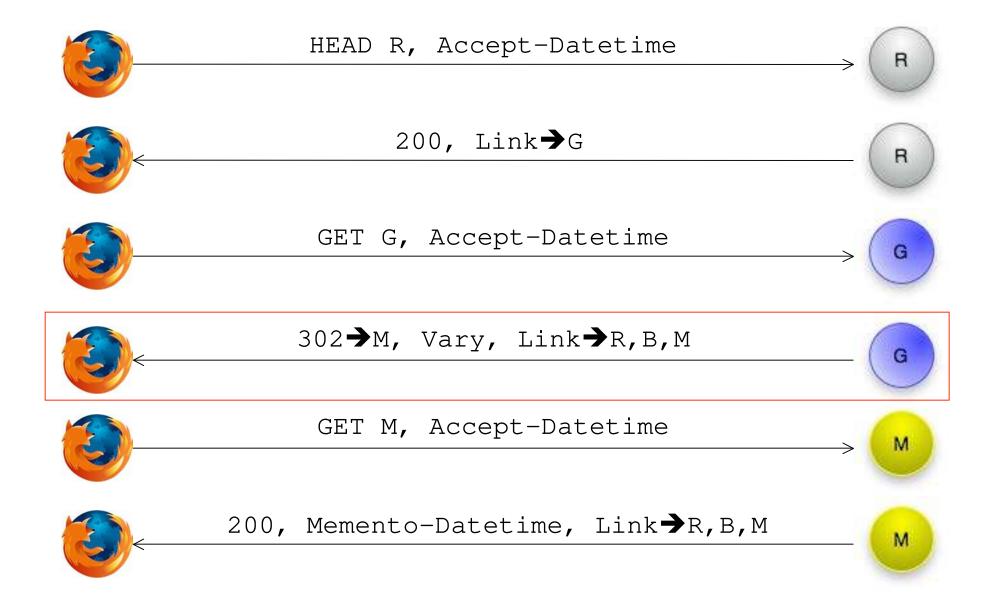


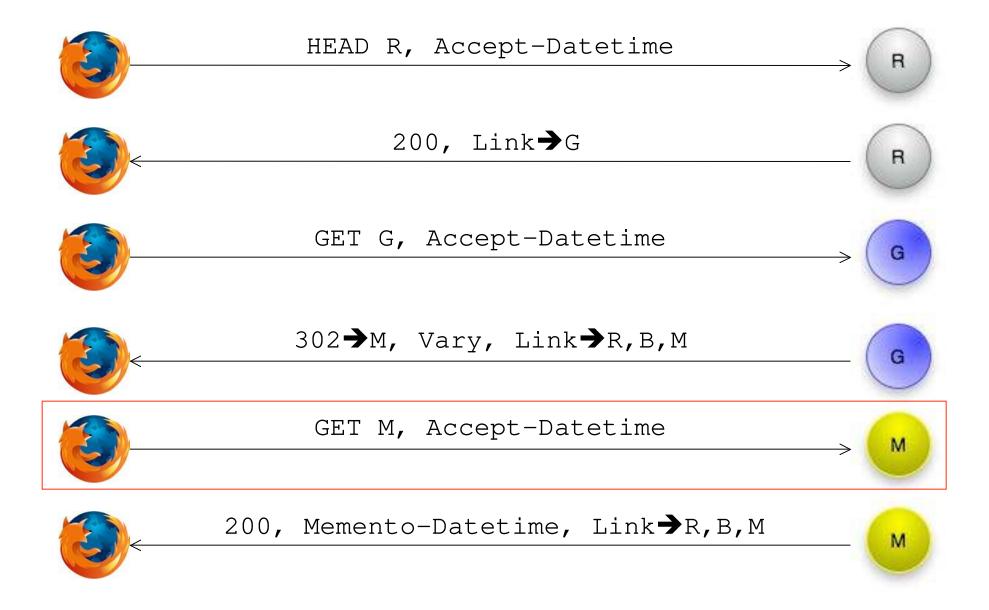
#### Memento HTTP Flow

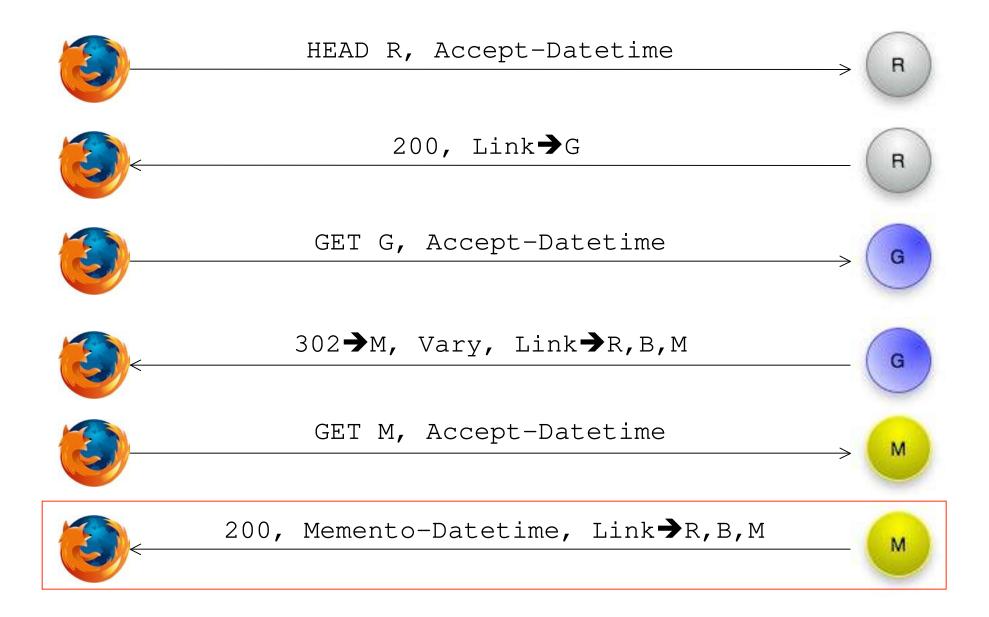




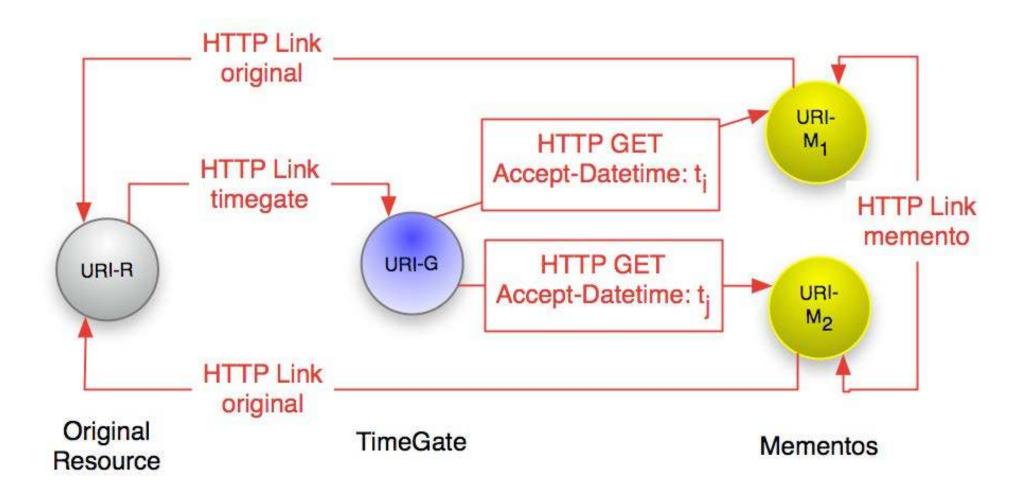








#### The Memento Framework











# Why Not Bypass URI-R and Go Directly to TimeGate?

- The Original Resource (URI-R) might know a "better"
   TimeGate than the default client value
  - a CMS (e.g., a wiki) or transactional archive will always have the most complete archival coverage for a URI-R
  - the client can always ignore the URI-R's TimeGate suggestion
- Summary: it is good design to check with URI-R before using the default the TimeGate

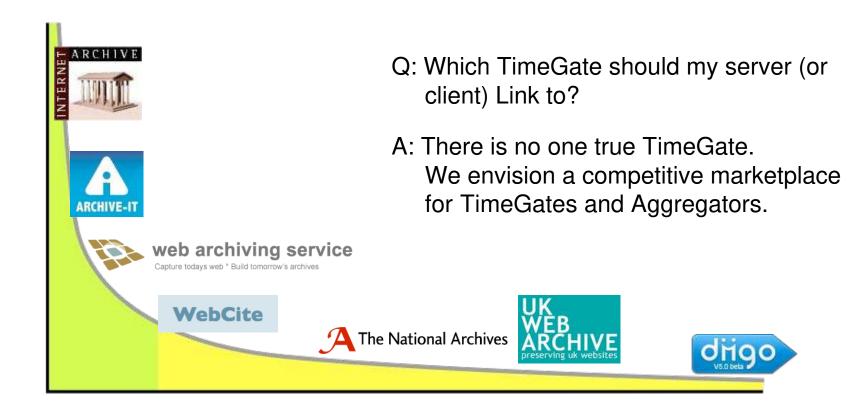








## Long Tail of Archives











#### **Transition Period**

- Up until now, we've presented the scenario where both the client and server are compliant with the Memento framework
- Good news: Memento elegantly handles scenarios where either the client or the server are not (yet) compliant









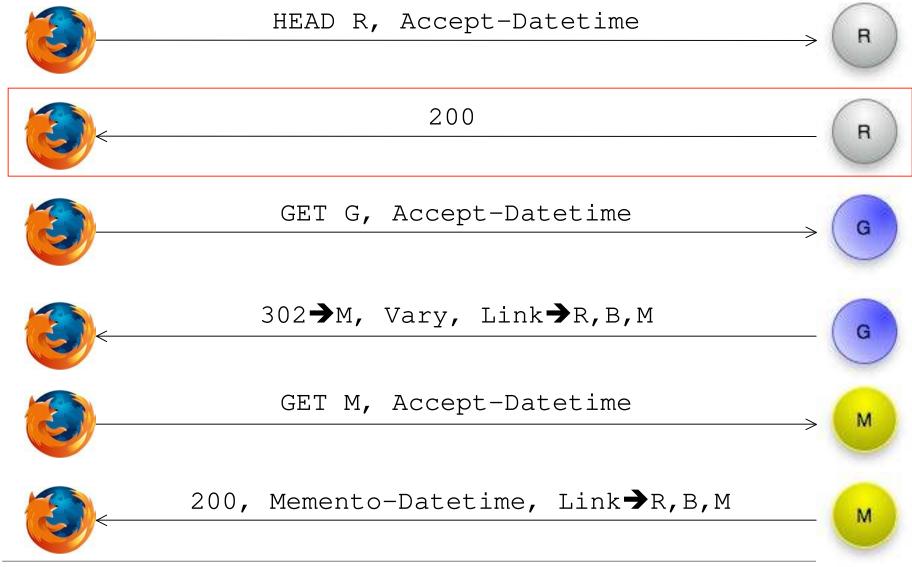
HEAD R, Accept-Datetime	$\rightarrow$ R
200	R
GET G, Accept-Datetime	$\rightarrow$ G
302→M, Vary, Link→R,B,M	G
GET M, Accept-Datetime	$\rightarrow$ $M$
200, Memento-Datetime, Link→R,B,M	M









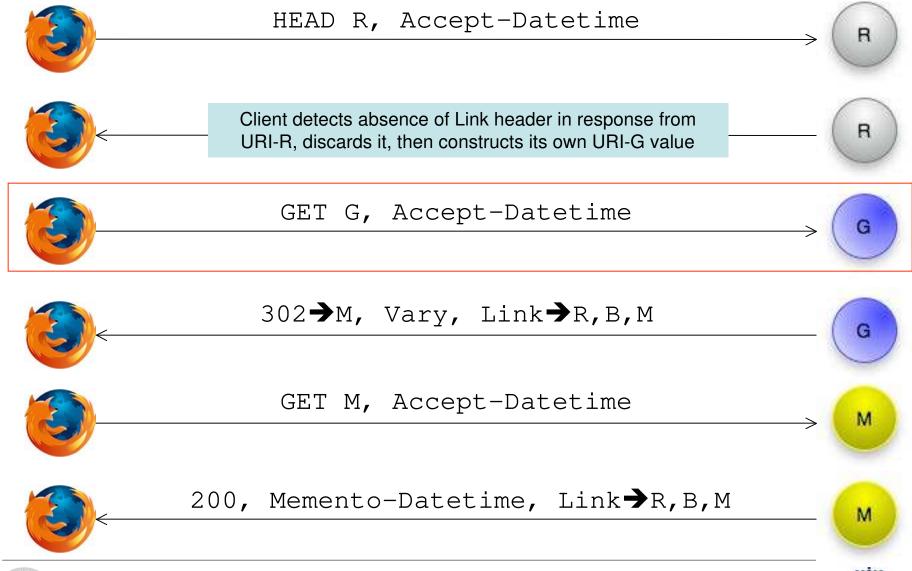










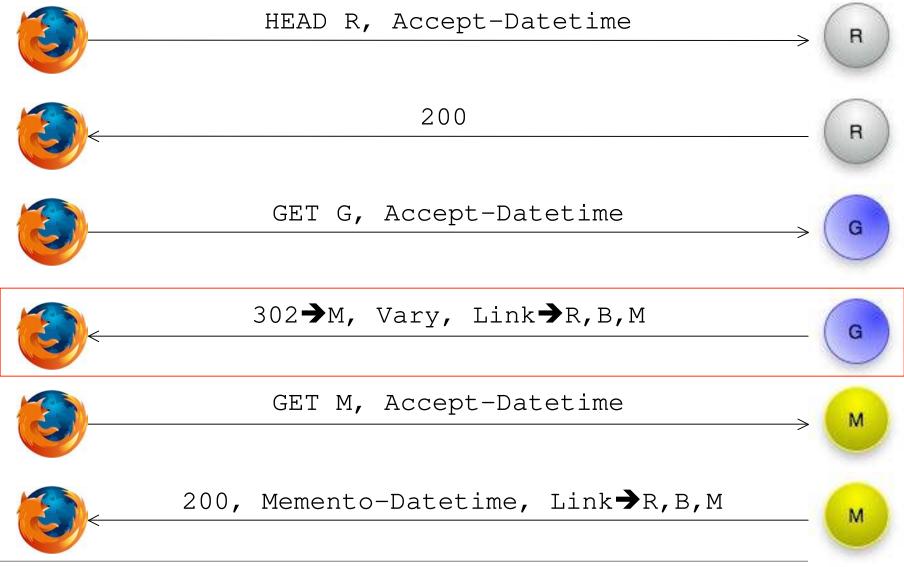










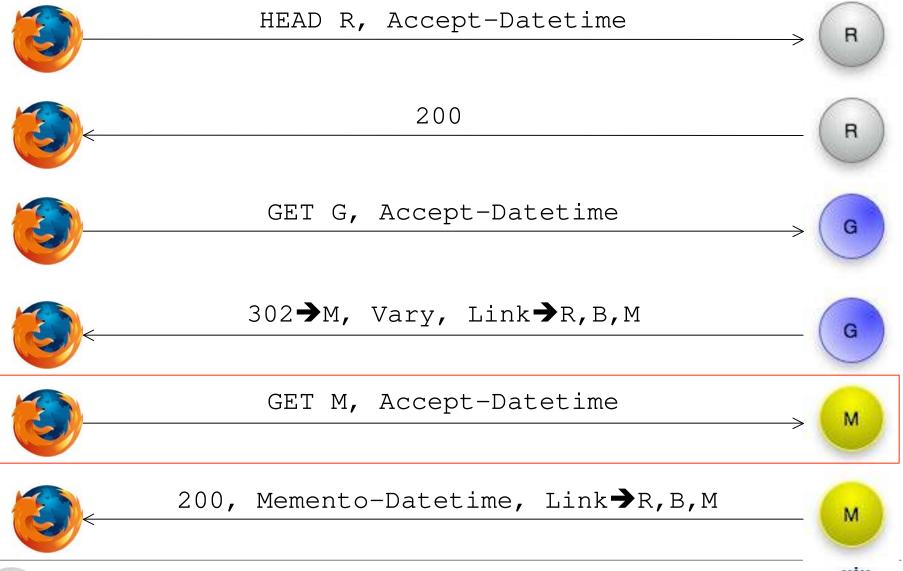










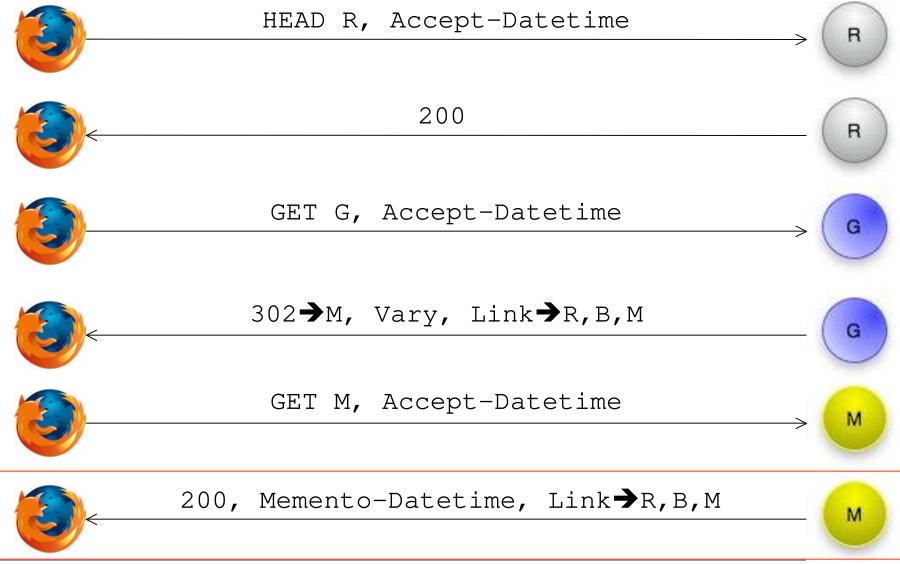












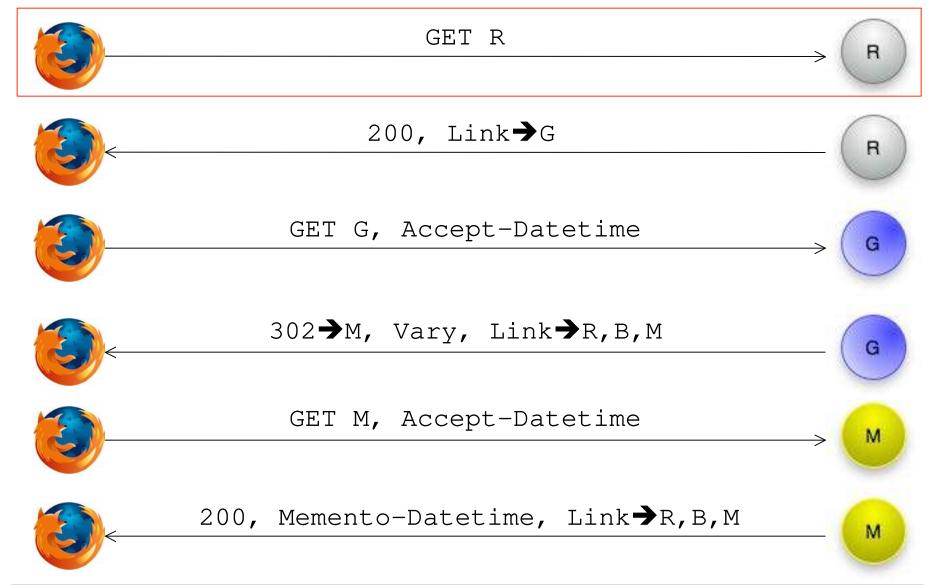








## Compliant Server, Non-Compliant Client



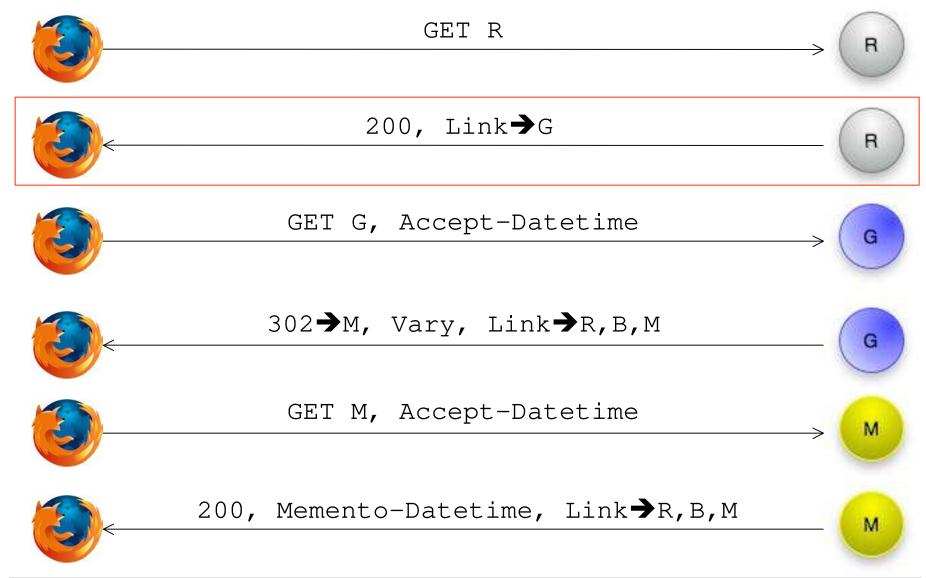








## Compliant Server, Non-Compliant Client



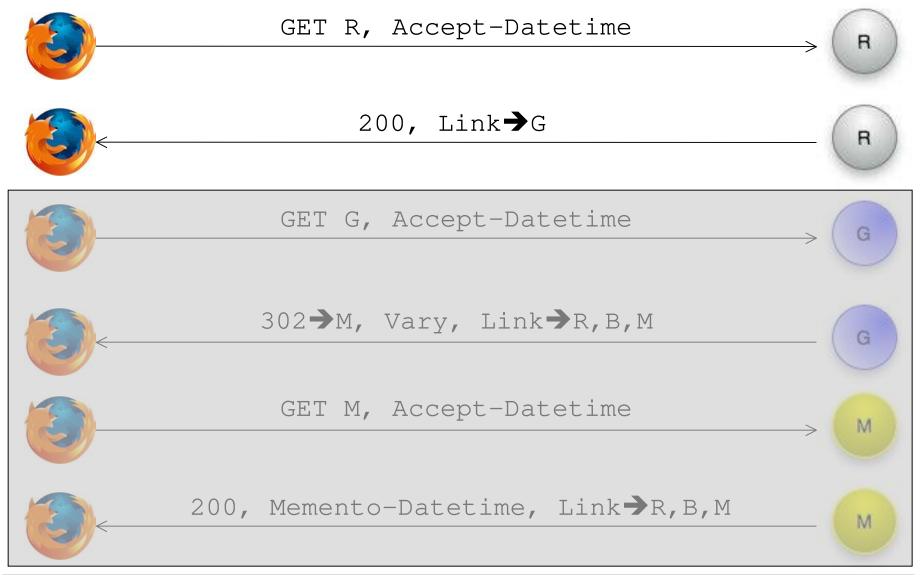








## Compliant Server, Non-Compliant Client











#### Some Issues

- Observational uncertainty
- Different notions of time
- When is the past?



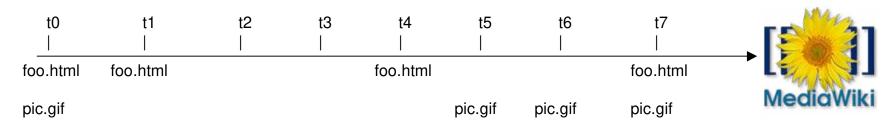






# No Uncertainty With Self-Archiving Systems

foo.html has <img src=pic.gif>



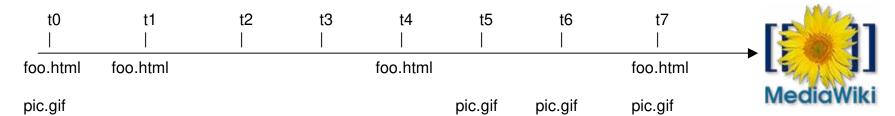








#### foo.html has <img src=pic.gif>



GET /foo.html

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t4

GET /pic.gif

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t0

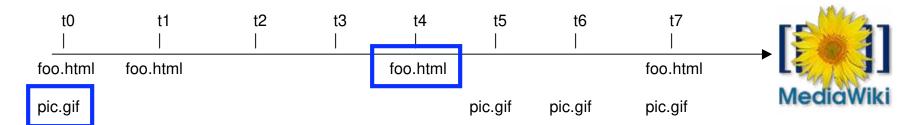








#### foo.html has <img src=pic.gif>



GET /foo.html

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t4

GET /pic.gif

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t0

foo.html correct

pic.gif correct



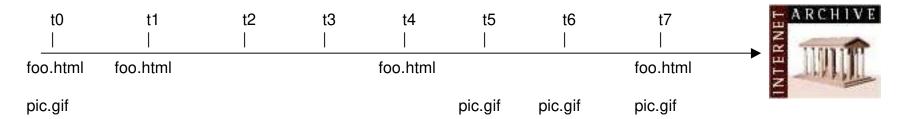






# Uncertainty in Third-Party Archives

foo.html has <img src=pic.gif>





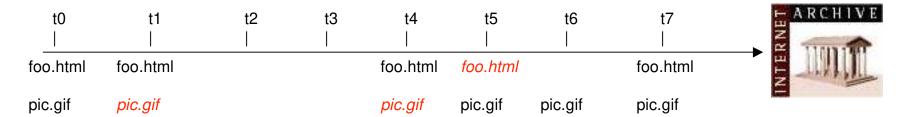






# Missed Updates

foo.html has <img src=pic.gif>



red italics = missed updates

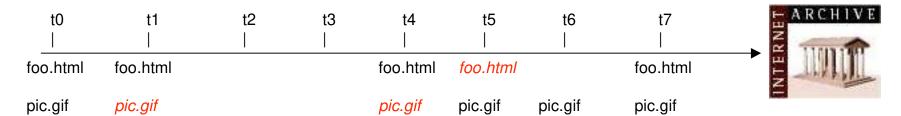








#### foo.html has <img src=pic.gif>



GET /foo.html

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t4

GET /pic.gif

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t0

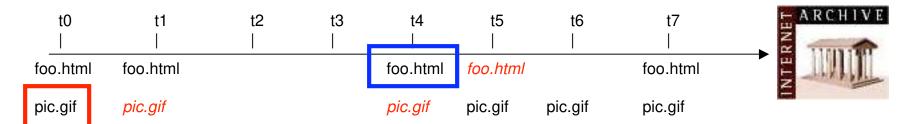








#### foo.html has <img src=pic.gif>



GET /foo.html

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t4

GET /pic.gif

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t0

foo.html correct

pic.gif incorrect (should be t4)

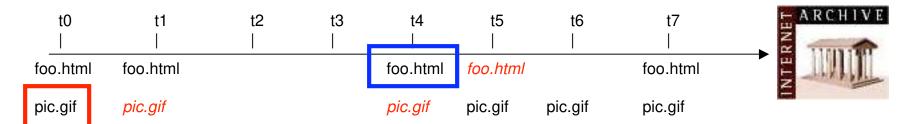








#### foo.html has <img src=pic.gif>



GET /foo.html

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t4

GET /pic.gif

Accept-Datetime: t4

HTTP/1.1 200 OK

Memento-Datetime: t0

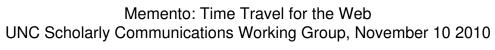
foo.html correct

pic.gif incorrect (should be t4)

this combination (foo@t4, pic@t0) never existed!





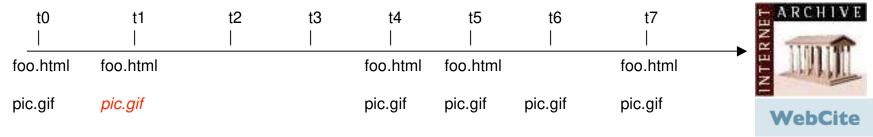






# Decrease Uncertainty With More Observations?

foo.html has <img src=pic.gif>



red italics = missed updates











### Three Notions of Time: Cr, LM, MD

- Creation (Cr): datetime when the resource first came into being
- Last-Modified (LM): when the resource was last changed
- Memento-Datetime (MD): the datetime that the resource was (meaningfully) observed on the web
  - not the same as the datetime that the resource is "about"; i.e. there is no

"Memento-Datetime: Thu, 04 Jul 1776 14:37:00"

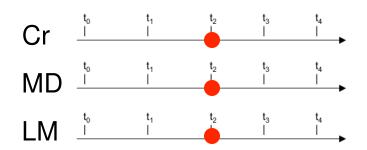


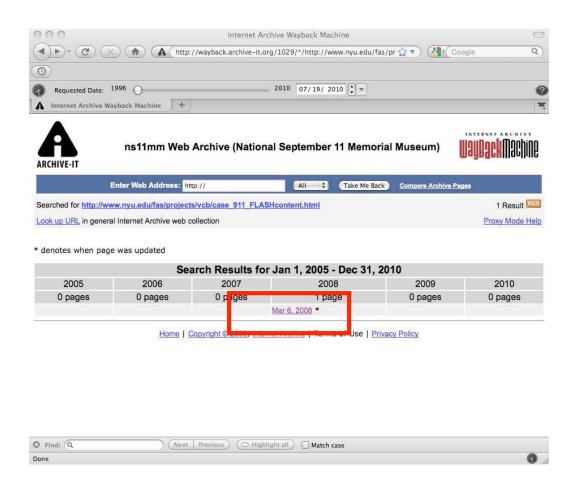






#### Cr == MD == LM





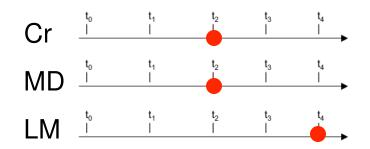








#### Cr == MD < LM





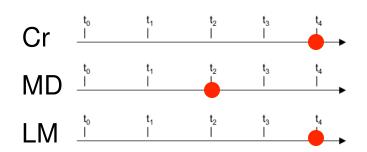


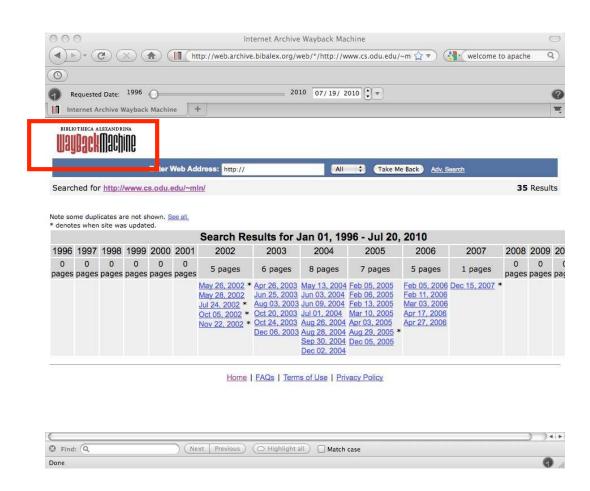






#### MD < Cr <= LM





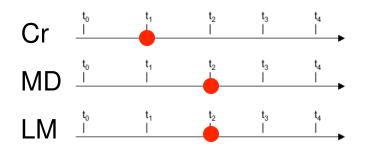


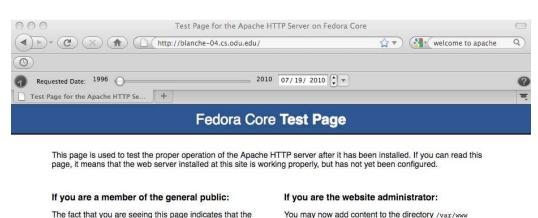






#### Cr < MD <= LM





website you just visited is either experiencing problems, or is undergoing routine maintenance.

If you would like to let the administrators of this website

know that you've seen this page instead of the page you expected, you should send them e-mail. In general, mail sent to the name "webmaster" and directed to the website's domain should reach the appropriate person.

For example, if you experienced problems while visiting www.example.com, you should send e-mail to "webmaster@example.com".

For information on Fedora Core, please visit the <u>Fedora</u> <u>Project website</u>.

You may now add content to the directory /var/www /html/. Note that until you do so, people visiting your website will see this page, and not your content. To prevent this page from ever being used, follow the instructions in the file /etc/httpd/conf.d/welcome.conf.

You are free to use the images below on Apache and Fedora Core powered HTTP servers. Thanks for using Apache and Fedora Core!







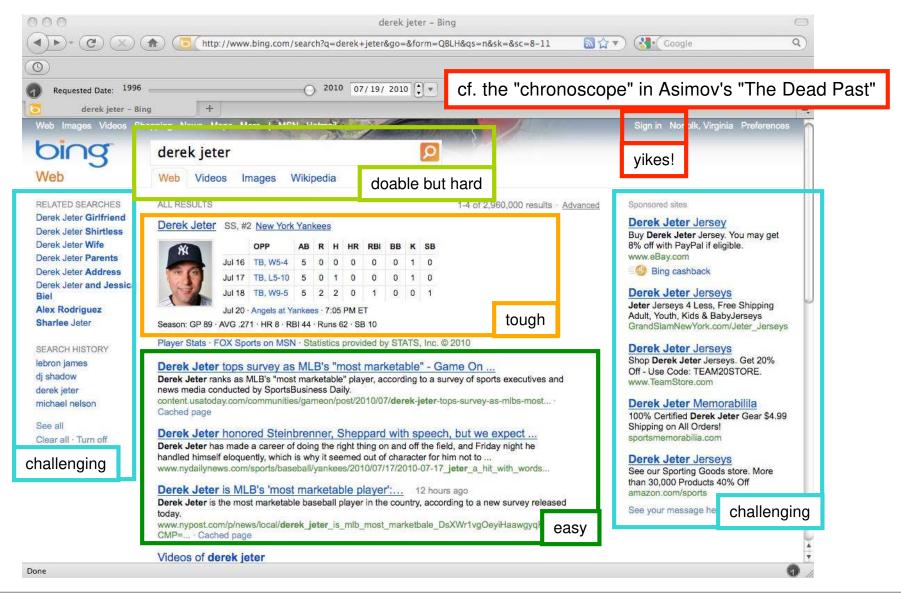








#### When is the Past?











## Why Care About The Past?

From an anonymous reviewer (emphasis mine):

"Is there any statistics to show that many or a good number of Web users would like to get obsolete data or resources?"









# Replaying the Experience...

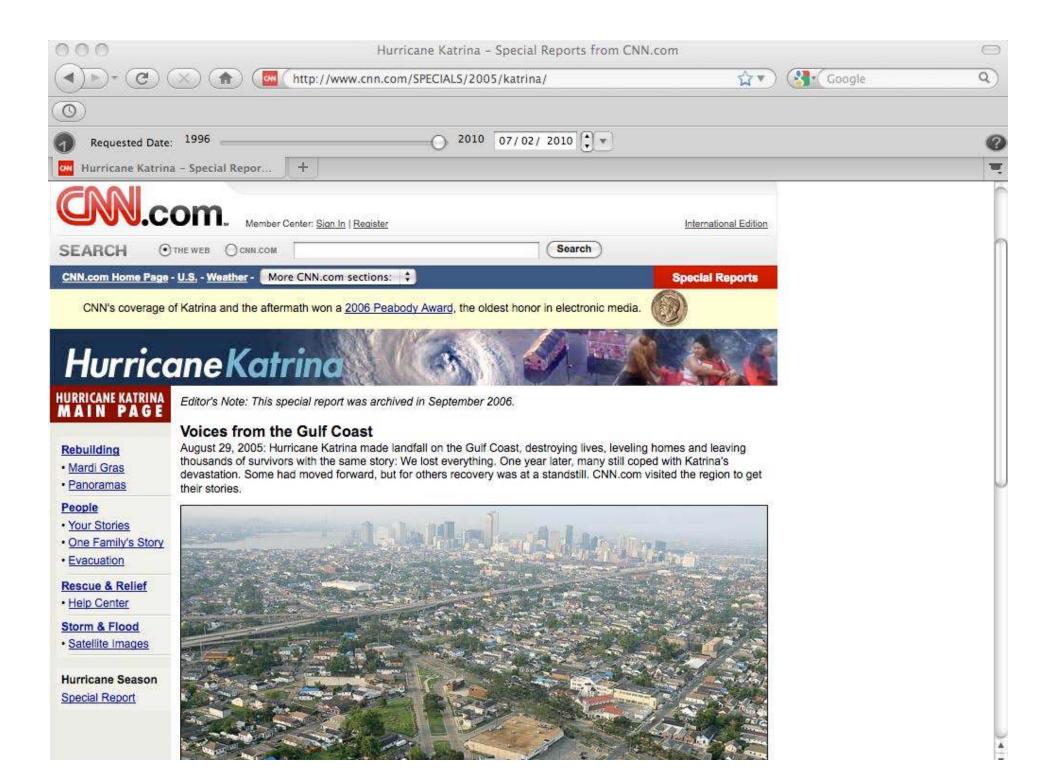
...can be more compelling than a summary











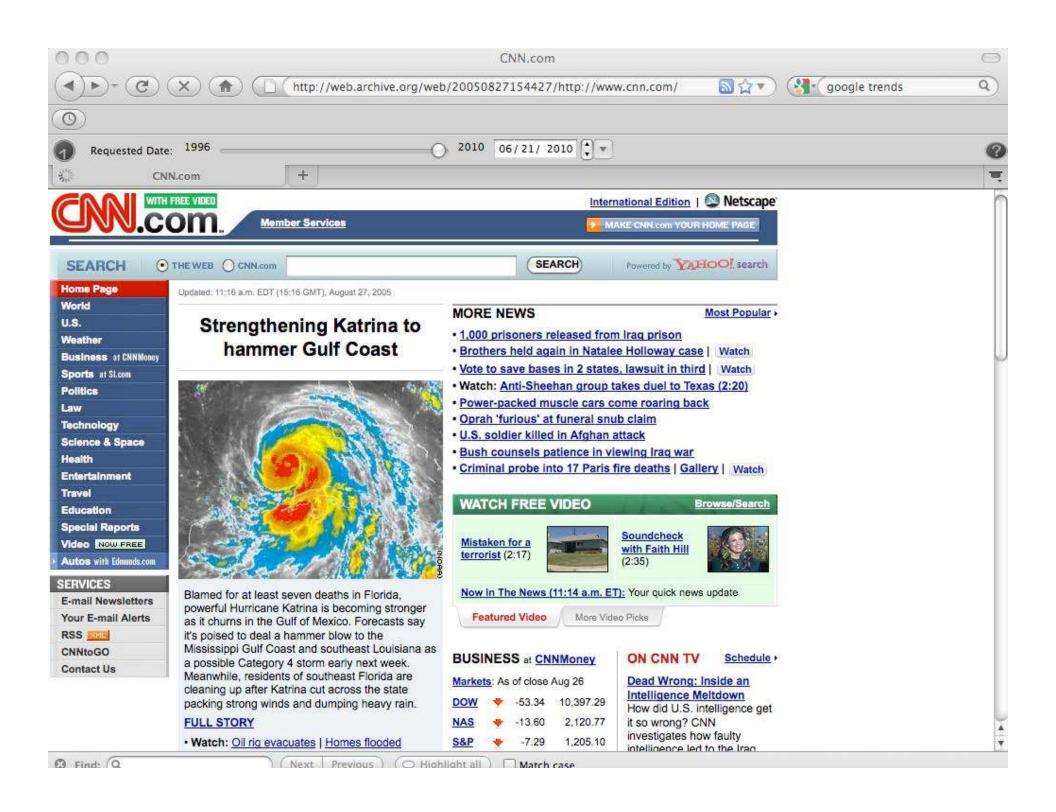
(thanks to Michele Weigle for the following Memento selection)

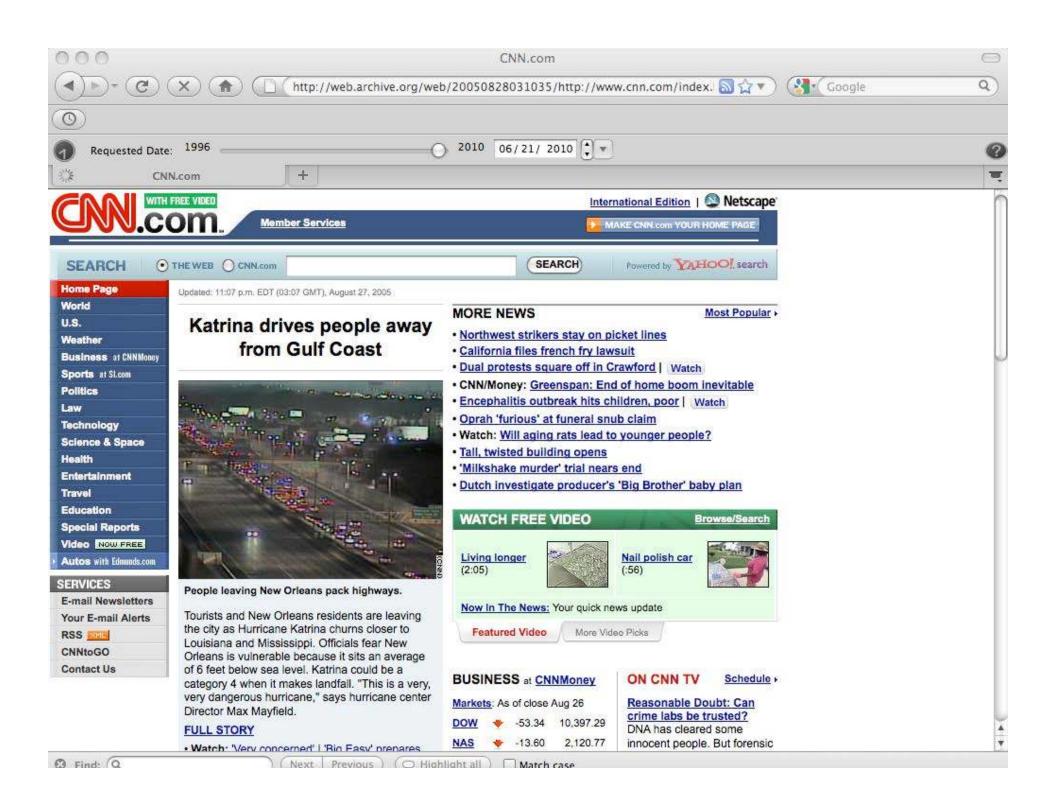


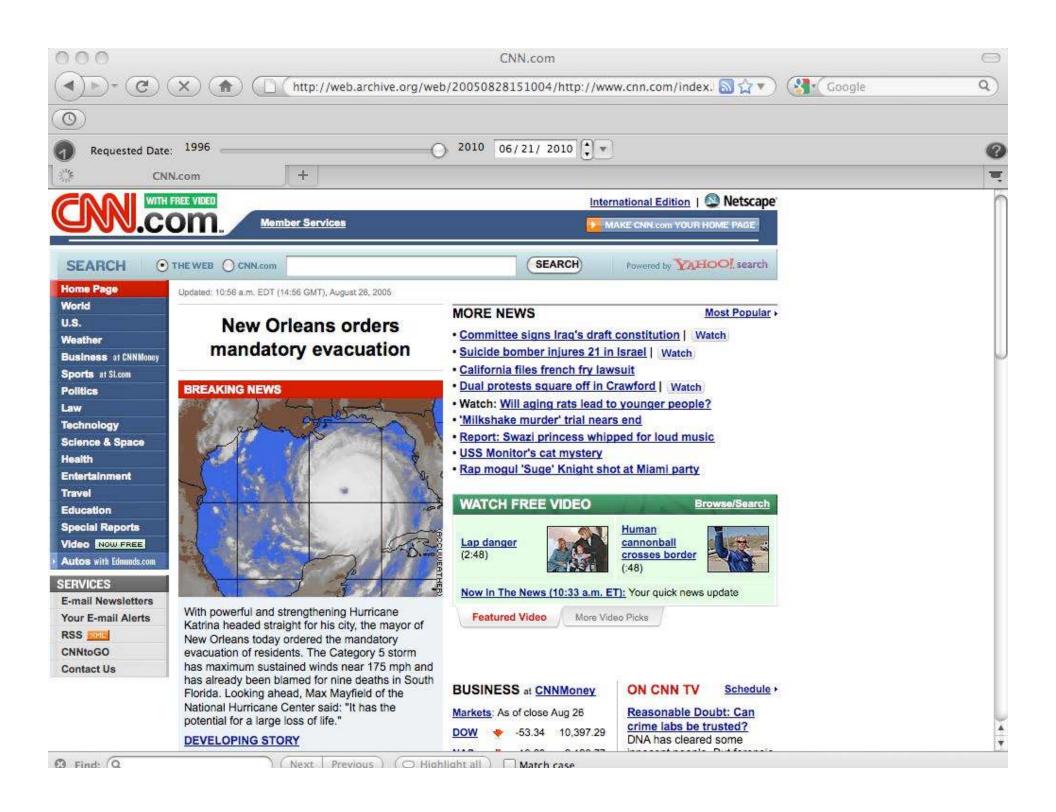


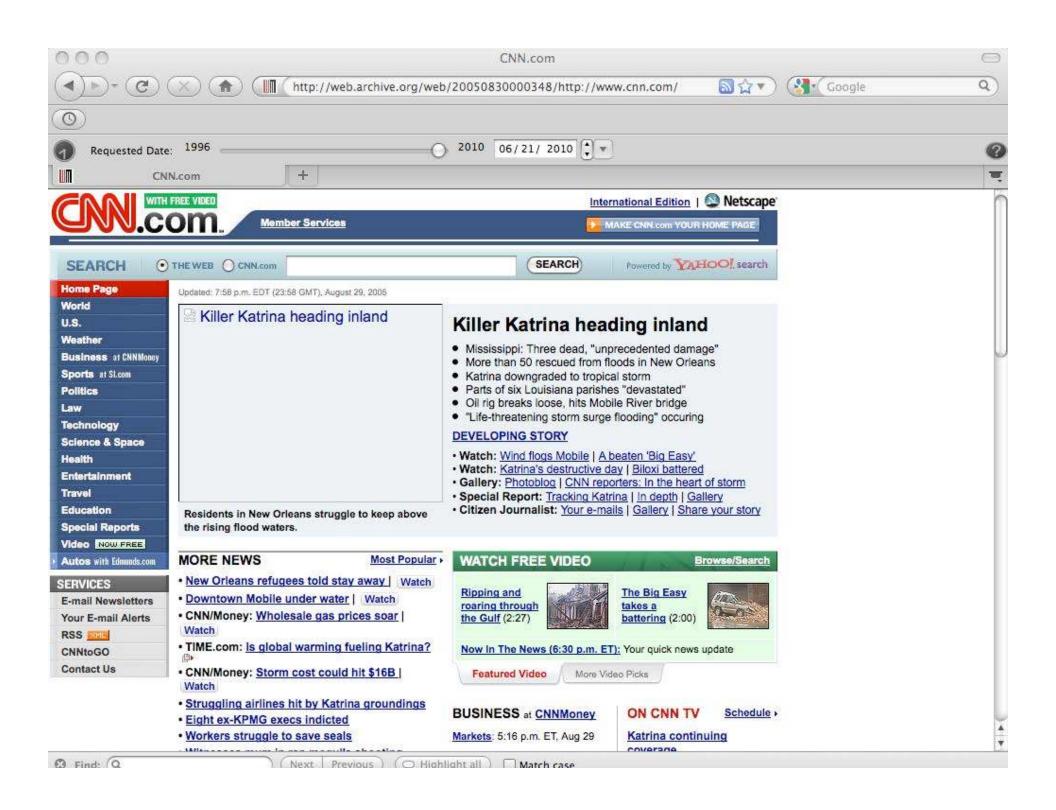


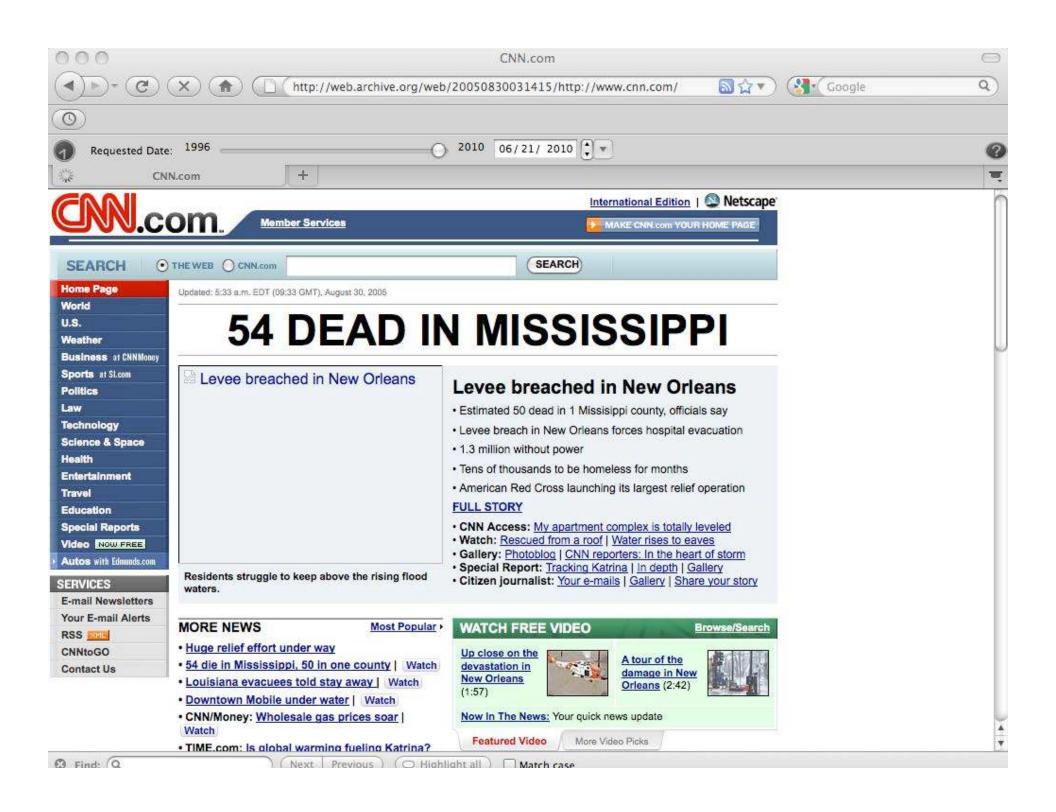


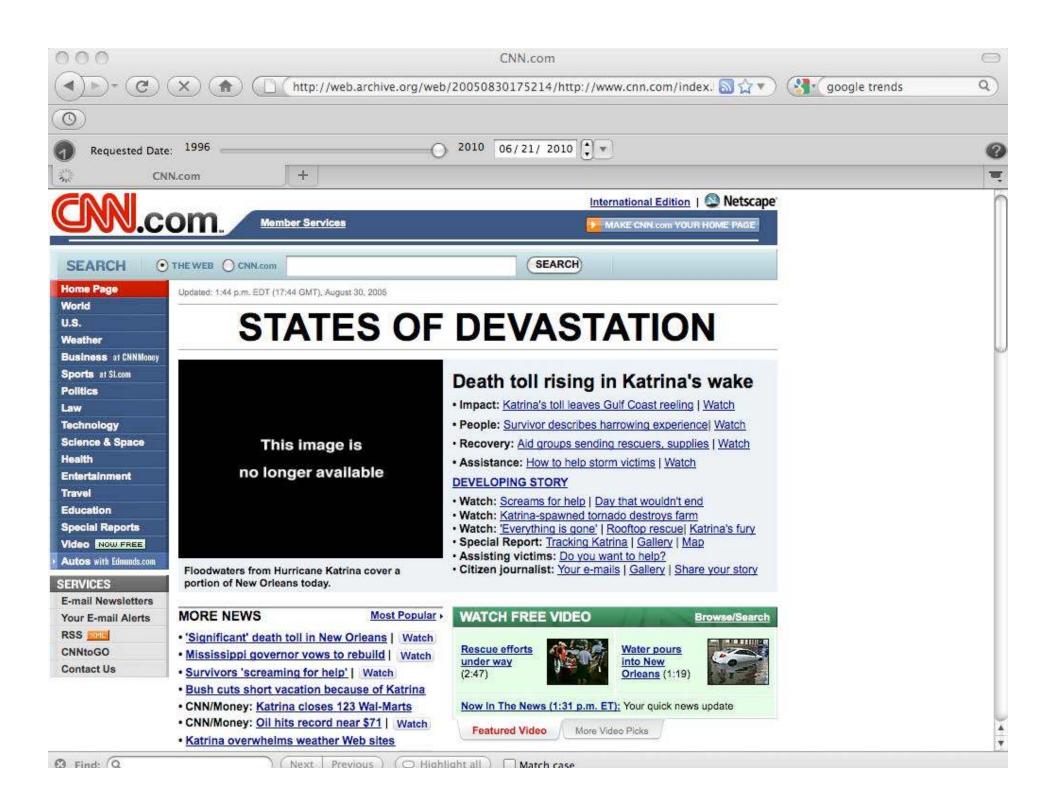






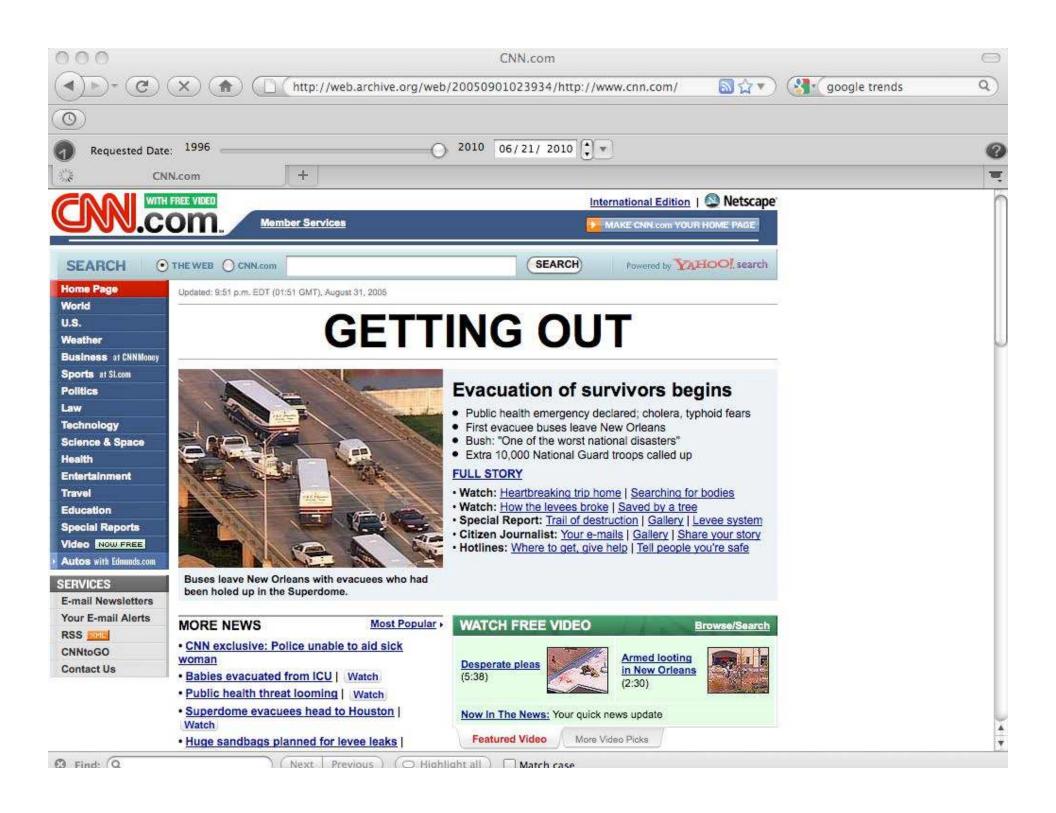


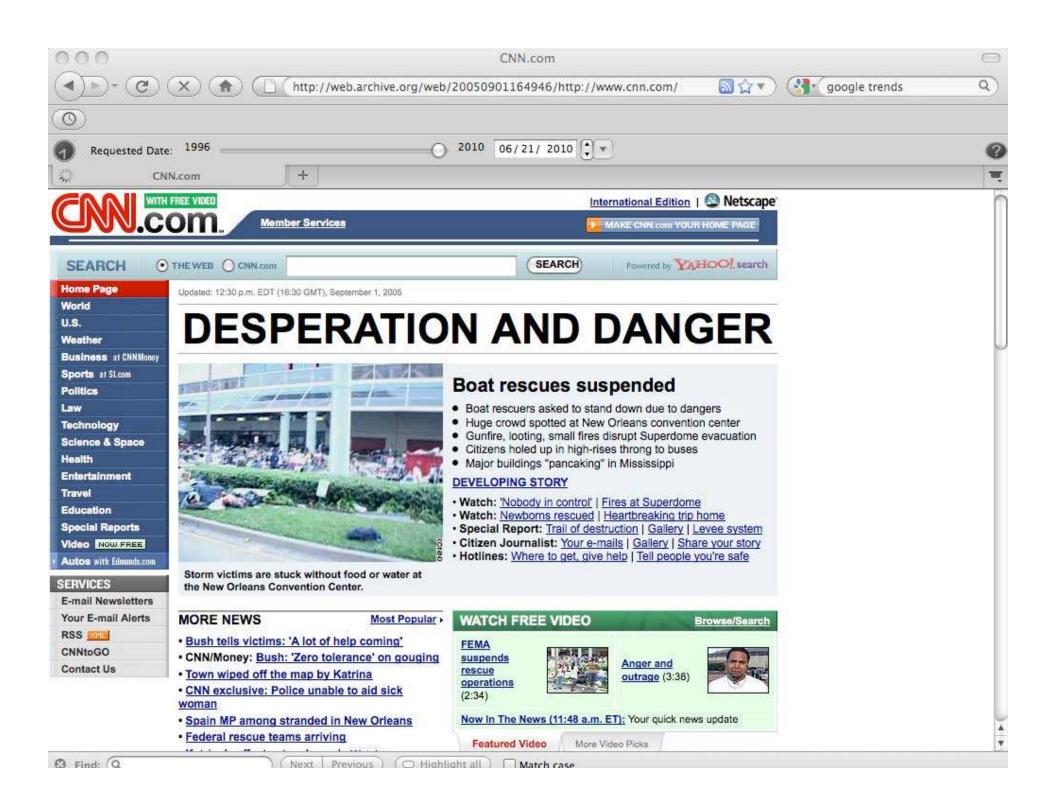


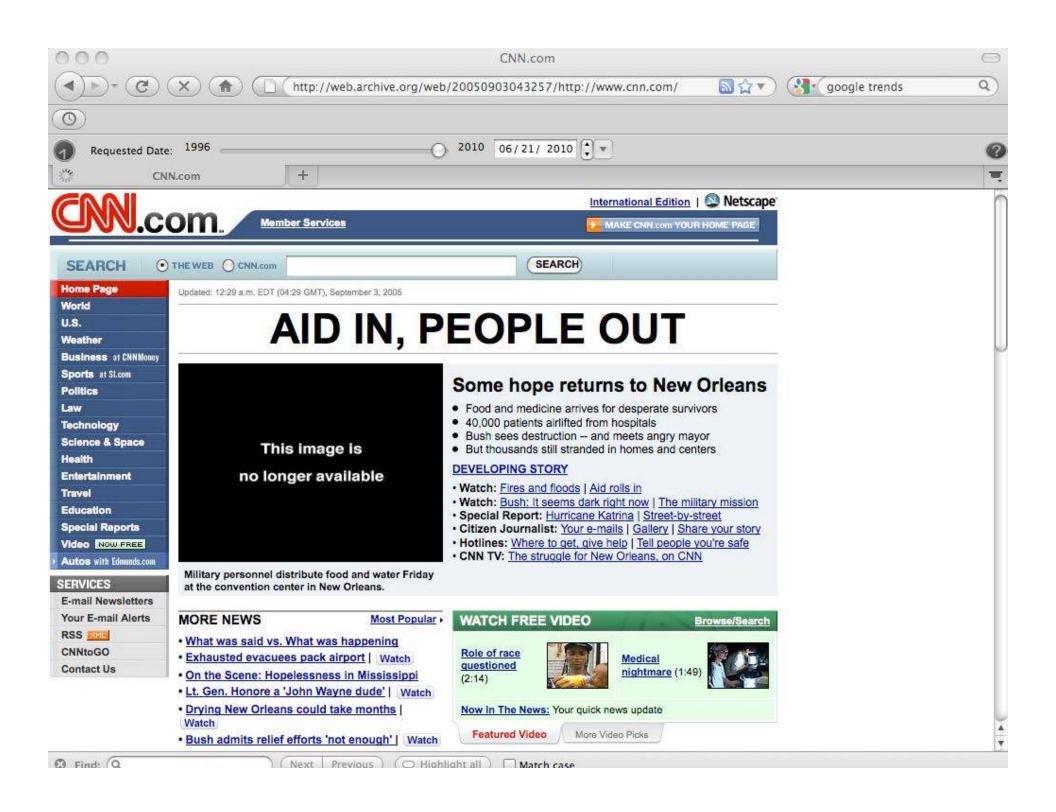


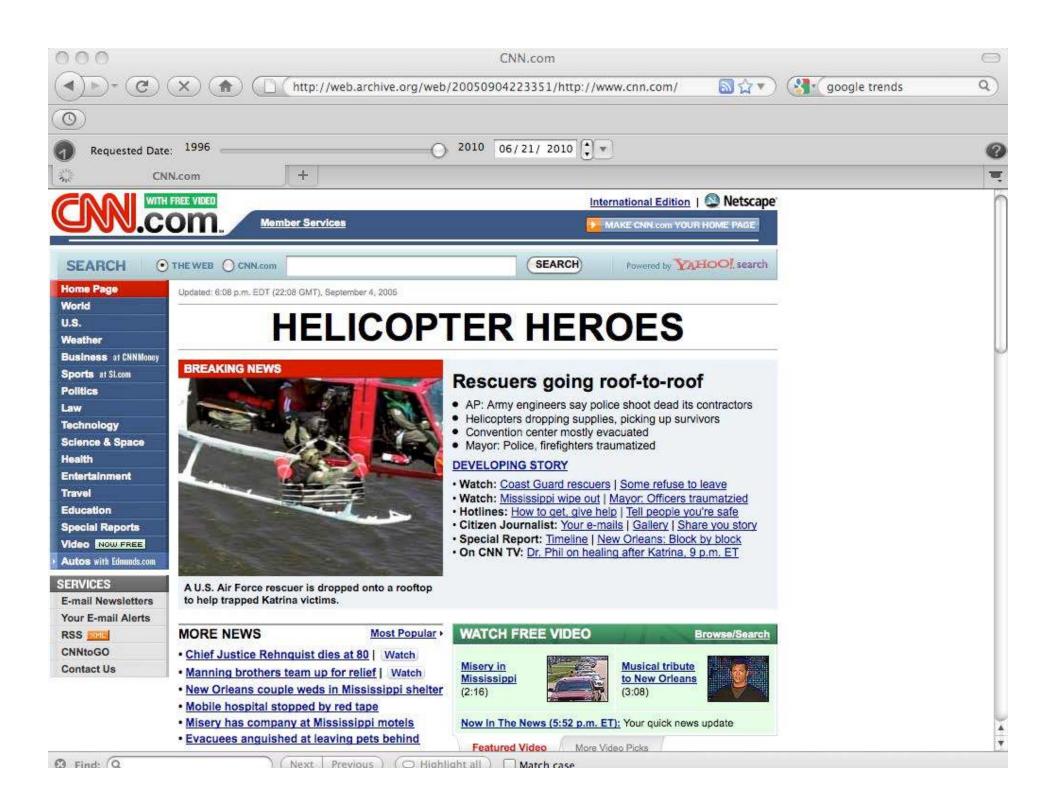












## Memento wants to make navigating the Web's Past Easy



http://www.mementoweb.org http://groups.google.com/group/memento-dev







