

9-9-2010

A Very Incomplete & Biased Review of Web Archiving

Michael L. Nelson

Old Dominion University, mnelson@odu.edu

Herbert Van de Sompel

Robert Sanderson

Frank McCown

Martin Kliein

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



Part of the [Archival Science Commons](#)

Recommended Citation

Nelson, Michael L.; de Sompel, Herbert Van; Sanderson, Robert; McCown, Frank; and Kliein, Martin, "A Very Incomplete & Biased Review of Web Archiving" (2010). *Computer Science Presentations*. 19.
https://digitalcommons.odu.edu/computerscience_presentations/19

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.

A Very Incomplete & Biased

Review of Web Archiving

Michael L. Nelson
Old Dominion University

Additional slides: Herbert Van de Sompel, Robert Sanderson,
Frank McCown, Martin Klein



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Outline

- Actors, technology, projects
- Conventional web archives
- Archives are silos
- Long tail of archives
- Memento



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Background

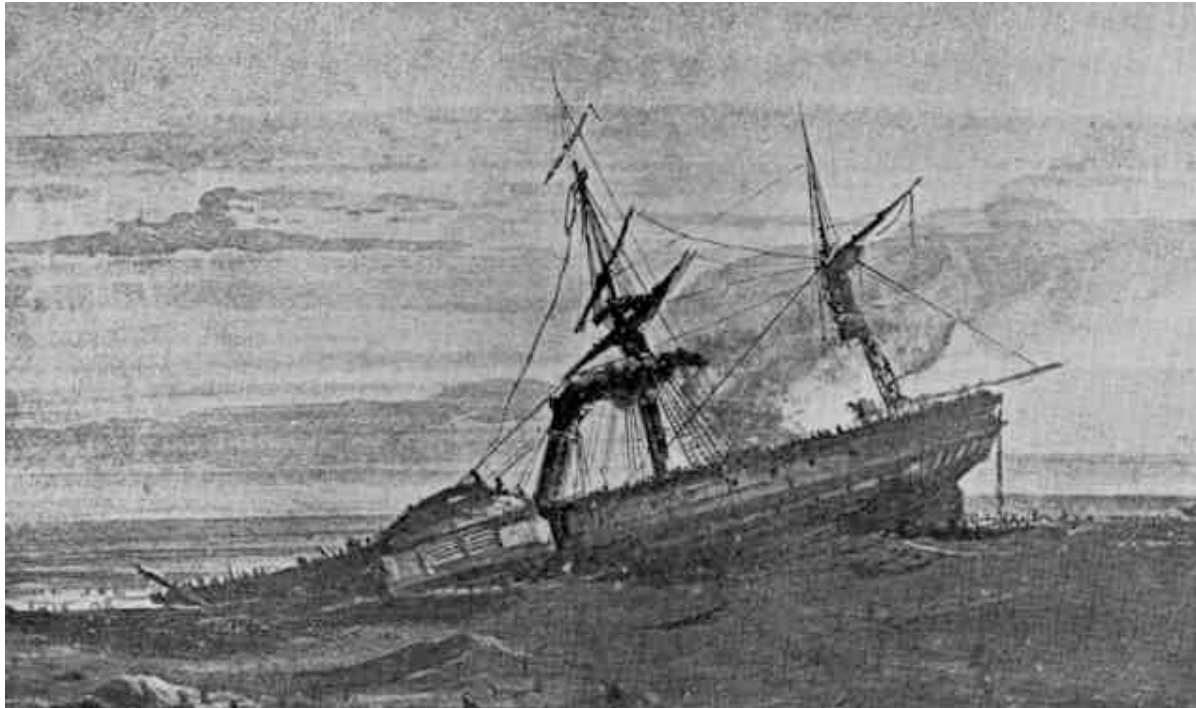
- “We can’t save everything!”
 - if not “everything”, then how much?
 - what does “save” mean?



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



“Women and Children First”



HMS Birkenhead, Cape Danger, 1852

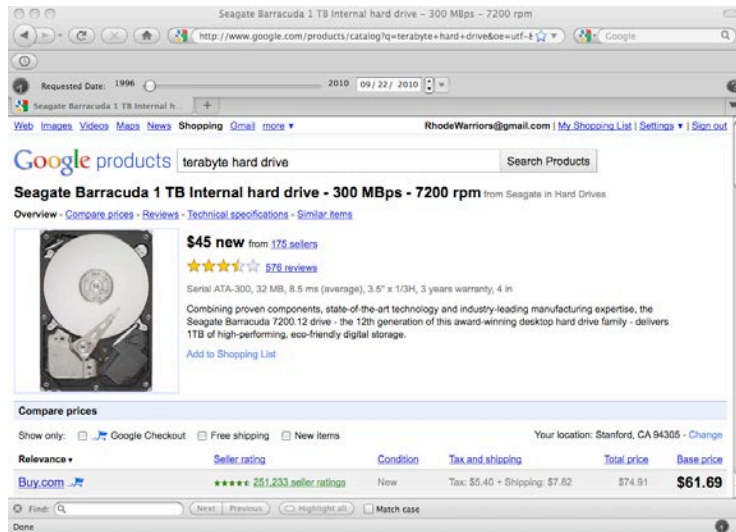
638 passengers 193 survivors all 7 women & 13 children 8 of 9 horses



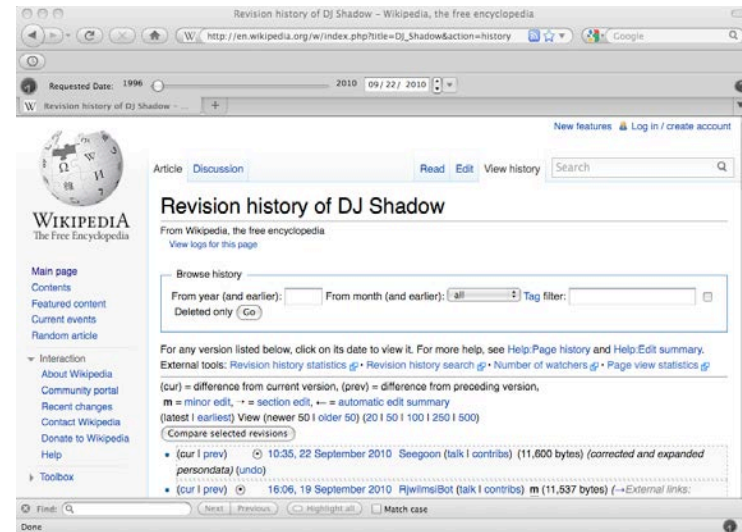
Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Time to Talk About Saving Everything?



Dinner for one or two costs more than 1TB disk



Wikis have popularized versioning

Cool URIs (<http://www.w3.org/Provider/Style/URI.html>) are widely adopted, e.g.:

http://news.yahoo.com/s/ap/20100920/ap_on_el_se/us_alaska_senate

<http://d.yimg.com/a/p/ap/20100918/capt.67567dbc0a874b689f0b4a5c392f379c-67567dbc0a874b689f0b4a5c392f379c-0.jpg>

http://d.yimg.com/a/p/afp/20100918/thumb.photo_1284846332993-1-0.jpg

Also related projects with cool URI / permalink focus:

<http://www.citability.org/>

<http://data.gov/>

<http://data.gov.uk/>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010





A Comparison of Queueing, Cluster and Distributed Computing Systems

Joseph A. Kaplan and Michael L. Nelson
Langley Research Center, Hampton, Virginia

<ftp://techreports.larc.nasa.gov/pub/techreports/larc/93/tm109025.ps.Z>
<http://techreports.larc.nasa.gov/ltrs/PDF/tm109025.pdf>

(NASA-TM-109025) A COMPARISON OF
QUEUEING, CLUSTER AND DISTRIBUTED
COMPUTING SYSTEMS (NASA, Langley
Research Center) 51 p

N94-36932

Unclass

G3/62 0015723

June 1994

National Aeronautics and
Space Administration
Langley Research Center
Hampton, Virginia 23681-0001



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Unguided Refreshing & Migrating

Google Scholar: ml nelson ja kaplan

Google Scholar BETA

ml nelson ja kaplan Search

Scholar Results 1 - 10 of about 9,270 for [ml nelson ja kaplan](#). (0.03 seconds)

12 versions found

[A comparison of queueing, cluster and distributed computing systems](#)
JA Kaplan, ML Nelson - [View as HTML](#) - [Cited by 71](#) - [Web Search](#)
 ... Joseph A. Kaplan (ja.kaplan@larc.nasa.gov) Michael L. Nelson (ml.nelson@larc.nasa.gov) NASA Langley Research Center June, 1994 Abstract ...
[NASA Technical Memorandum, 1993 - techreports.larc.nasa.gov - tworoads.net - cmpharm.ucsf.edu - phi.sinica.edu.tw - all 12 versions »](#)

[CITATION] A Comparison of Queueing
JA Kaplan, ML Nelson - [Cited by 6](#) - [Web Search](#)
 Cluster and Distributed Computing Systems, NASA Langley ..., 1994

Internet Archive Wayback Machine

Enter Web Address: <http://techreports.larc.nasa.gov/ltrs/PDF/tm109025.pdf> Take Me Back

Searched for <http://techreports.larc.nasa.gov/ltrs/PDF/tm109025.pdf> 2 Results

* denotes when site was updated.

Search Results for Jan 01, 1996 - May 03, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	1 pages	0 pages	1 pages	0 pages	0 pages	0 pages	0 pages	0 pages
		May 03, 1998 *		Sep 25, 2000 *					

2 cached PDF versions

[Home](#) | [Help](#)

Copyright © 2001, [Internet Archive](#) | [Terms of Use](#) | [Privacy Policy](#)

A Comparison of Queueing, Cluster and Distributed Computing Systems (1994) (Make Corrections) (8 citations)

Joseph A. Kaplan, Michael L. Nelson

View or download:
[fsu.edu/pub/drargo/dep/kaplan.ps.Z](#)
[kari.re.kr/NASA/larc/94...tm109025.ps.Z](#)
[nasa.gov/pub/techreport...tm109025.ps.Z](#)
 Cached: [PS.gz](#) [PS](#) [PDF](#) [Image](#) [Update](#) [Help](#)

From: [fsu.edu \(more\)](#)
 From: [kari.re.kr/ltrs/1994cit](#)
 (Enter author homepages)

(Enter summary)

Abstract: Using workstations clusters for distributed computing has become popular with the proliferation of inexpensive, powerful workstations. Workstation clusters offer both a cost effective alternative to batch processing and an easy entry into parallel computing. However, a number of workstations on a network does not constitute a cluster. Cluster management software is necessary to harness the collective computing power. In this paper, we compare a variety of cluster management and queueing...

3 remote & 4 cached versions

YAHOO! SEARCH

Web Images Video Directory Local News Product

A Comparison of Queueing, Cluster and Distributed Computing Systems

Shortcuts
[ng Systems - citeseer](#)

3 versions (2 nasa.gov & 1 mpg.de)

- [A comparison of queueing, cluster and distributed computing systems](#)
 A comparison of queueing, cluster and distributed computing systems Using workstation clusters for distributed computing has become popular with the proliferation of ...
[ntrs.nasa.gov/archive/nasa/.../19940032425_1994032425.pdf](#) - [More from this site](#)
- [A Comparison of Queueing, Cluster and Distributed Computing Systems, NASA TM-109025 \(Revision 1\), June 1994](#)
 Joseph A. Kaplan and Michael L. Nelson, A Comparison of Queueing, Cluster and Distributed Computing Systems, NASA TM-109025 (Revision 1), June 1994, pp. 60 - 69, (95KB PS, 143KB PDF) ... of cluster management and queueing systems: Computing in Distributed Networked ...
[techreports.larc.nasa.gov/ltrs/94...tm109025.refer.html](#) - 3k - [Cached](#) - [More from this site](#)
- [A Comparison of Queueing, Cluster and Distributed Computing Systems \(PDF\)](#)
 ... A Comparison of Queueing, Cluster and Distributed Computing Systems ... a variety of cluster management and queueing systems: Computing in Distributed Networked Environments (CODINE ...
[www.mpi-mainz.mpg.de/theory/general/tm109025.pdf](#) - 143k - [View as html](#) - [More from this site](#)



Review of Web Archiving: Michael L. Nelson
 Web Archiving Cooperative, Stanford, Sep 09 2010



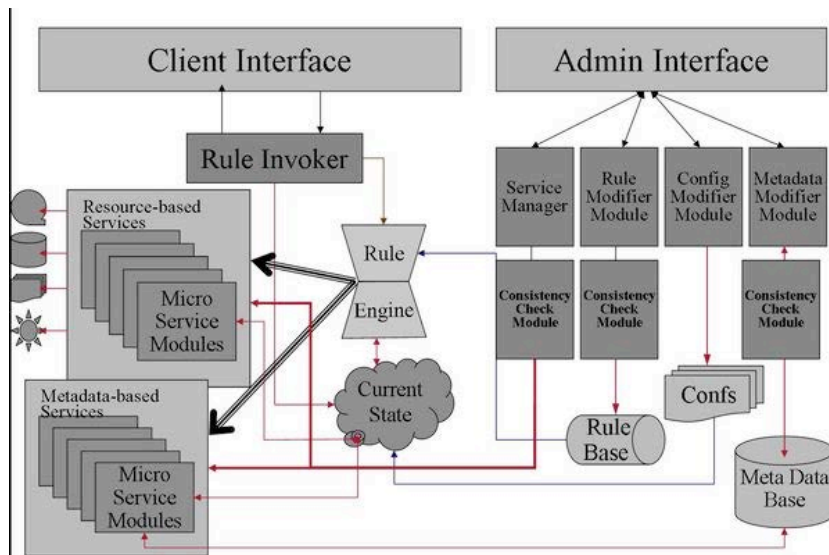
Who are the actors?



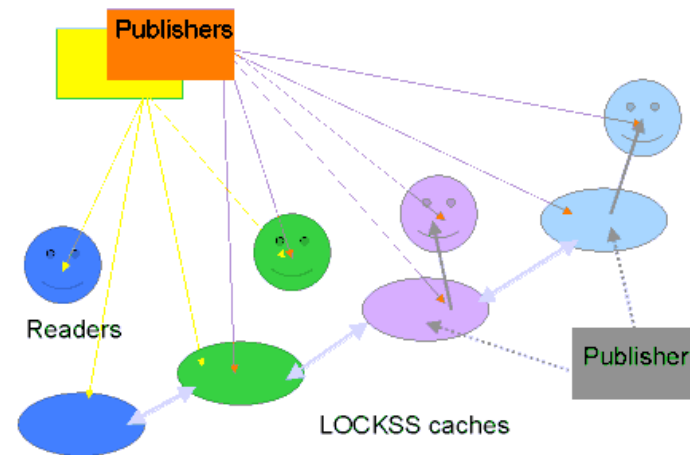
Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Archiving Frameworks



iRODS (nee SRB)
<http://www.irods.org/>



LOCKSS/CLOCKSS
<http://www.lockss.org/>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Web 2.0 Related Preservation

The screenshot shows a web browser window with two tabs. The first tab is titled "301Works.org : Free Software : Download & Streaming : Internet Archive" and displays the 301works.org website. The website has a green header with the text "Welcome to 301Works.org" and "301works.org". Below the header, there is a section titled "Spotlight Item" with the text "301WORKS.ORG" and a description of the service. There is also a section titled "About the Internet Archive" and a section titled "Rights". The second tab is titled "How Tweet It Is!: Library Acquires Entire Twitter Archive" and displays a blog post from the Library of Congress. The blog post is dated April 14th, 2010, and is by Matt Raymond. The post title is "How Tweet It Is!: Library Acquires Entire Twitter Archive". The post content includes an update about a FAQ and a description of the Twitter archive project. A photo of a man wearing a tinfoil hat and holding a cat is visible on the right side of the blog post.

<http://www.archive.org/details/301works>

<http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Conventional Web Archives

Requested Date: 1996 2010 09/22/2010

netpreserve.org
international internet preservation consortium

About:

- mission
- who we are
- member archives
- for members
- join the iipc
- working groups
- press releases

publications:
reports

events:
conferences and workshops

software:
toolkit
downloads

Member Archives

Archive List

Alphabetically (All archives are open onsite and offsite (via the Internet) unless otherwise noted after the archive name.)

Bibliothèque nationale de France - Archives de l'Internet (Bibliothèque nationale de France Web Archives) - (Access is open to researchers onsite at the institution)

- Collecting institution:** Bibliothèque nationale de France (National Library of France) - <http://www.bnf.fr>
- Start Date:** 2002
- Archive interface language(s):** French
- Access methods:** URL Search, Keyword Search, Full-Text Search, Topical Collections
- Harvesting methods:** National Domain, Bulk, Selective, Event, Thematic
- Description:** Since 2006, BnF shares with INA responsibility for the legal deposit of the French online publications and web material at large. BnF web archiving program started in 2002 with

20+ (light & dark): <http://netpreserve.org/about/archiveList.php>



web archiving service
Capture today's web * Build tomorrow's archives

WebCite



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010

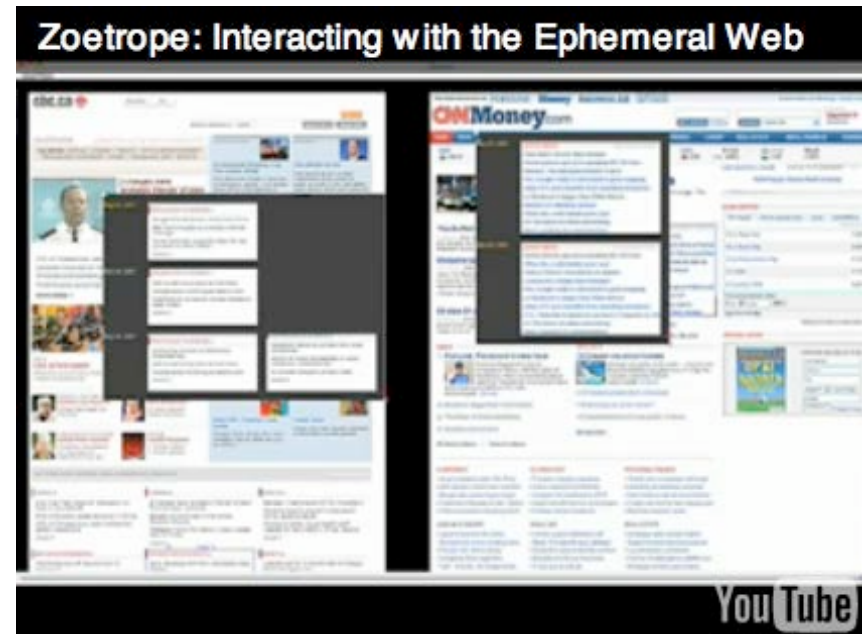


Visualization/Exploratory Services Built on Top of Archives



Past Web Browser: Adam Jatowt

<http://www.dl.kuis.kyoto-u.ac.jp/~adam/pastwebbrowser.html>



Zoetrope: Eytan Adar

<http://www.cond.org/zoetrope.html>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Tools for Batch/Site & Real-time/URI Recovery

The screenshot shows the 'Warrick - Start New Recovery' page in a Mozilla Firefox browser. The page has a blue header with the 'WARRICK' logo and links for 'Home' and 'Check on Status'. Below the header, a form is provided for recovering a lost website. The form includes fields for 'First name', 'Last name', 'Email address', and 'Website URL to recover'. There are also checkboxes for 'Recover entire website?' and 'Use Windows filenames?'. A section for 'Repositories to use' includes checkboxes for 'Internet Archive', 'Google', 'Yahoo!', and 'Live Search'. A 'Submit' button and a 'Clear' button are at the bottom of the form. A status bar at the bottom of the browser window shows 'Done'.

Warrick - Start New Recovery - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

http://warrick.cs.odu.edu/new.html

WARRICK

[Home](#) | [Check on Status](#)

In order to recover a lost website, please fill out the information below:

First name: *

Last name:

Email address: *

Website URL to recover: *

Example: <http://www.example.com/>

Recover entire website? ☒ Yes ☐ No, only recover the single page [Help](#)

Repositories to use: ** ☒ ☐ Limit dates From: To: [Help](#)

☒ ☒ ☒

Use Windows filenames? ☒ Yes ☐ No [Help](#)

* A required field.
** Only the latest resources from the Internet Archive will be returned unless dates are specified.

Done

Lazy Preservation: Frank McCown
<http://warrick.cs.odu.edu/>

The screenshot shows the 'Synchronicity' web archiving tool interface in a Google Chrome browser window. The browser's address bar shows a Google search for a cached page. The page displays the search results, indicating that the search did not match any documents. Below the search results, there are suggestions for improving the search. At the bottom of the page, there is a section for 'Synchronicity' with buttons for 'Archived Version' and 'New Version'. Below these buttons, there are tabs for 'Title Search', 'LS Search', and 'Tag Search'. A search input field is present, along with 'Retry' and 'Reset' buttons. Below the search field, there is a text area for the search query and a 'Get Title' button. At the bottom of the page, there are buttons for 'Title in Google', 'Title in Yahoo', and 'Title in Bing'. A status bar at the bottom of the browser window shows 'Done'.

cache: http://www.pspcentral.org/events/annual_meeting_2003.html - Google Search

http://webcache.googleusercontent.com/search?q=cache:http://www.pspcentral.org/events/annual_meeting_2003.html

Web Images Videos Maps News Shopping Gmail more

Google

cache: http://www.pspcentral.org/events/annual_meeting_2003.html Search [Advanced Search](#) [Preferences](#)

Web

Your search - **cache: http://www.pspcentral.org/events/annual_meeting_2003.html** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.

Synchronicity

On this screen you can choose to have Synchronicity query search engines with the title of the latest available copy of the missing page provided by the Internet Archive (IA). Synchronicity displays the title and automatically predicts its discovery performance. You can edit the title to refine the query.

Done

Just-in-Time Preservation: Martin Klein
Synchronicity



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



How do we measure success in reconstruction?

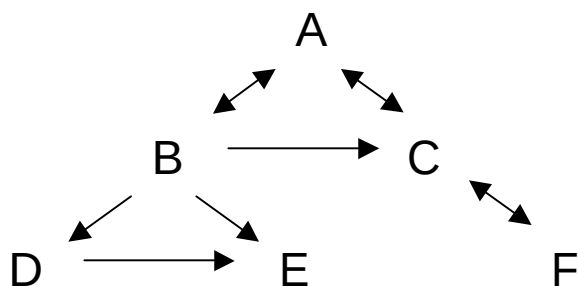


Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010

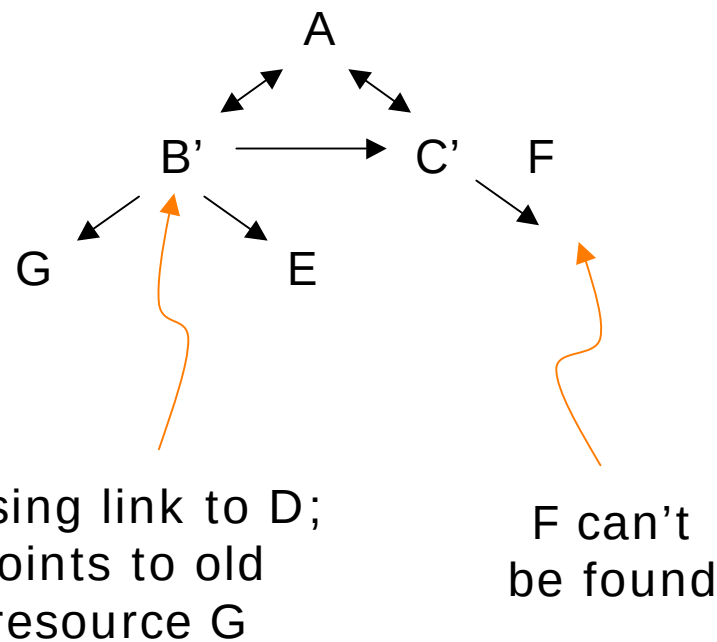


How Much Did We Reconstruct?

“Lost” web site

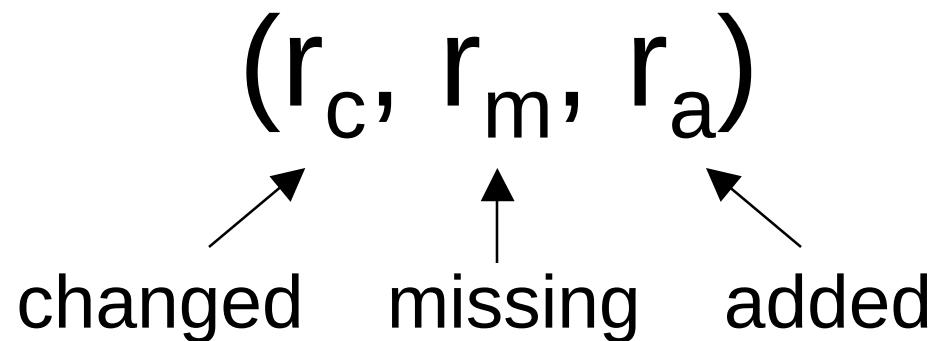


Reconstructed web site



Measuring the Difference

Apply **Recovery Vector** for each resource



Compute **Difference Vector** for website

$$\text{difference}(L, R) = \left(\frac{r_c}{|L|}, \frac{r_m}{|L|}, \frac{r_a}{|R|} \right)$$



Reconstruction Diagram

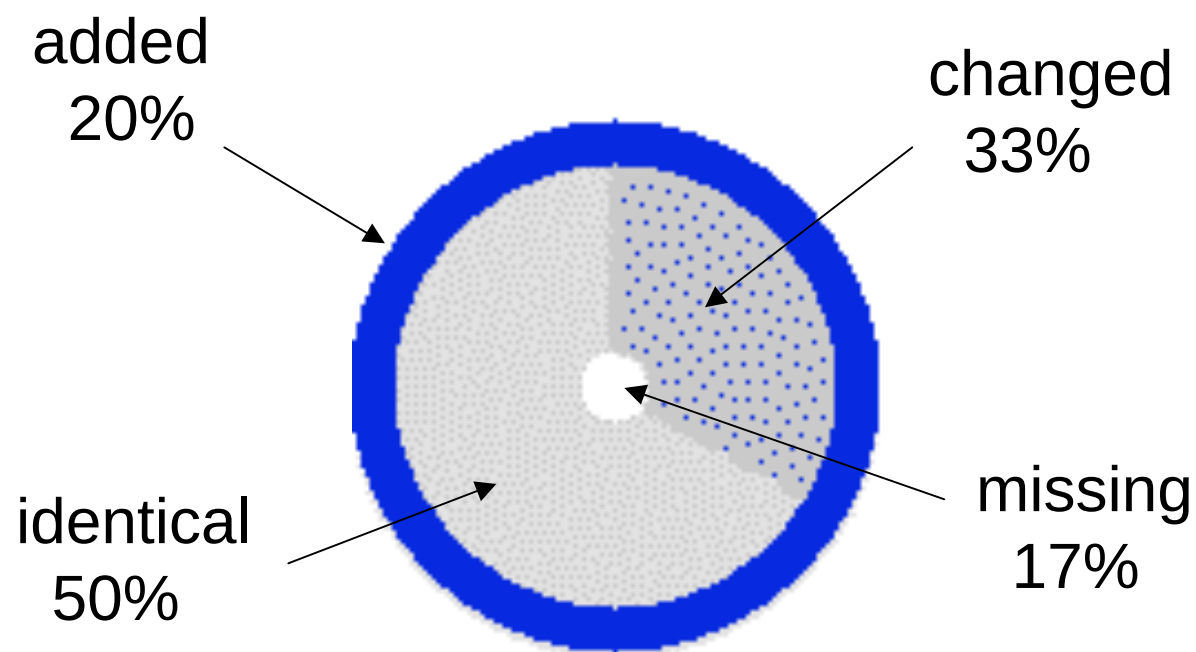


TABLE 17: Results of crawling-policy website reconstructions (Part 1).

Website	URIs	G1					
		Naïve	Knowledgeable		Exhaustive		
1. www.techlocker.com	1	(0.000, 0.000, 0.000)		(0.000, 0.000, 0.000)		(0.000, 0.000, 1.000)	
2. www.harding.edu/hr	50	(0.720, 0.220, 0.328)		(0.640, 0.280, 0.163)		(0.620, 0.240, 0.591)	
3. www.smoky.ccsd.k12.co.us	57	(0.298, 0.509, 0.000)		(0.298, 0.509, 0.000)		(0.316, 0.491, 0.970)	
4. www.genesis427.com	65	(0.508, 0.077, 0.016)		(0.523, 0.077, 0.016)		(0.538, 0.062, 0.500)	
5. englewood.k12.co.us/schools/clayton	77	(0.247, 0.286, 0.000)		(0.247, 0.286, 0.000)		(0.247, 0.286, 0.304)	
6. www.raitinvestmenttrust.com	79	(0.228, 0.127, 0.014)		(0.278, 0.253, 0.033)		(0.253, 0.127, 0.859)	
7. otago.settlers.museum	120	(0.208, 0.525, 0.017)		(0.208, 0.542, 0.286)		(0.208, 0.533, 0.341)	
8. www.usamriid.army.mil	121	(0.397, 0.413, 0.220)		(0.364, 0.512, 0.253)		(0.364, 0.471, 0.880)	
9. www.mie2005.net	136	(0.699, 0.199, 0.009)		(0.801, 0.096, 0.000)		(0.801, 0.096, 0.335)	

McCown and Nelson, **Evaluation of Crawling Policies for a Web-Repository Crawler**, HYPERTEXT 2006



Content and URIs are Orthogonal



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Lapsed Website

Internet Archive
WayBackMachine

Enter Web Address: All

Searched for <http://www.dl00.org/> 56 Results

Note some duplicates are not shown. [See all](#).
* denotes when site was updated.

Search Results for Jan 01, 1996 - May 03, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	0 pages	1 pages	12 pages	11 pages	5 pages	13 pages	11 pages	0 pages
			Oct 01, 1999 *	Jan 05, 2000 Mar 06, 2000 Mar 10, 2000 May 11, 2000 May 20, 2000 Jun 20, 2000 Jun 21, 2000 Aug 15, 2000 Aug 16, 2000 Oct 09, 2000 Oct 18, 2000 Dec 04, 2000	Jan 24, 2001 Feb 02, 2001 Feb 04, 2001 Feb 13, 2001 Mar 01, 2001 Apr 01, 2001 Apr 05, 2001 Apr 14, 2001 * Apr 21, 2001 Aug 31, 2001 Nov 27, 2001 *	Jan 18, 2002 * May 31, 2002 Jun 06, 2002 Nov 25, 2002 Nov 27, 2002	Feb 05, 2003 * May 02, 2003 * Jun 10, 2003 * Jul 30, 2003 * Aug 05, 2003 Aug 08, 2003 Sep 27, 2003 Oct 05, 2003 * Oct 07, 2003 * Dec 13, 2003 Dec 17, 2003 Dec 27, 2003 * Dec 28, 2003 *	Jan 01, 2004 * Jan 03, 2004 * Jan 30, 2004 * Apr 03, 2004 * Apr 11, 2004 * May 25, 2004 * Jun 08, 2004 * Jun 09, 2004 * Jun 10, 2004 * Jun 12, 2004 Jun 16, 2004 *	

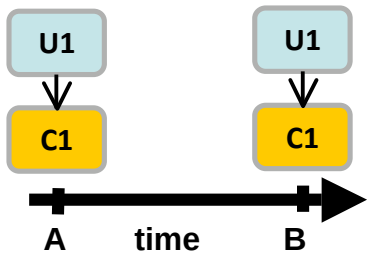
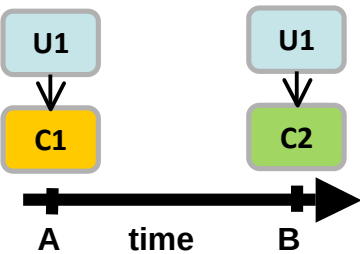
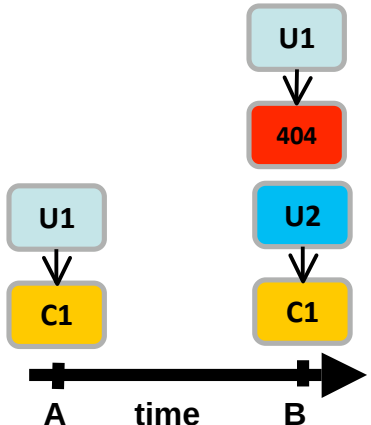
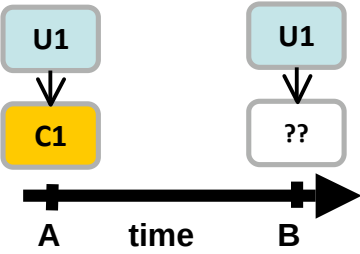
ACM DL Conference → [Oct 01, 1999](#)

Porn → [Nov 27, 2001](#)

Gambling → [Nov 25, 2002](#)

Search Engine Portal → [Jan 01, 2004](#)

URI Content Mapping Problem

1		same URI maps to same or very similar content at a later time	2		same URI maps to different content at a later time
3		different URI maps to same or very similar content at the same or at a later time	4		the content can not be found at any URI

Let's examine conventional archives in more detail



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Wayback Machine

Internet Archive Wayback Machine

Requested Date: 1996 2010 09 / 21 / 2010

Enter Web Address: All Take Me Back Adv. Search Compare Archive Pages

ched for <http://www.cnn.com/> **7783** Results

me duplicates are not shown. [See all.](#)
 es when site was updated.
 I typically becomes available here 6 months or more after collection, with some exceptions [See FAQ.](#)

Archived Results from Jan 01, 1996 - latest

1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
0 pages	0 pages	0 pages	30 pages	2734 pages	231 pages	156 pages	625 pages	1186 pages	886 pages
			Jun 20, 2000 *	Mar 01, 2001 *	Jan 23, 2002 *	Jan 29, 2003 *	Jan 01, 2004 *	Jan 01, 2005 *	Jan 01, 2006
			Jun 21, 2000	Mar 02, 2001	Feb 02, 2002	Jan 31, 2003 *	Jan 02, 2004 *	Jan 01, 2005 *	Jan 01, 2006
			Jun 21, 2000 *	Mar 05, 2001 *	Feb 09, 2002 *	Feb 01, 2003 *	Jan 03, 2004 *	Jan 02, 2005 *	Jan 02, 2006
			Jun 21, 2000 *	Mar 07, 2001 *	Feb 10, 2002 *	Feb 01, 2003 *	Jan 10, 2004 *	Jan 02, 2005 *	Jan 02, 2006
			Jun 21, 2000 *	Mar 09, 2001 *	Feb 11, 2002 *	Feb 02, 2003 *	Jan 11, 2004 *	Jan 03, 2005 *	Jan 03, 2006
			Jul 06, 2000 *	Mar 31, 2001 *	Feb 13, 2002 *	Feb 02, 2003 *	Jan 12, 2004 *	Jan 04, 2005 *	Jan 03, 2006
			Jul 07, 2000	Apr 10, 2001 *	Feb 13, 2002 *	Feb 03, 2003 *	Jan 12, 2004 *	Jan 05, 2005 *	Jan 03, 2006
			Jul 11, 2000	Apr 13, 2001 *	Feb 14, 2002 *	Feb 05, 2003 *	Jan 13, 2004 *	Jan 06, 2005 *	Jan 04, 2006
			Aug 04, 2000	Apr 13, 2001 *	Feb 15, 2002 *	Feb 06, 2003 *	Jan 15, 2004 *	Jan 06, 2005 *	Jan 04, 2006
			Aug 15, 2000	Apr 13, 2001 *	Feb 16, 2002 *	Feb 07, 2003 *	Jan 16, 2004 *	Jan 06, 2005 *	Jan 04, 2006

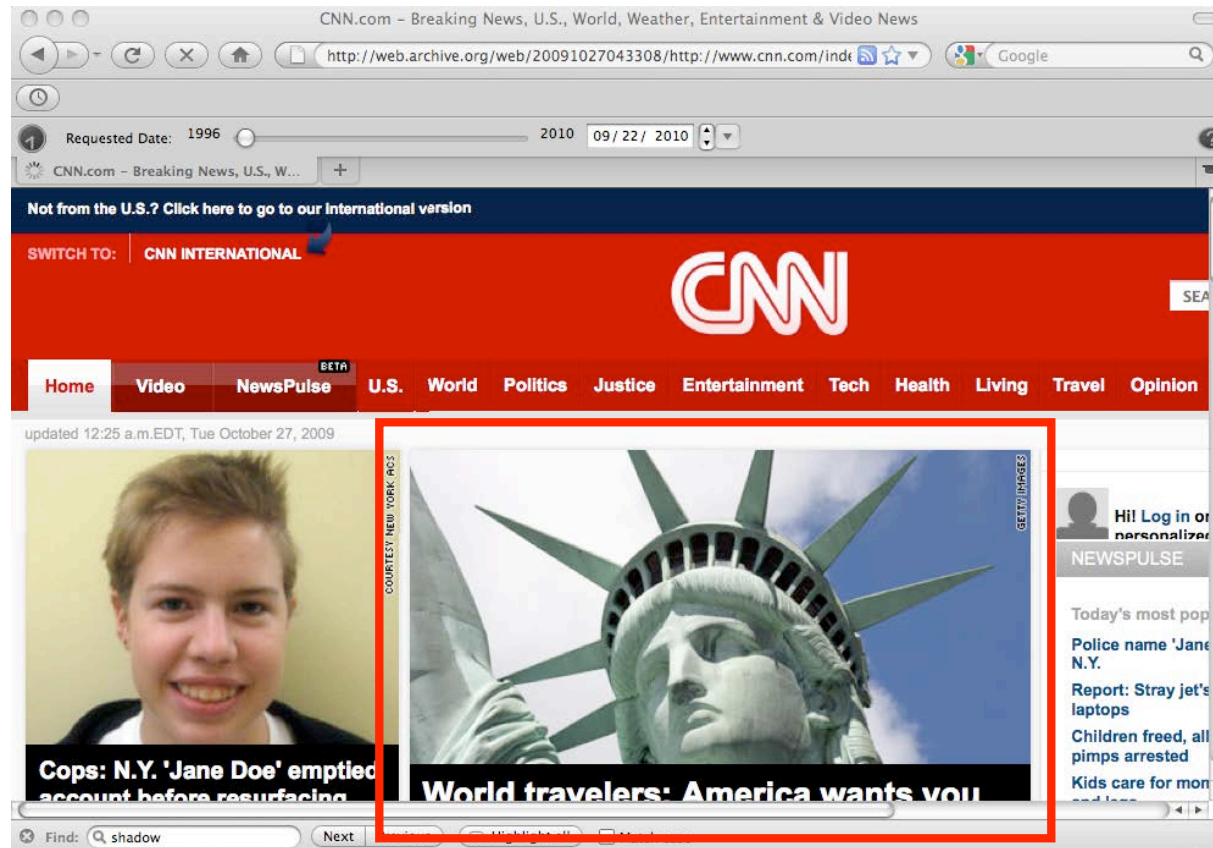
<http://web.archive.org/web/20030129185239/http://www4.cnn.com/>
<http://web.archive.org/web/20030131093102/http://cnn.com/>
<http://web.archive.org/web/20040102095249/http://www3.cnn.com/>
 etc.



Review of Web Archiving: Michael L. Nelson
 Web Archiving Cooperative, Stanford, Sep 09 2010



URI Rewriting Makes for Nice Archives



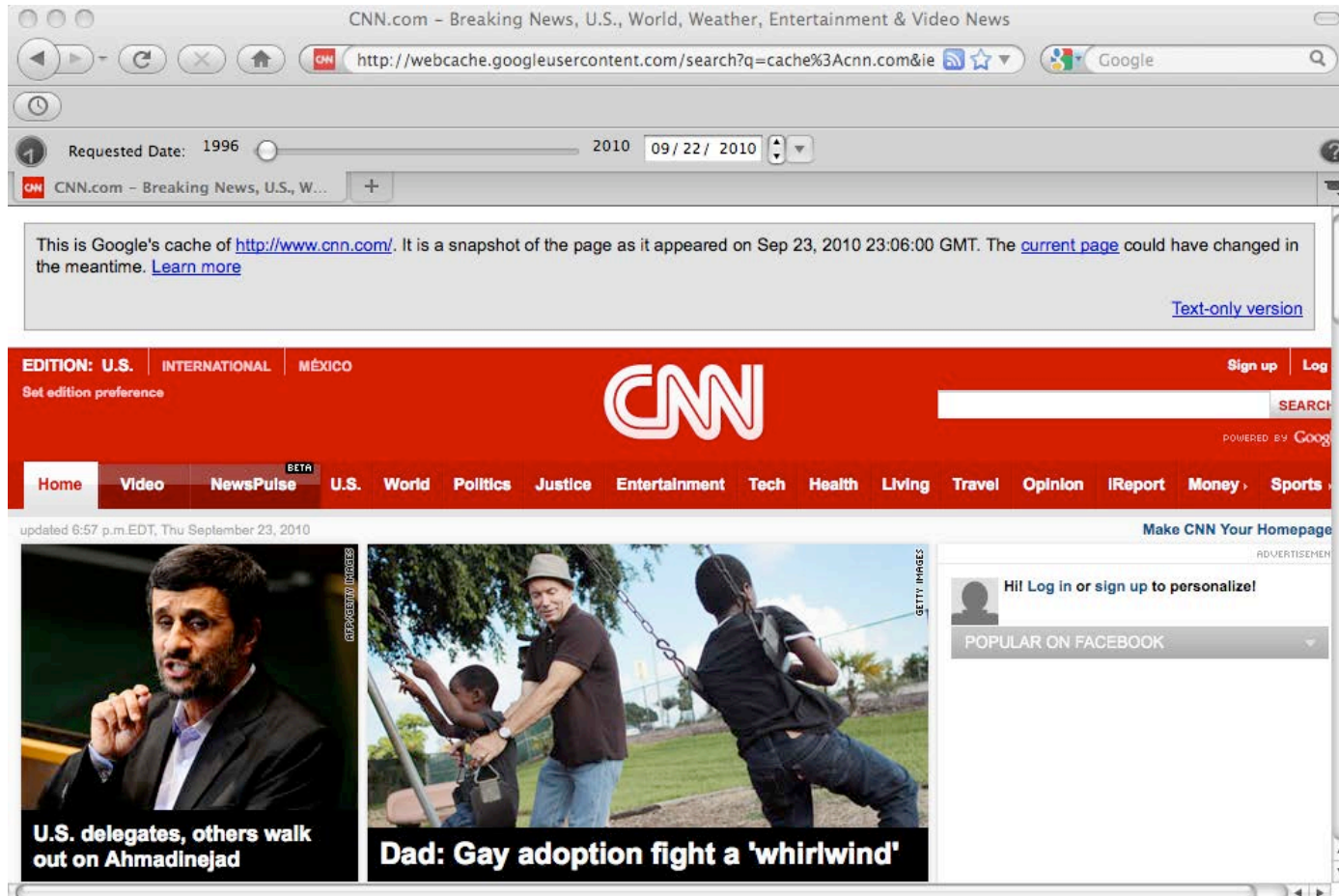
The link to: <http://i.cdn.turner.com/cnn/2009/TRAVEL/10/26/overseas.visitors.travel/c1main.liberty.gi.jpg>
using Javascript is dynamically rewritten to:
<http://web.archive.org/web/20091027043308/http://i.cdn.turner.com/cnn/2009/TRAVEL/10/26/overseas.visitors.travel/c1main.liberty.gi.jpg>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



SE Caches Do Not Rewrite URIs



Cached version of cnn.com (html only):

<http://webcache.googleusercontent.com/search?q=cache%3Acnn.com>

But images, for example, are not relative to SE cache; they're still at:

<http://i2.cdn.turner.com/cnn/2010/POLITICS/09/23/un.ahmadinejad.walkouts/t1main.ahmadinejad.afp.gi.jpg>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



URI Rewriting is Great -- Until Something Goes Wrong...

Requested Date: 1996 2010 09/22/2010

Internet Archive Wayback Machine

Enter Web Address: http:// All Take Me Back Adv. Search

Not in Archive.

No archived versions of the page you requested are available. If the page is still available on the Internet, we will begin archiving it during our next crawl.

You may want to:

- [Search for all pages](#) on the site [thecribs.com/](#)
- Try a different page address, at top

See the [FAQs](#) for more info and help, or [contact](#) us.

Find: Next Previous Highlight all Match case

Done

<http://web.archive.org/web/20080302121117/http://www.thecribs.com/>

<http://web.archive.org/web/20100923232312/http://www.thecribs.com/aa/banners/itunes.gif>



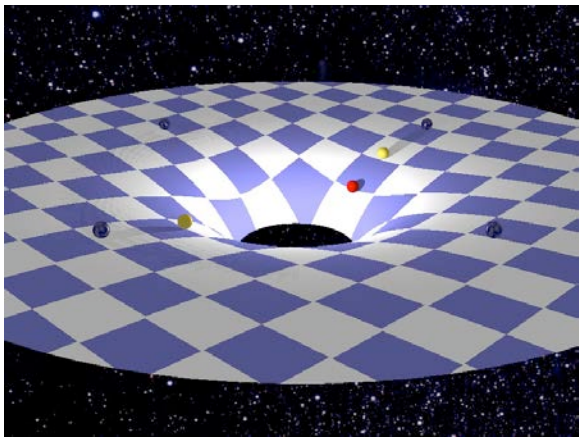
Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Where Else Could .../itunes.gif Be?



Paradox: URI rewriting makes archives useful for interactive browsing, but it actively inhibits interoperability -- your session becomes trapped in an archive



How can you escape the gravitational pull of IA's Wayback Machine and other large archives? You'd like to start an archive, but yours will never be as "good" as theirs...



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Long Tail of Archives



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Memento wants to make navigating the Web's Past Easy



<http://www.mementoweb.org>

<http://groups.google.com/group/memento-dev>



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010

29



Some more background...



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



TBL on Generic vs. Specific Resources

Web Architecture: Generic Resources

http://web.archive.org/web/19970731003253/www.w3.org/DesignIssues/Generic.html

Requested Date: 1996 2010 06/05/1997 Displayed Date: 1997-07-31

resources too, and there is a relationship of genericity between the generic and the relatively specific resource.

As an example, successively specific resources might be

1. The Bible
2. The Bible, King James Version
3. The Bible, KJV, in English
4. A particular ASCII rendering of the KJV Bible in English

http://www.w3.org/DesignIssues/Generic.html

Each resource may have a URI. The authority which allocates the URI is the authority which determines to what it refers: Therefore, that authority determines to what extent that resource is generic or specific.

This model is more of an observation of a requirement than an implementation decision. Multilevel genericity clearly exists in all our current life with books and electronic documents. Adoption of this model simply follows from the rule that Web design should not arbitrarily seek to constrain life in general for its own purposes.

Dimensions of genericity

When we discuss electronic resources, an interesting fact is that a small number of dimensions of genericity emerge.

Time	A resource may vary with time. For example, "The Wall Street Journal" varies with time. Each issue is a time-specific resource, which does not change with time. Most home pages on the Web change with time, in a less periodic way.
Language	When a document is translated, it is useful to be able to refer to it either in the generic, or to a particular specific translation.
Representation	A given resource may have many ways in which it can be represented on the wire, using different Content-types (in HTTP terms). As an example, an image may be represented in PNG or JFIF format.

The fact that there are such a small number of dimensions currently apparent suggests that Web software should handle them individually in its interface with the user, even though the architecture should handle them as a single concept.

Find: [] Next Previous Highlight all Match case

Done



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.



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



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.



Sep 11 2001, 20:36:10 UTC

Dec 20 2001, 4:51:00 UTC

Archived Resources

The screenshot shows the CNN.com website's 'SPECIAL REPORT' section for the September 11 attacks. The main headline is 'AMERICA UNDER ATTACK' with a sub-headline 'CNN EXCLUSIVE'. A large image shows the World Trade Center towers on fire. The text describes the attacks at 8:45 a.m. EDT. The page includes a 'COMPLETE COVERAGE' section with links to various news stories, a 'U.S. SCENE' section, a 'WORLD SCENE' section, and a 'RESOURCES AND RESCUE' section. There are also sections for 'VIDEO', 'PHOTO GALLERY', 'CHRONOLOGY', and 'EXTRA INFO'.

The screenshot shows the Wikipedia page for 'September 11 attacks'. The page is titled 'September 11 attacks' and includes a notice that it is an old revision from December 20, 2001. The main text describes the attacks on the World Trade Center and the Pentagon. It mentions that four passenger jets were hijacked and that the towers collapsed. The page also includes a section on the aftermath, mentioning the rescue and recovery efforts, and a section on the global impact, mentioning the 2001 U.S. Attack on Afghanistan.

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

http://en.wikipedia.org/w/index.php?title=September_11_attacks&oldid=282333 archived resource for http://en.wikipedia.org/wiki/September_11_attacks



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010

35



Finding Archived Resources

able here 6 months after collection. See FAQ.

Search Results for Jan 01, 1996 - May 14, 2009							
2000	2001	2002	2003	2004	2005	2006	2007
30 pages	2734 pages	231 pages	156 pages	625 pages	1186 pages	886 pages	953 pages
Jun 20, 2000	Mar 01, 2001	Jan 23, 2002	Jan 29, 2003	Jan 01, 2004	Jan 01, 2005	Jan 01, 2006	Jan 01, 2007
Jun 21, 2000	Mar 02, 2001	Feb 02, 2002	Jan 31, 2003	Jan 02, 2004	Jan 01, 2005	Jan 01, 2006	Jan 01, 2007
Jun 21, 2000	Mar 05, 2001	Feb 09, 2002	Feb 01, 2003	Jan 03, 2004	Jan 02, 2005	Jan 02, 2006	Jan 02, 2007
Jun 21, 2000	Mar 07, 2001	Feb 10, 2002	Feb 01, 2003	Jan 10, 2004	Jan 02, 2005	Jan 02, 2006	Jan 02, 2007
Jun 21, 2000	Mar 09, 2001	Feb 11, 2002	Feb 02, 2003	Jan 11, 2004	Jan 03, 2005	Jan 03, 2006	Jan 02, 2007
Jul 06, 2000	Mar 31, 2001	Feb 13, 2002	Feb 02, 2003	Jan 12, 2004	Jan 04, 2005	Jan 03, 2006	Jan 03, 2007
Jul 07, 2000	Apr 10, 2001	Feb 13, 2002	Feb 03, 2003	Jan 12, 2004	Jan 05, 2005	Jan 03, 2006	Jan 03, 2007
Jul 11, 2000	Apr 13, 2001	Feb 14, 2002	Feb 05, 2003	Jan 13, 2004	Jan 06, 2005	Jan 04, 2006	Jan 04, 2007
Aug 04, 2000	Apr 13, 2001	Feb 15, 2002	Feb 06, 2003	Jan 15, 2004	Jan 06, 2005	Jan 04, 2006	Jan 05, 2007
Aug 15, 2000	Apr 13, 2001	Feb 16, 2002	Feb 07, 2003	Jan 16, 2004	Jan 06, 2005	Jan 04, 2006	Jan 05, 2007
Aug 15, 2000	Apr 14, 2001	Feb 20, 2002	Feb 10, 2003	Jan 17, 2004	Jan 07, 2005	Jan 04, 2006	Jan 06, 2007
Aug 15, 2000	Apr 14, 2001	Feb 22, 2002	Feb 10, 2003	Jan 20, 2004	Jan 07, 2005	Jan 05, 2006	Jan 06, 2007
Aug 15, 2000	Apr 14, 2001	Feb 22, 2002	Feb 11, 2003	Jan 20, 2004	Jan 08, 2005	Jan 05, 2006	Jan 06, 2007
Aug 16, 2000	Apr 18, 2001	Feb 23, 2002	Feb 12, 2003	Jan 20, 2004	Jan 09, 2005	Jan 05, 2006	Jan 07, 2007
Aug 16, 2000	May 03, 2001	Feb 23, 2002	Feb 12, 2003	Jan 24, 2004	Jan 09, 2005	Jan 06, 2006	Jan 07, 2007
Aug 17, 2000	May 03, 2001	Feb 23, 2002	Jan 27, 2003	Jan 27, 2004	Jan 10, 2005	Jan 06, 2006	Jan 07, 2007
Aug 17, 2000	May 03, 2001	Feb 24, 2002	Feb 15, 2003	Jan 28, 2004	Jan 10, 2005	Jan 06, 2006	Jan 07, 2007
Aug 18, 2000	May 03, 2001	Feb 26, 2002	Mar 20, 2003	Jan 30, 2004	Jan 10, 2005	Jan 06, 2006	Jan 07, 2007
Aug 28, 2000	May 03, 2001	Feb 27, 2002	Mar 21, 2003	Jan 30, 2004	Jan 11, 2005	Jan 07, 2006	Jan 08, 2007
Sep 20, 2000	May 03, 2001	Mar 27, 2002	Mar 21, 2003	Jan 31, 2004	Jan 11, 2005	Jan 07, 2006	Jan 08, 2007
Sep 30, 2000	May 03, 2001	Mar 29, 2002	Mar 24, 2003	Feb 02, 2004	Jan 12, 2005	Jan 07, 2006	Jan 08, 2007
Oct 18, 2000	May 04, 2001	May 25, 2002	Mar 25, 2003	Feb 02, 2004	Jan 12, 2005	Jan 07, 2006	Jan 08, 2007
Nov 09, 2000	May 04, 2001	May 27, 2002	Mar 26, 2003	Feb 03, 2004	Jan 13, 2005	Jan 08, 2006	Jan 08, 2007
Nov 10, 2000	May 04, 2001	May 27, 2002	Mar 29, 2003	Feb 04, 2004	Jan 13, 2005	Jan 09, 2006	Jan 08, 2007
Nov 10, 2000	May 04, 2001	May 28, 2002	Mar 29, 2003	Feb 07, 2004	Jan 13, 2005	Jan 09, 2006	Jan 08, 2007
Nov 17, 2000	May 04, 2001	May 29, 2002	Mar 30, 2003	Feb 10, 2004	Jan 14, 2005	Jan 10, 2006	Jan 08, 2007
Nov 18, 2000	May 04, 2001	May 30, 2002	Mar 31, 2003	Feb 10, 2004	Jan 14, 2005	Jan 10, 2006	Jan 09, 2007
Dec 04, 2000	May 04, 2001	May 30, 2002	Apr 01, 2003	Feb 11, 2004	Jan 14, 2005	Jan 10, 2006	Jan 09, 2007
Dec 04, 2000	May 04, 2001	May 30, 2002	Apr 04, 2003	Feb 13, 2004	Jan 15, 2005	Jan 10, 2006	Jan 09, 2007
Dec 04, 2000	May 05, 2001	May 31, 2002	Apr 04, 2003	Feb 13, 2004	Jan 15, 2005	Jan 10, 2006	Jan 09, 2007
	May 05, 2001	Jun 02, 2002	Apr 05, 2003	Feb 14, 2004	Jan 15, 2005	Jan 10, 2006	Jan 10, 2007
	May 05, 2001	Jun 03, 2002	Apr 08, 2003	Feb 14, 2004	Jan 16, 2005	Jan 10, 2006	Jan 10, 2007
	May 05, 2001	Jun 03, 2002	Apr 09, 2003	Feb 17, 2004	Jan 16, 2005	Jan 11, 2006	Jan 11, 2007
	May 05, 2001	Jun 04, 2002	Apr 09, 2003	Feb 20, 2004	Jan 17, 2005	Jan 11, 2006	Jan 12, 2007
	May 05, 2001	Jun 06, 2002	Apr 11, 2003	Feb 21, 2004	Jan 18, 2005	Jan 11, 2006	Jan 12, 2007
	May 05, 2001	Jul 01, 2002	Apr 20, 2003	Feb 28, 2004	Jan 18, 2005	Jan 11, 2006	Jan 12, 2007
	May 05, 2001	Jul 02, 2002	Apr 23, 2003	Mar 02, 2004	Jan 19, 2005	Jan 11, 2006	Jan 12, 2007
	May 05, 2001	Jul 03, 2002	Apr 23, 2003	Mar 07, 2004	Jan 19, 2005	Jan 12, 2006	Jan 13, 2007

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

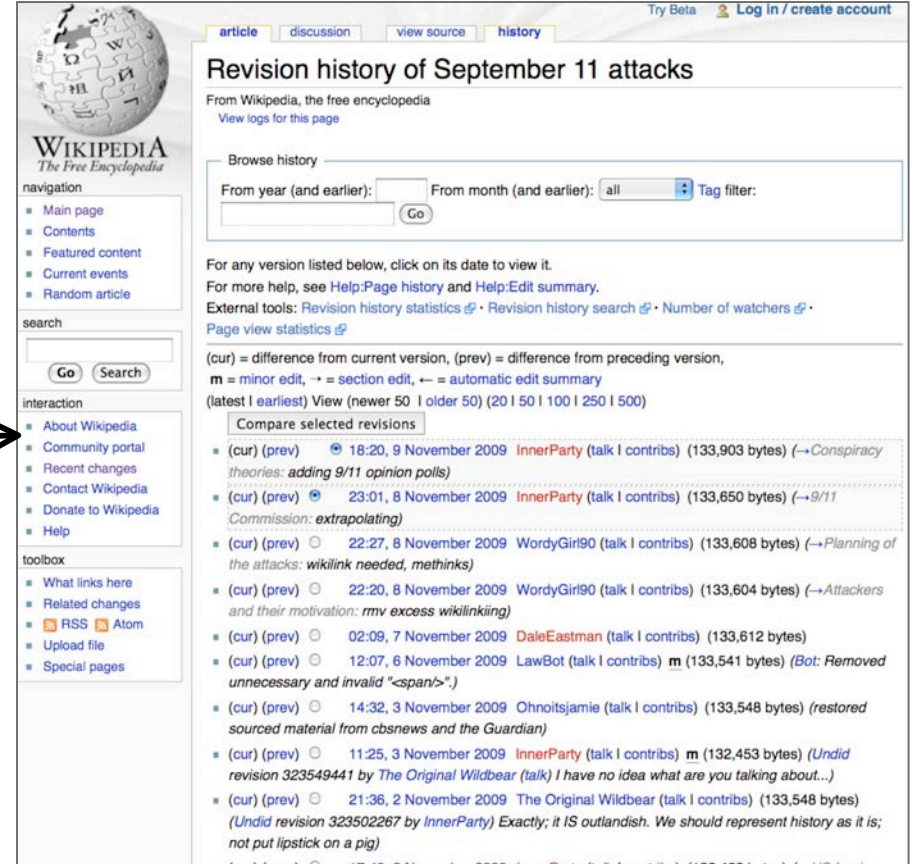
On http://web.archive.org/web/*/http://cnn.com, select desired datetime



Finding Archived Resources



The screenshot shows the Wikipedia article for "September 11 attacks". The article text describes the coordinated suicide attacks by Al-Qaeda on September 11, 2001, in New York City and Arlington, Virginia. It mentions the hijacking of four commercial passenger jet airliners, the collapse of the Twin Towers of the World Trade Center, and the crash of a third airliner into the Pentagon. The article also notes the death of 2,995 people, including the hijackers, and the overwhelming majority of casualties were civilians. It mentions the United States' response, including the "War on Terrorism" and the USA PATRIOT Act.



The screenshot shows the "Revision history of September 11 attacks" page on Wikipedia. It displays a list of revisions, including the current version (cur) and previous versions (prev). The list includes the date, time, and user for each revision, along with the size of the revision in bytes. The current version is 18:20, 9 November 2009 by InnerParty (133,903 bytes). The previous version is 23:01, 8 November 2009 by InnerParty (133,650 bytes). The list also includes a "Compare selected revisions" button and a "Browse history" button.

Go to

http://en.wikipedia.org/wiki/September_11_attacks
and click History

Browse History



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010

37



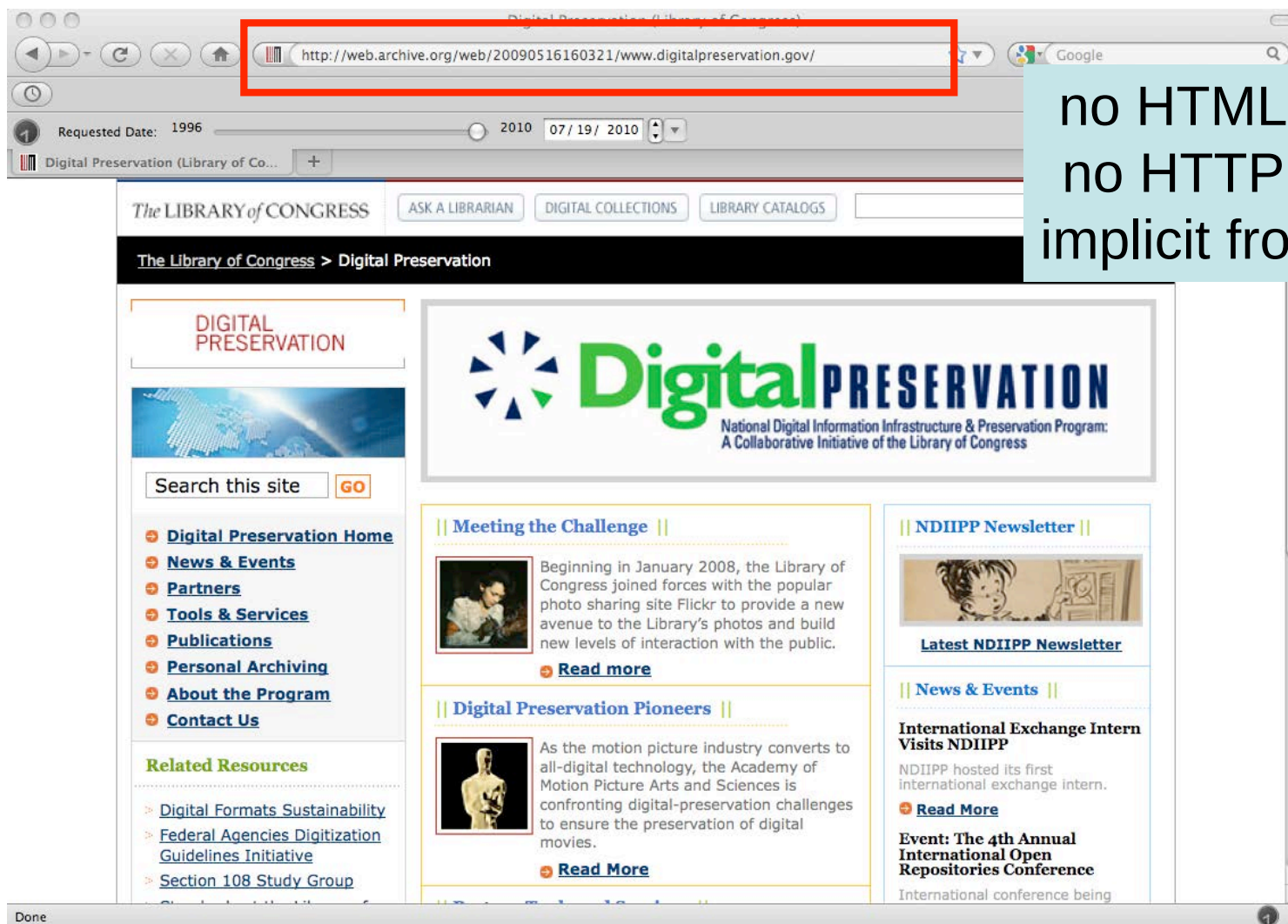
The Past Links to the Present...

The screenshot shows a WebCite query result page. The browser's address bar is highlighted with a red box, showing the URL: <http://webcitation.org/query?id=1257028234035091>. Below the address bar, a date range is set from 1996 to 2010, with a specific date of 07/19/2010 selected. The WebCite logo is visible, along with a link to "What's this?". A red box highlights the text: "Showing WebCite for URL: <http://www.digitalpreservation.gov/> [Permalink to this cache]". Below this, the cache date is shown as "Cached: 2009-10-31 18:30:35". The main content area shows the "Digital Preservation" page from the Library of Congress, with a search bar and a list of links including "Digital Preservation Home", "News & Events", "Partners", "Tools & Services", "Publications", "Video Presentations", "Personal Archiving", "About the Program", and "Contact Us". A "Related Resources" section is also present, listing "Digital Formats Sustainability", "Federal Agencies Digitization Guidelines Initiative", "Section 108 Study Group", and "Standards at the Library of Congress".

explicit HTML link;
no HTTP links;
opaque URI



The Past Links to the Present...



no HTML links;
no HTTP links;
implicit from URI



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



But the Present Does Not Link to the Past

Browser window showing the Digital Preservation (Library of Congress) website. The address bar displays `http://www.digitalpreservation.gov/`. The page content includes the Library of Congress logo, navigation links, and a search bar. The main content area features a large "Digital Preservation" header and several article teasers.

no hints in HTML, HTTP, or URI

```
% 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.
```



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



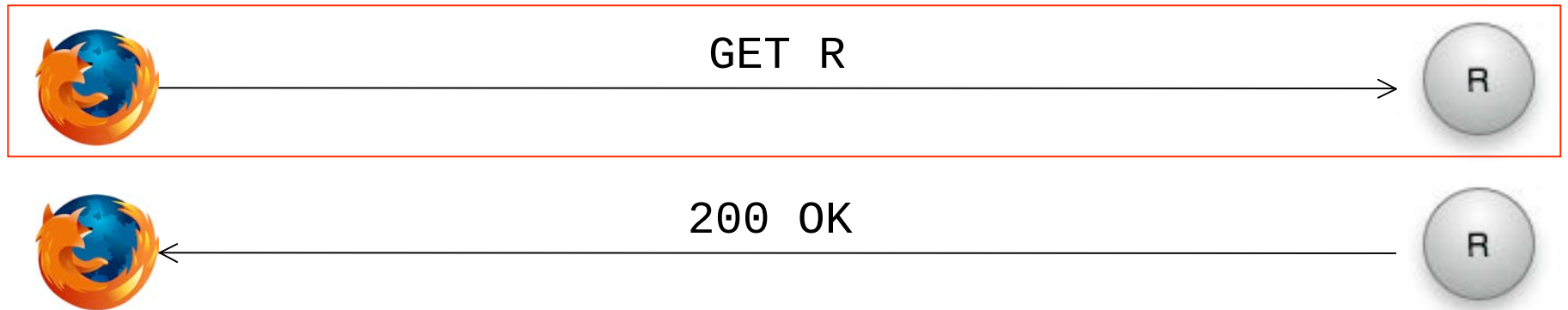
How does Memento do This?

There are two components to the Memento Solution:

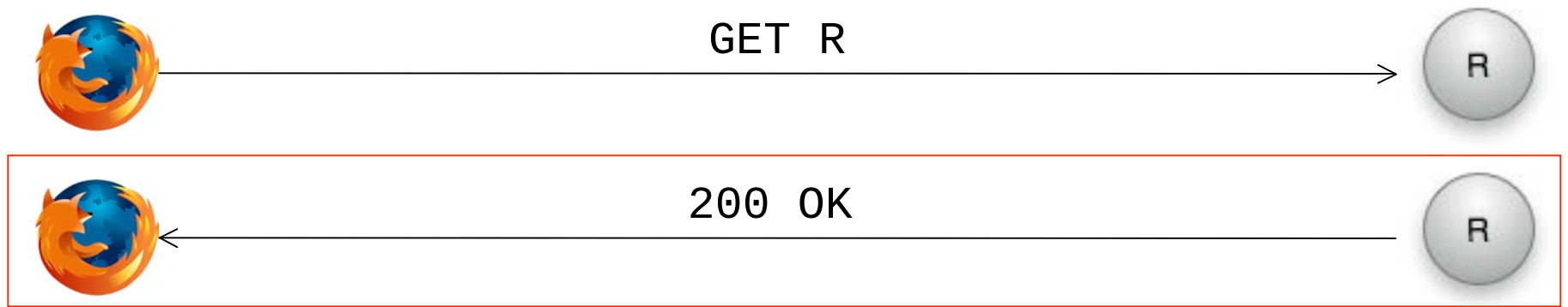
- *Component 1: Navigation towards an archived resource via its original resource, by leveraging content negotiation.*
- *Component 2: A discovery API for archives that allows requesting a list of all archived versions it holds for a resource with a given URI.*



Normal HTTP Flow



Normal HTTP Flow



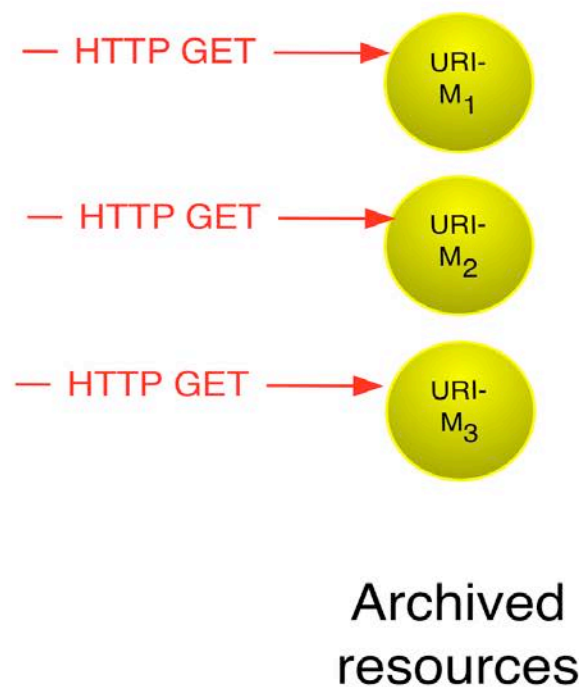
The Web without a Time Dimension



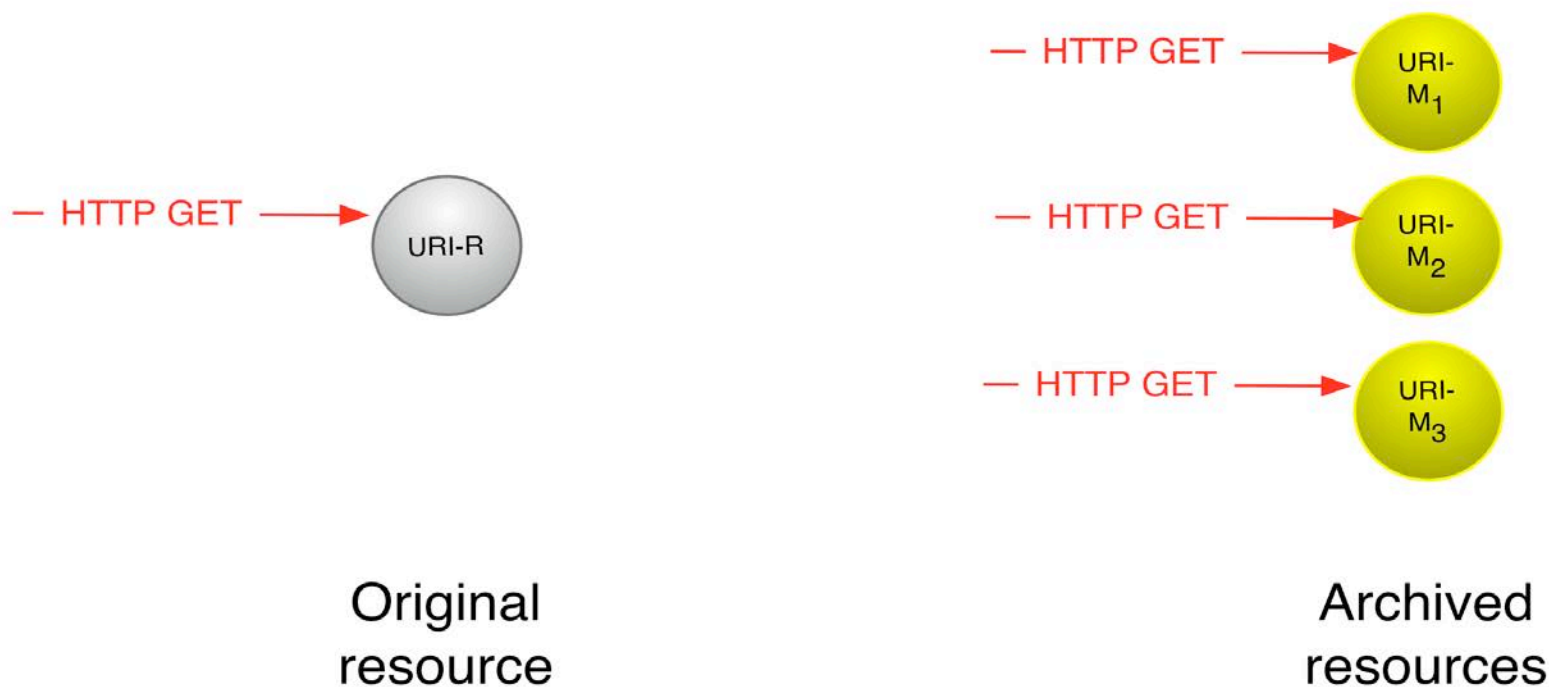
Original
resource



The Web without a Time Dimension



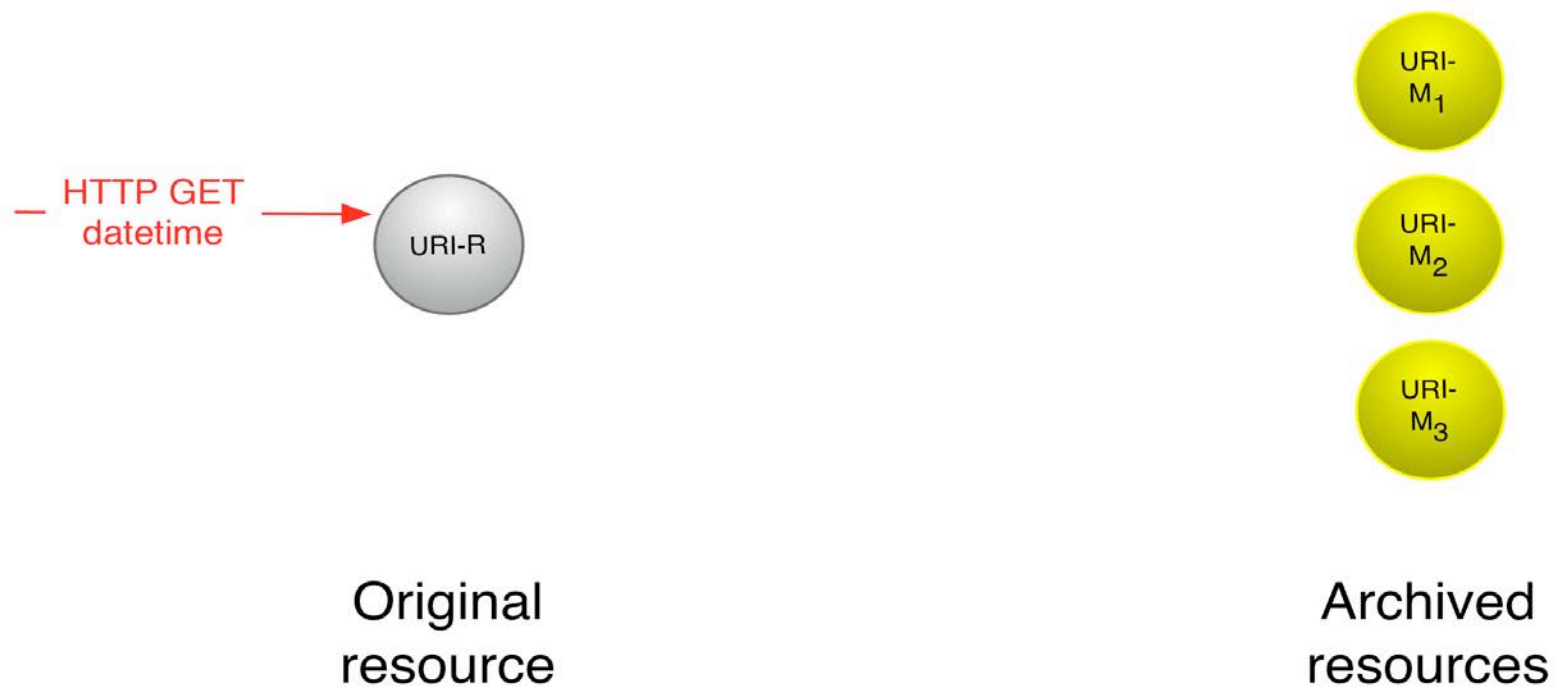
The Web without a Time Dimension



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



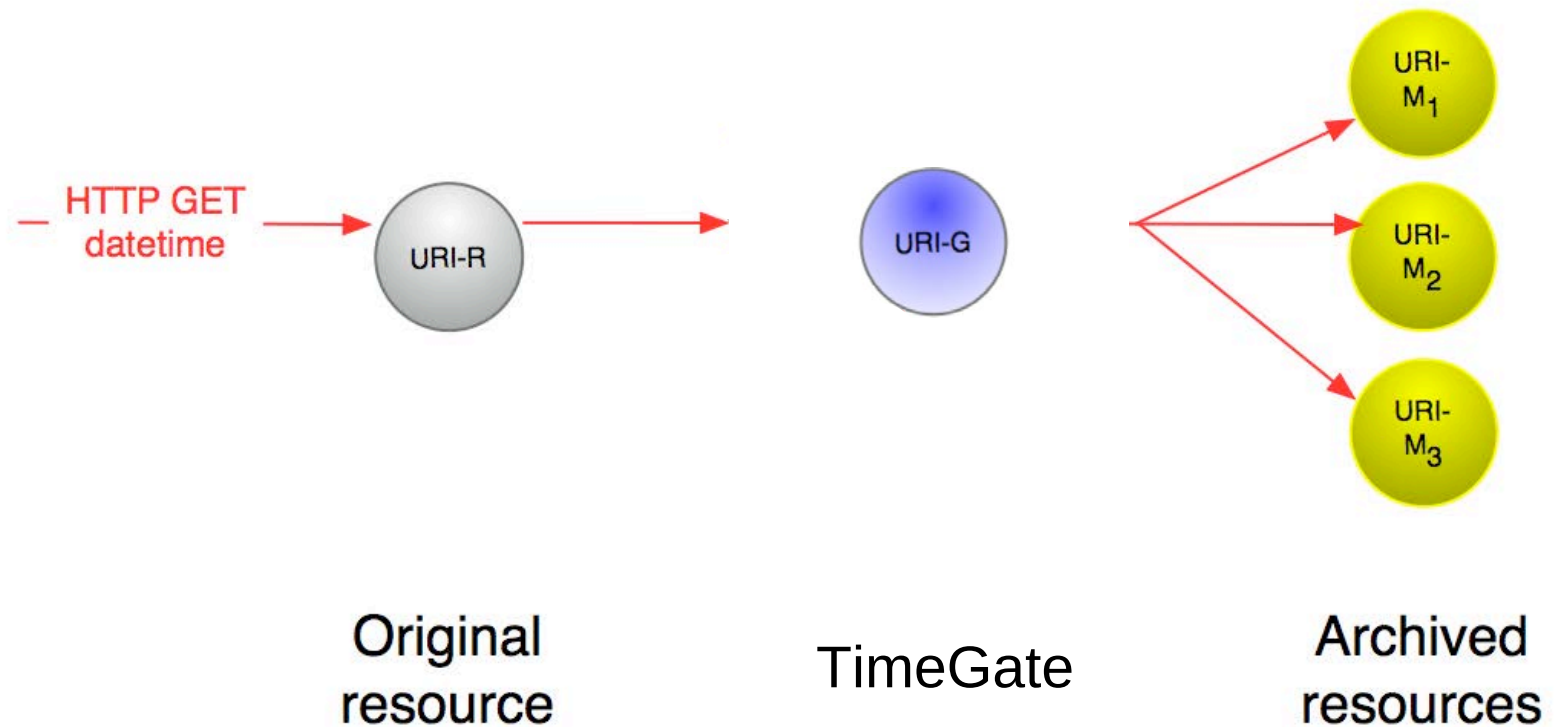
The Web with Time Dimension added by Memento



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

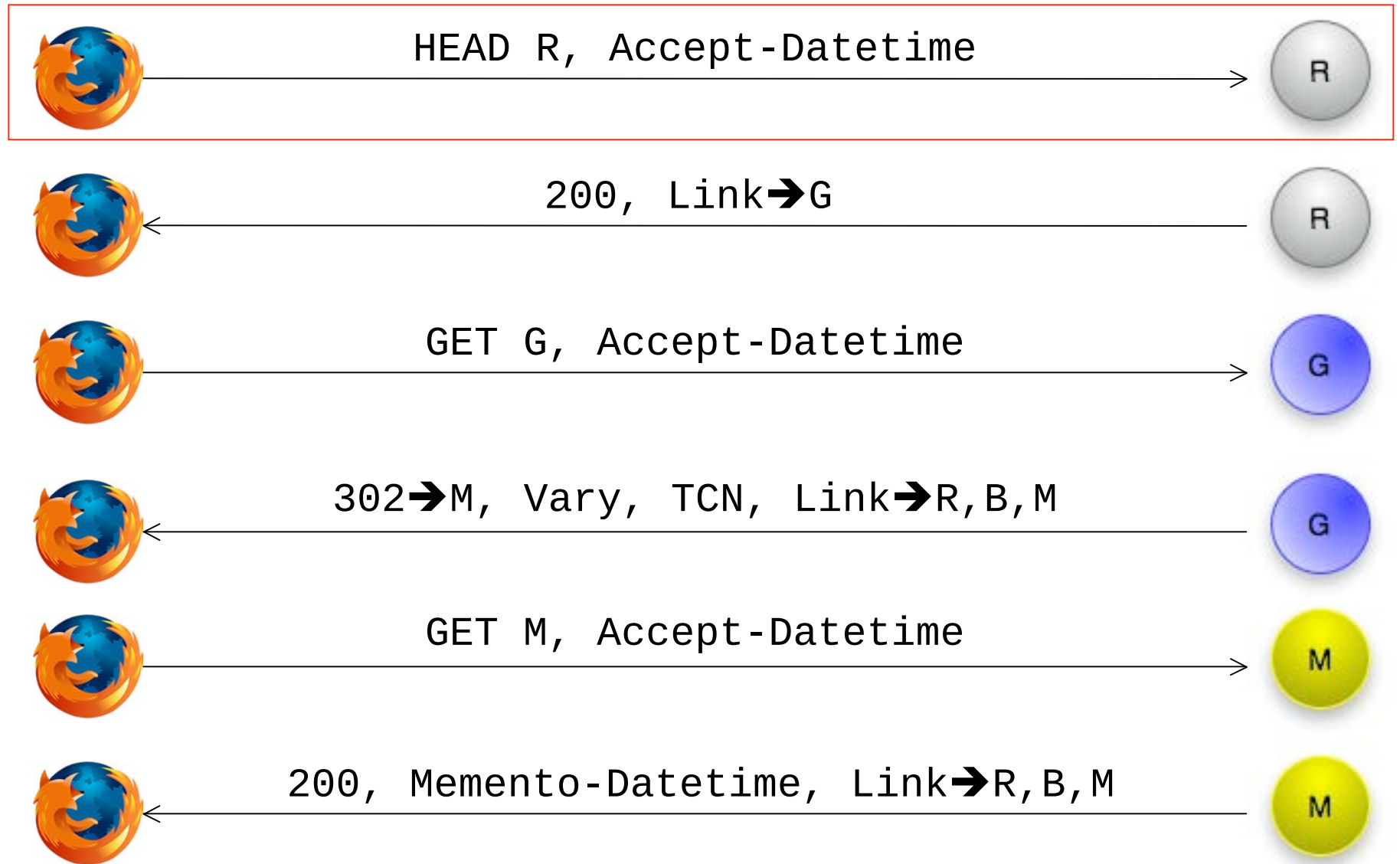


Content Negotiation in the datetime dimension

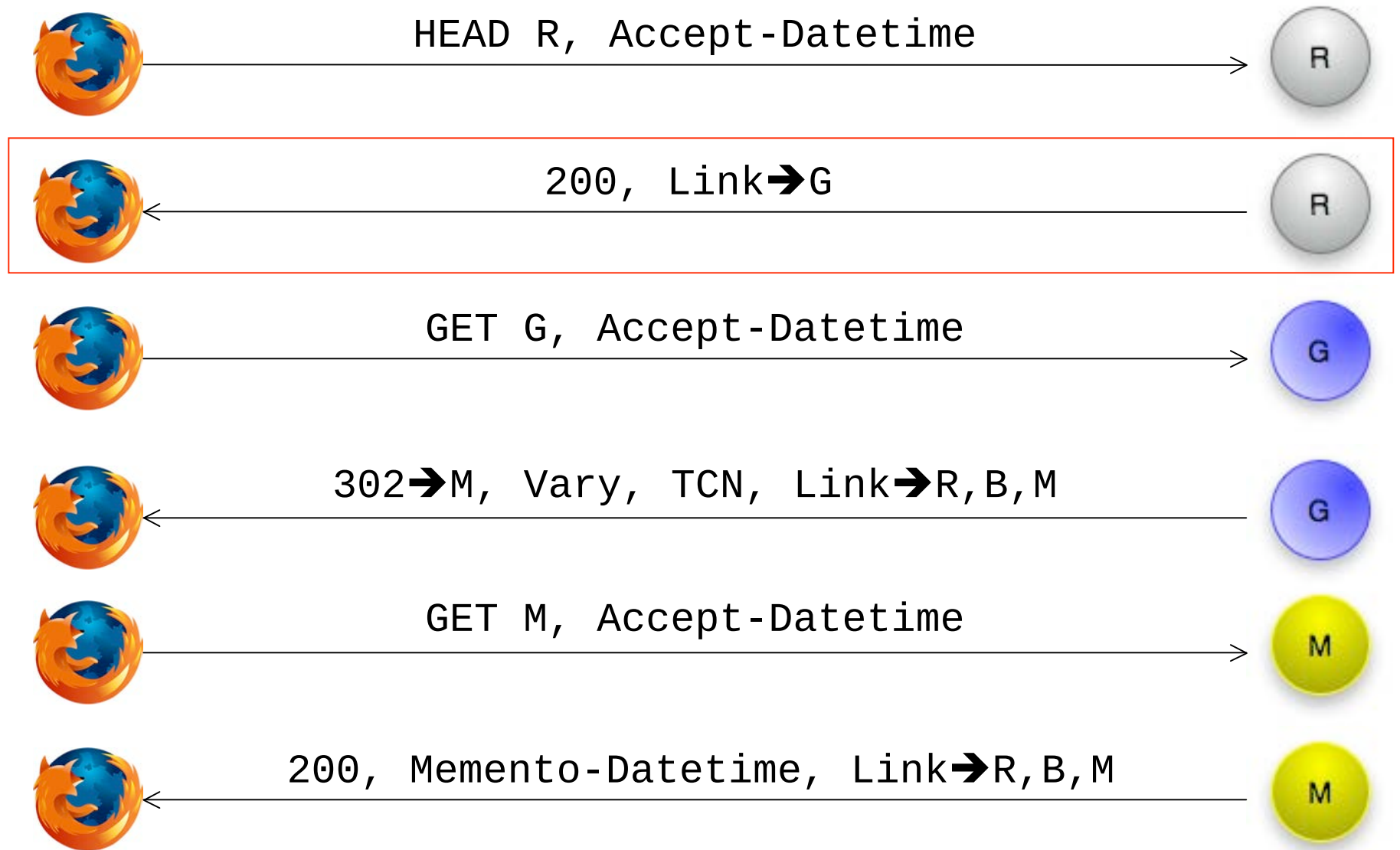
- Many systems support content negotiation for media type:
 - Your client by default asks for HTML and gets HTML;
 - But it could get PDF *via the same URI*.
- Memento proposes a new dimension for content negotiation – time:
 - Your client by default asks for the current time, and gets it
 - But it could 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



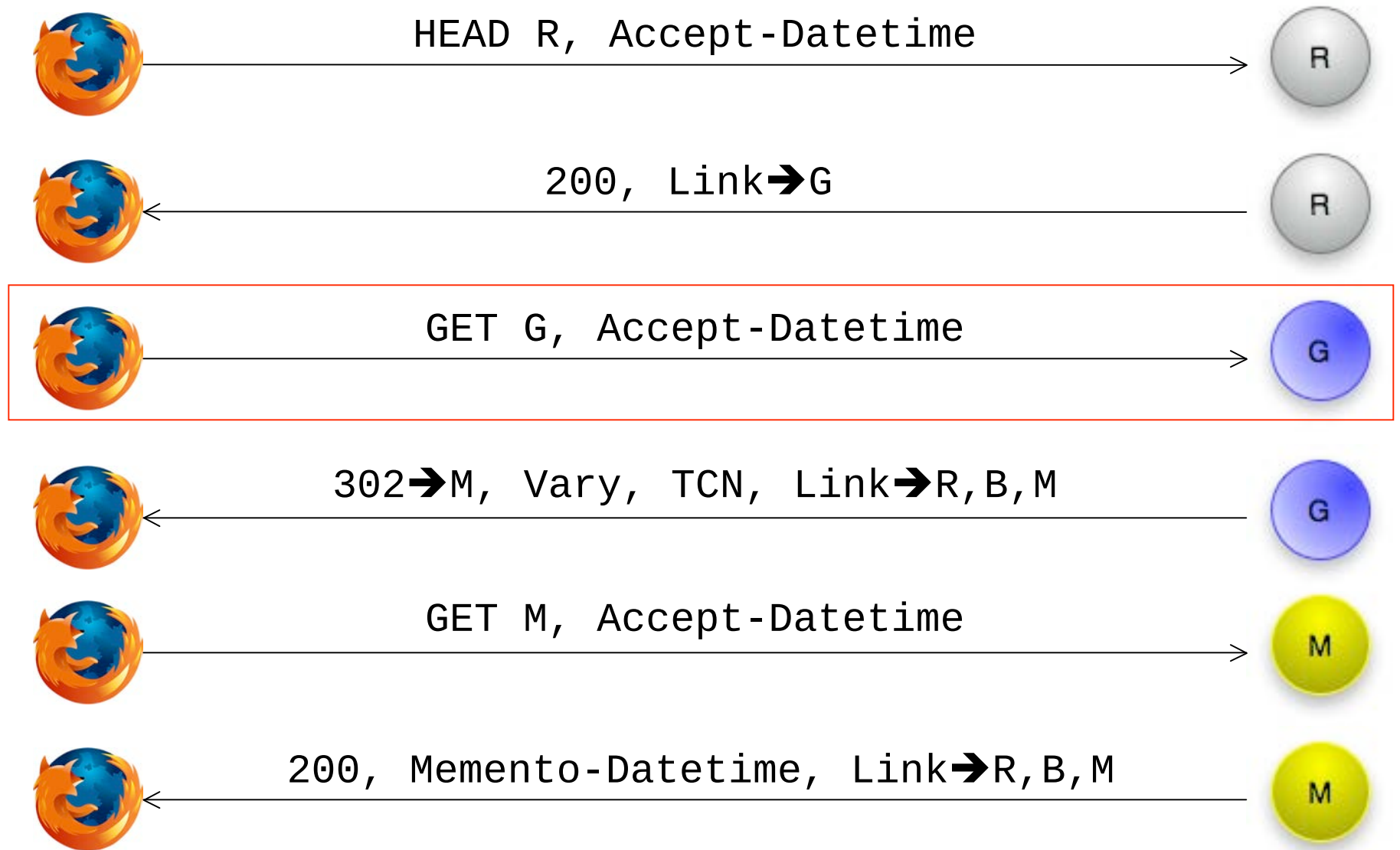
Memento HTTP Flow



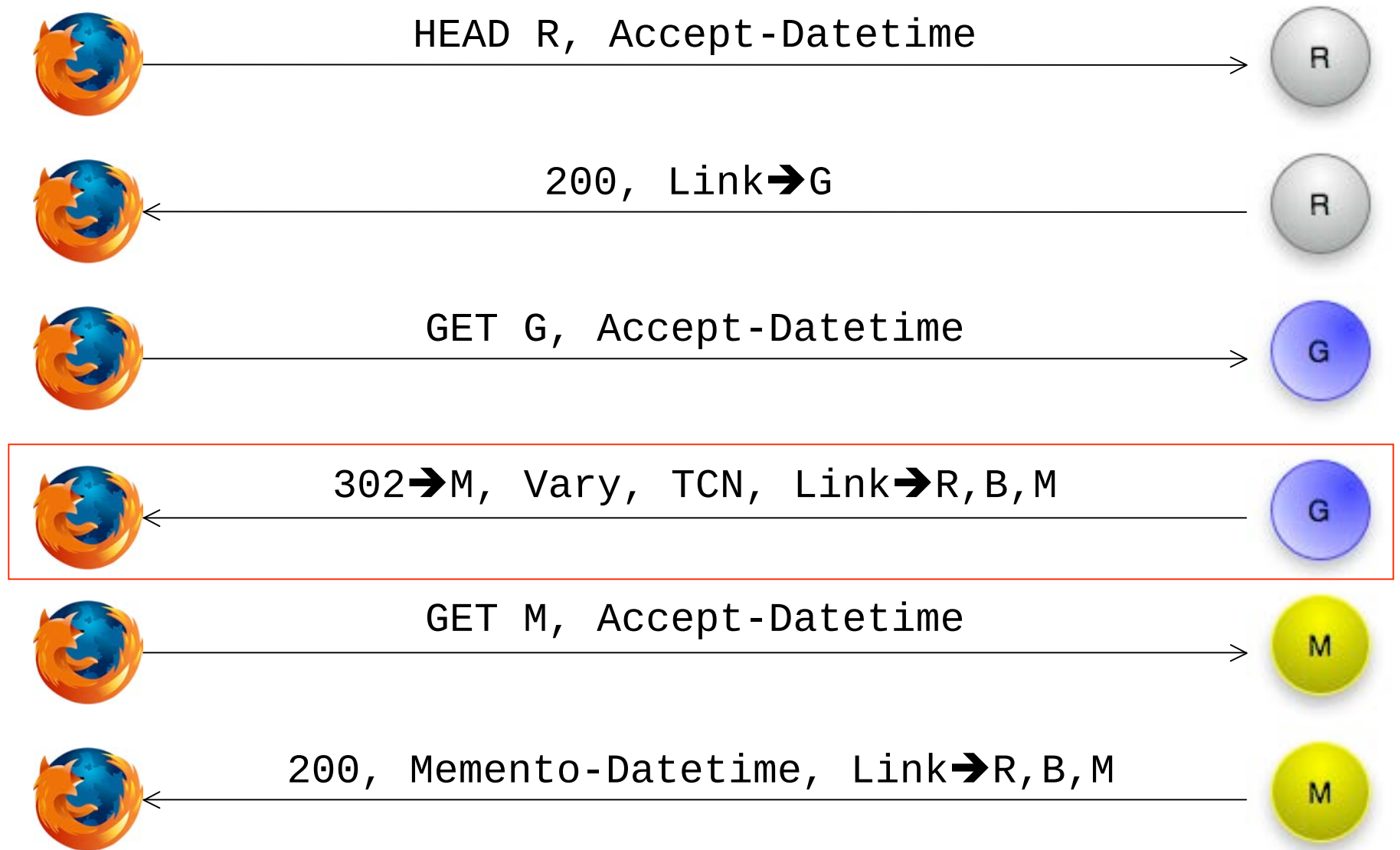
Memento HTTP Flow



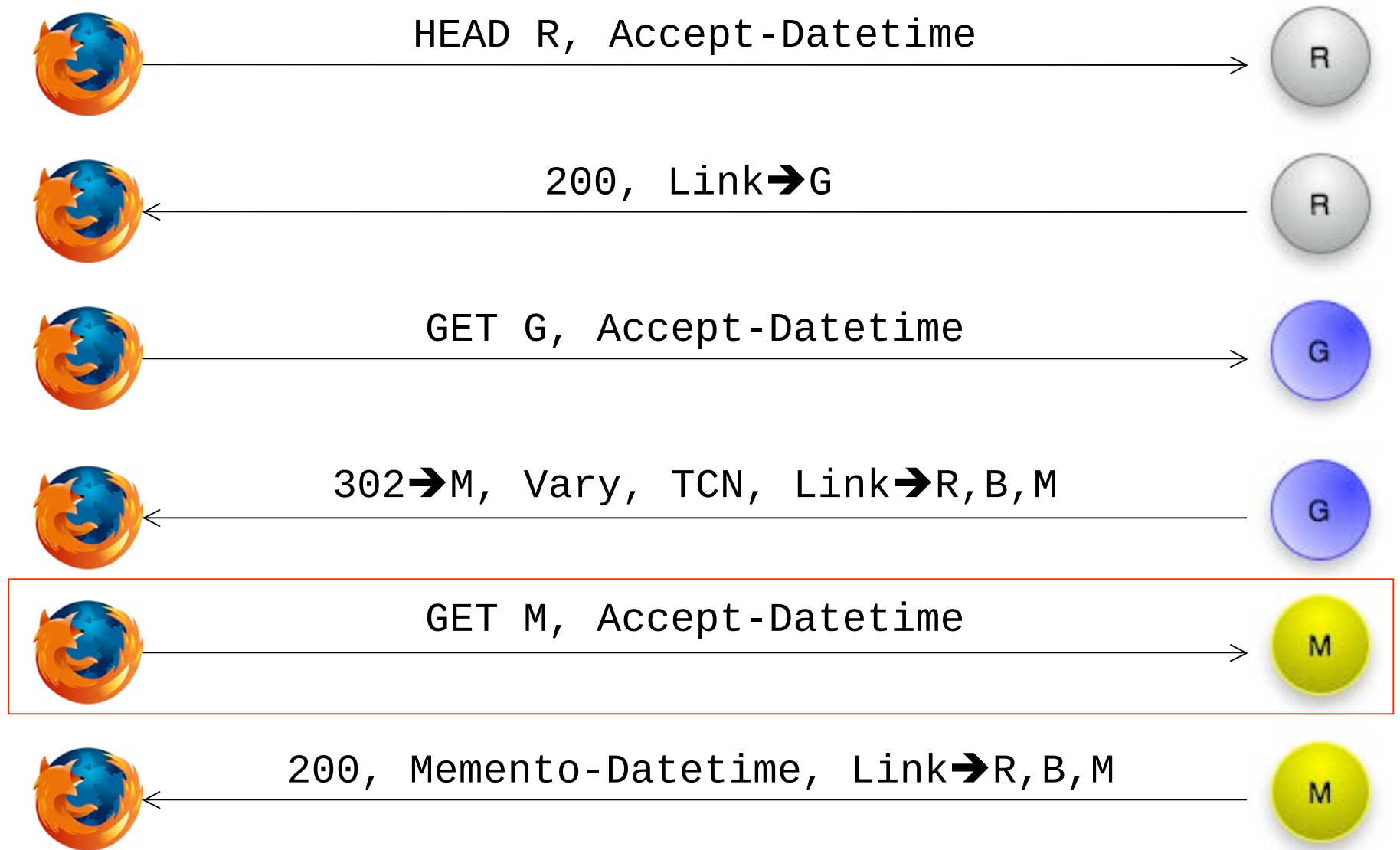
Memento HTTP Flow



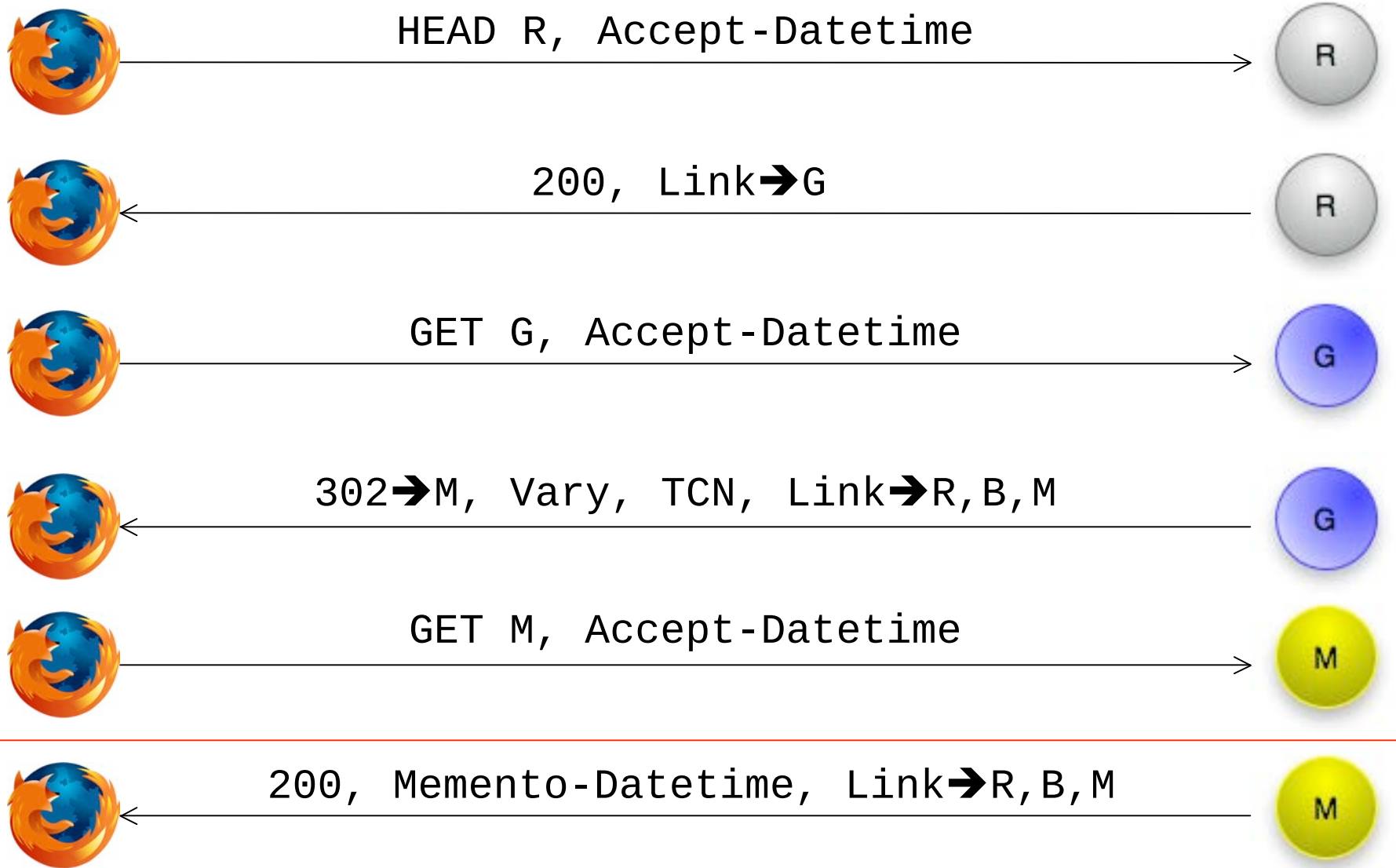
Memento HTTP Flow



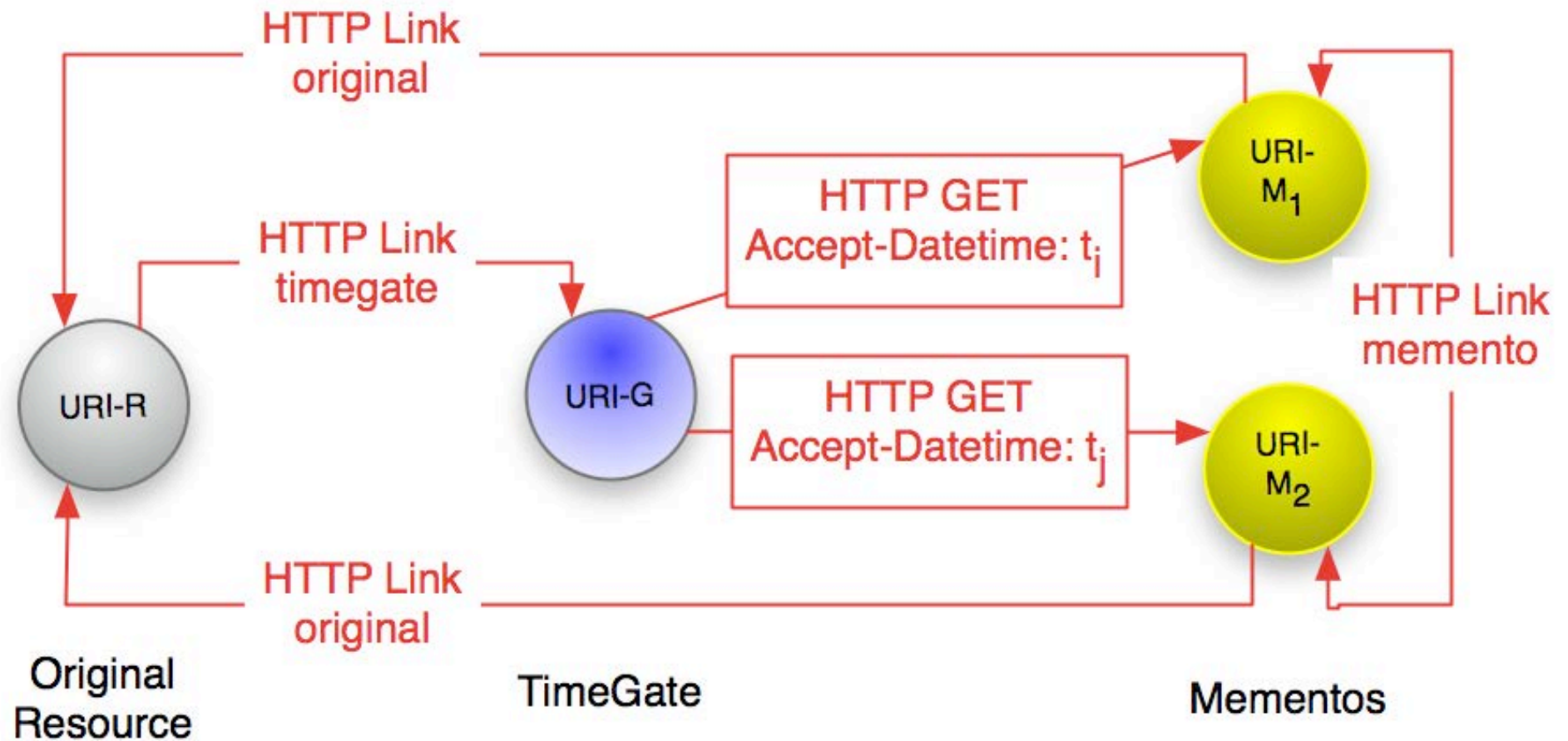
Memento HTTP Flow



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



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



Compliant Client, Non-Compliant Server



HEAD R, Accept-Datetime



200



GET G, Accept-Datetime



302→M, Vary, TCN, Link→R,B,M



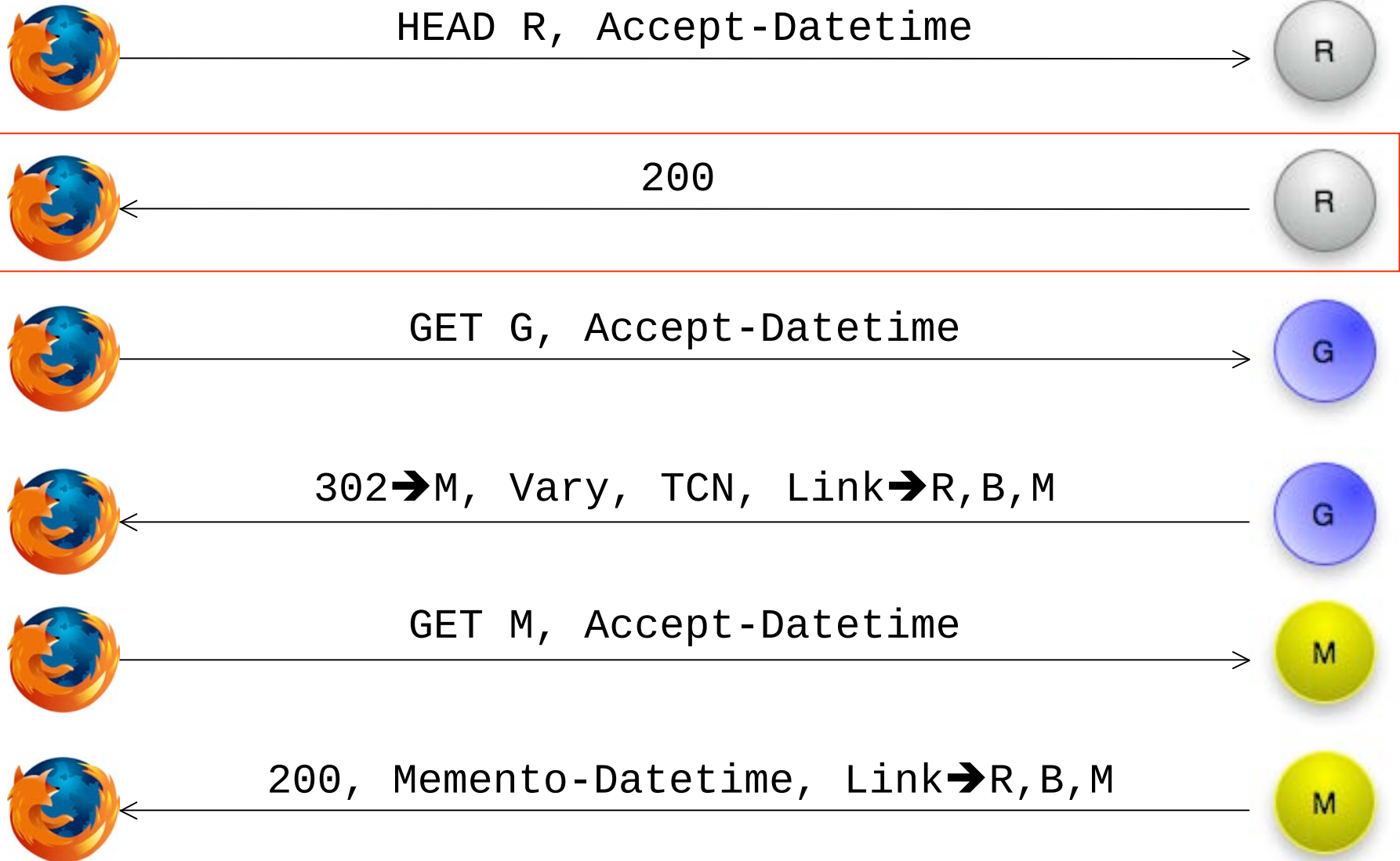
GET M, Accept-Datetime



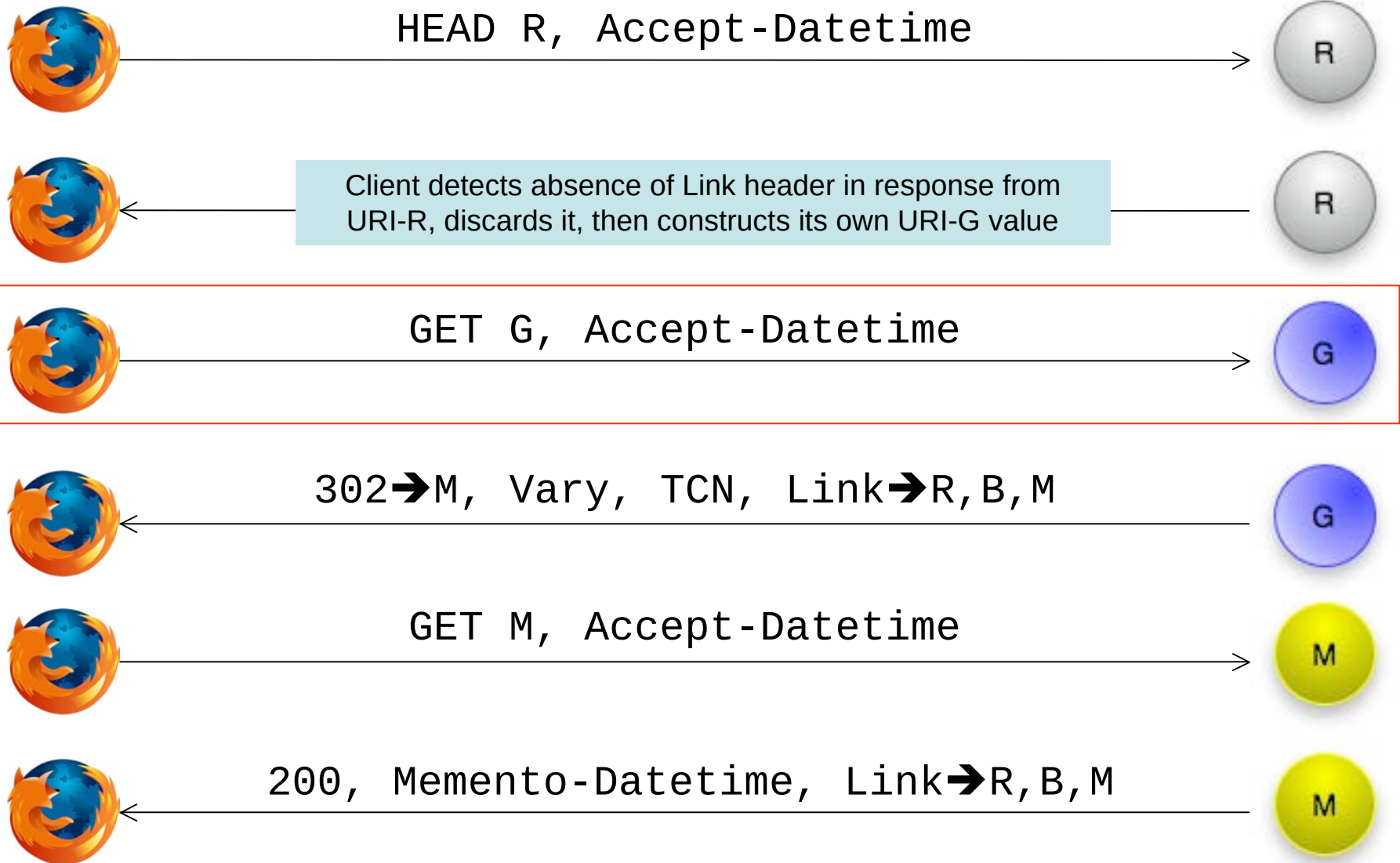
200, Memento-Datetime, Link→R,B,M



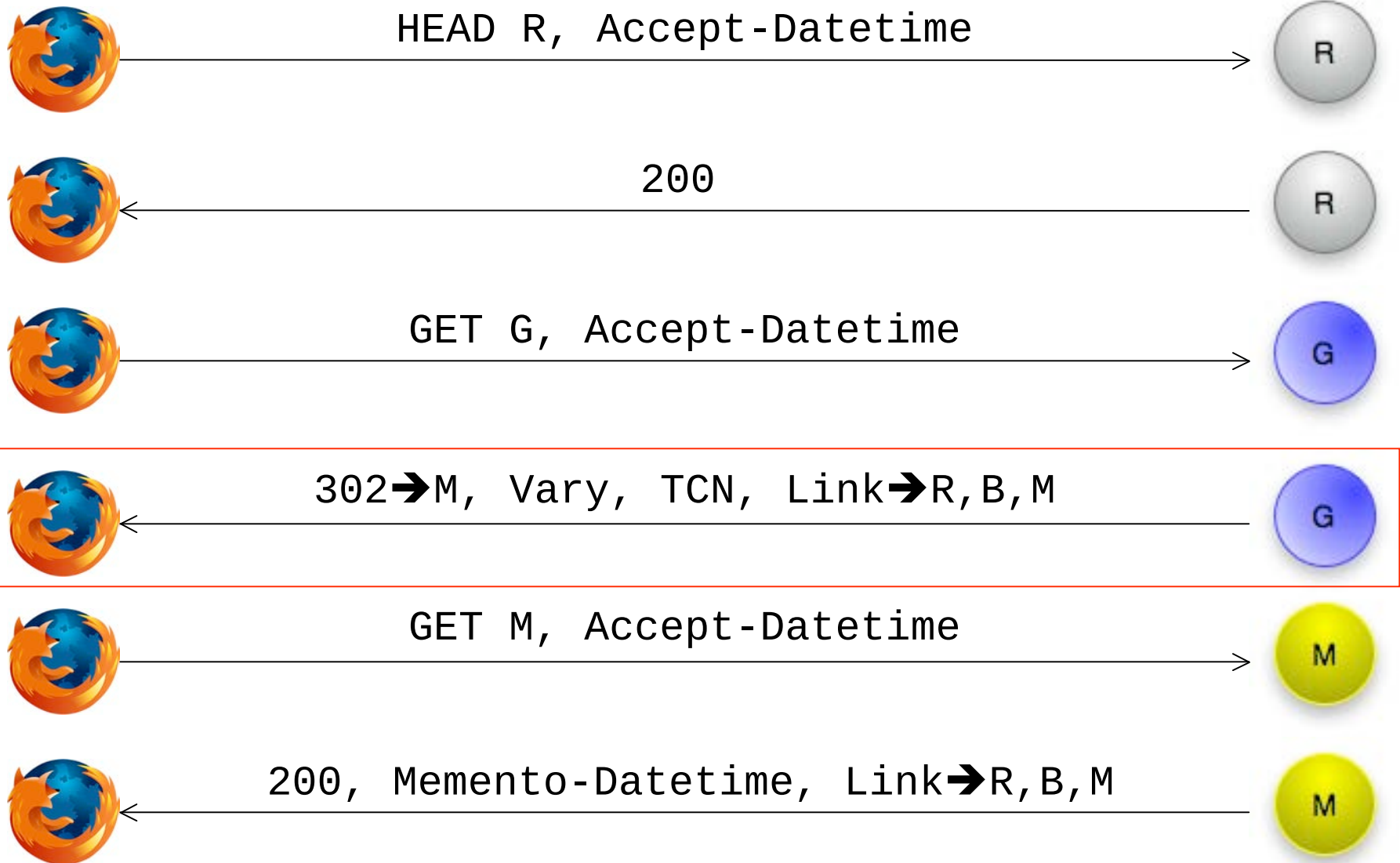
Compliant Client, Non-Compliant Server



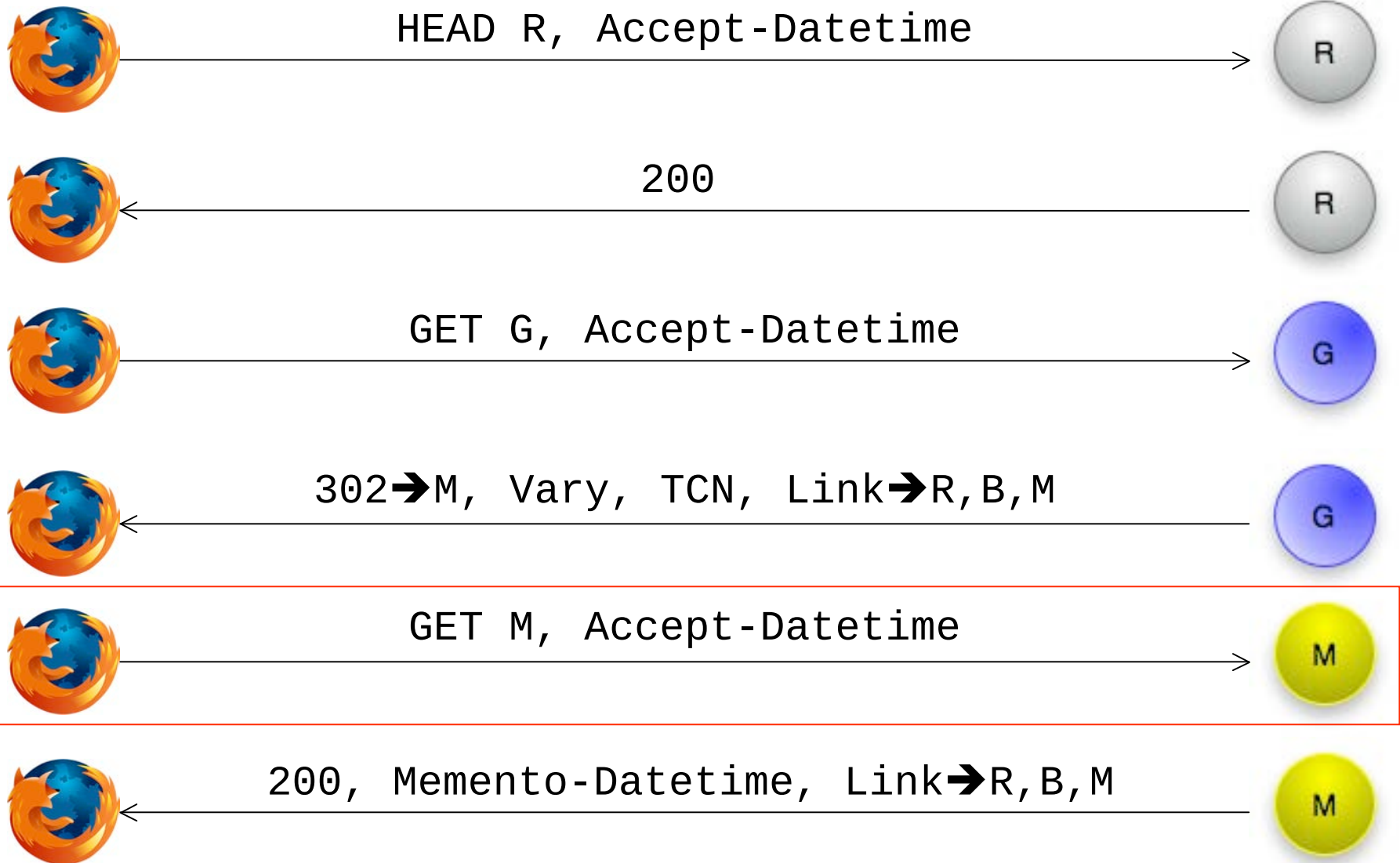
Compliant Client, Non-Compliant Server



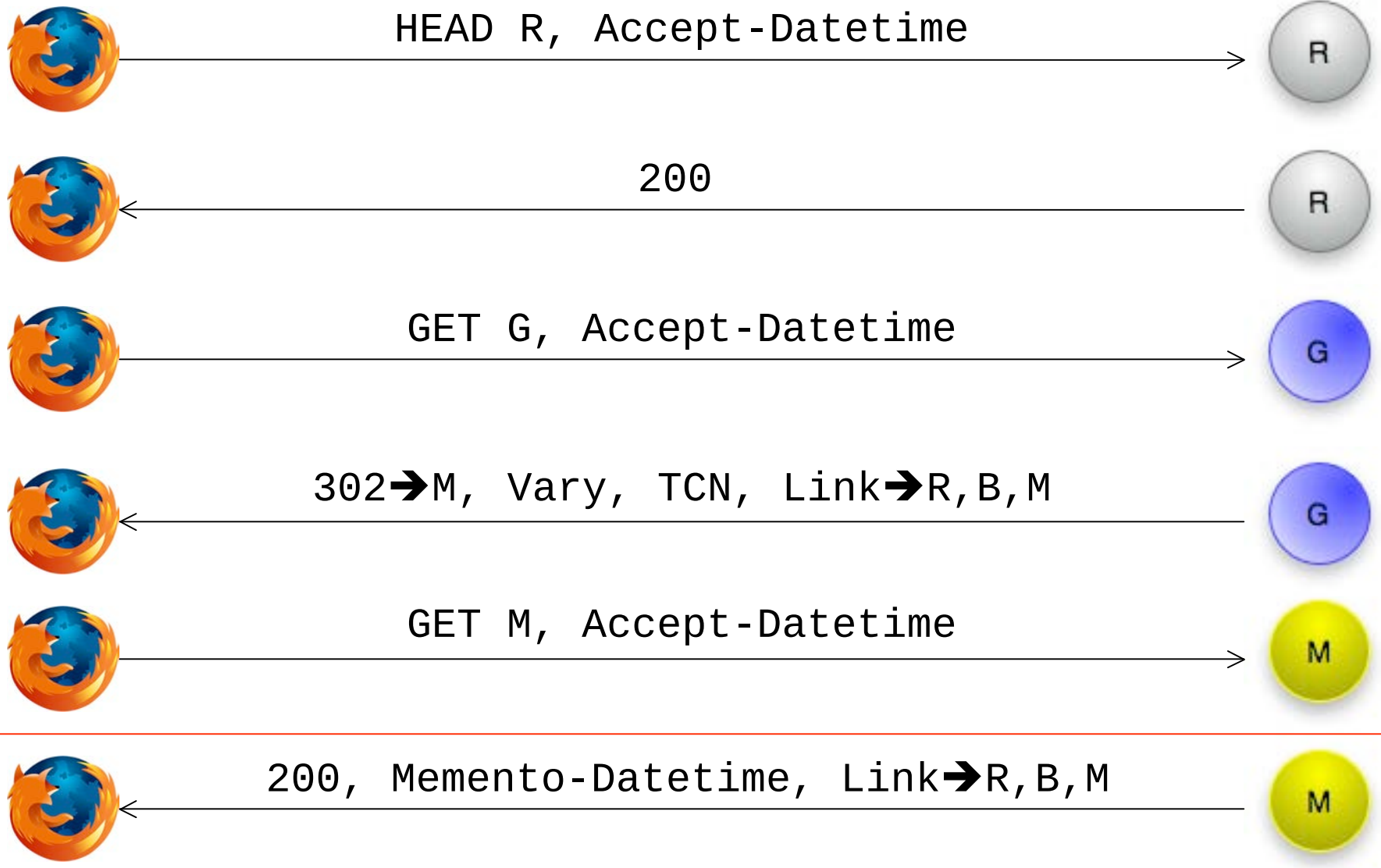
Compliant Client, Non-Compliant Server



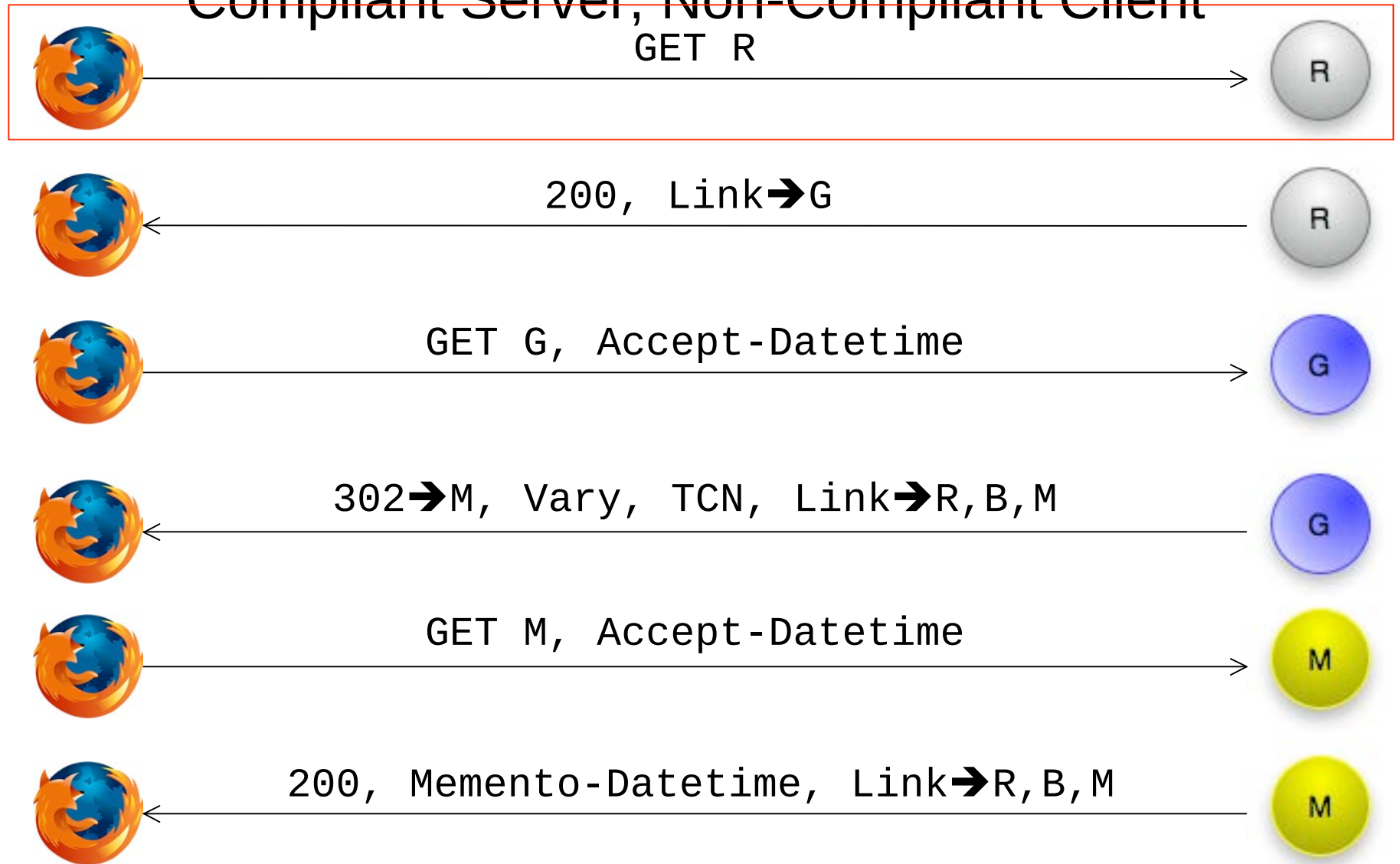
Compliant Client, Non-Compliant Server



Compliant Client, Non-Compliant Server



Compliant Server, Non-Compliant Client



Compliant Server, Non-Compliant Client



GET R



200, Link→G



GET G, Accept-Datetime



302→M, Vary, TCN, Link→R,B,M



GET M, Accept-Datetime



200, Memento-Datetime, Link→R,B,M



Compliant Server, Non-Compliant Client



GET R, Accept-Datetime



200, Link→G



GET G, Accept-Datetime



302→M, Vary, TCN, Link→R,B,M



GET M, Accept-Datetime



200, Memento-Datetime, Link→R,B,M



Some Issues

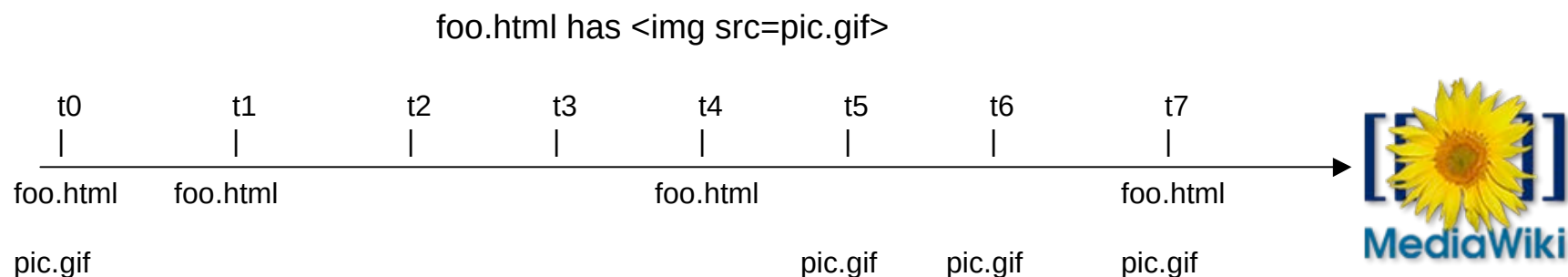
- Observational uncertainty
- Different notions of time
- When is the past?



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



No Uncertainty With Self-Archiving Systems

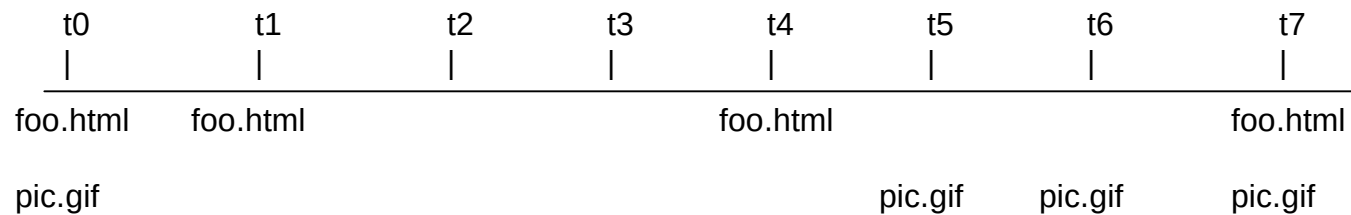


Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



foo.html @ t4

foo.html has



```
GET /foo.html
Accept-Datetime: t4

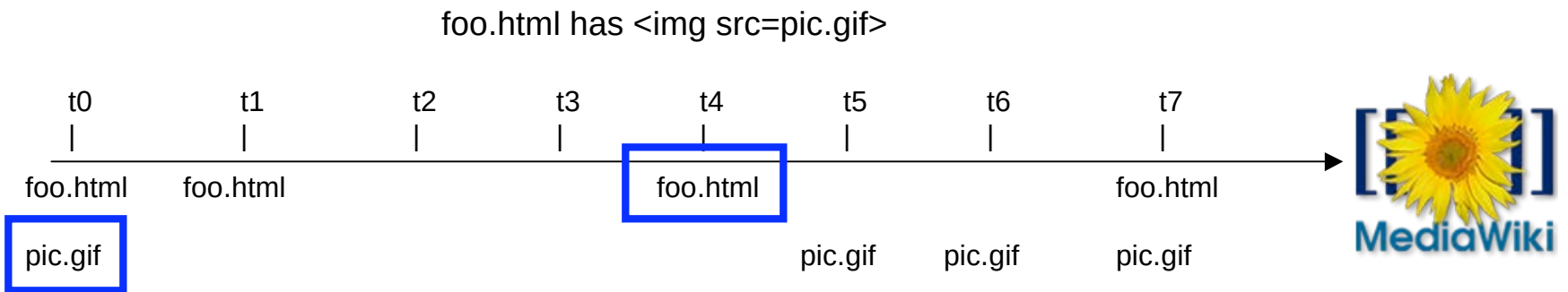
HTTP/1.1 200 OK
Memento-Datetime: t4
```

```
GET /pic.gif
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t0
```



foo.html @ t4



```
GET /foo.html
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t4
```

foo.html correct

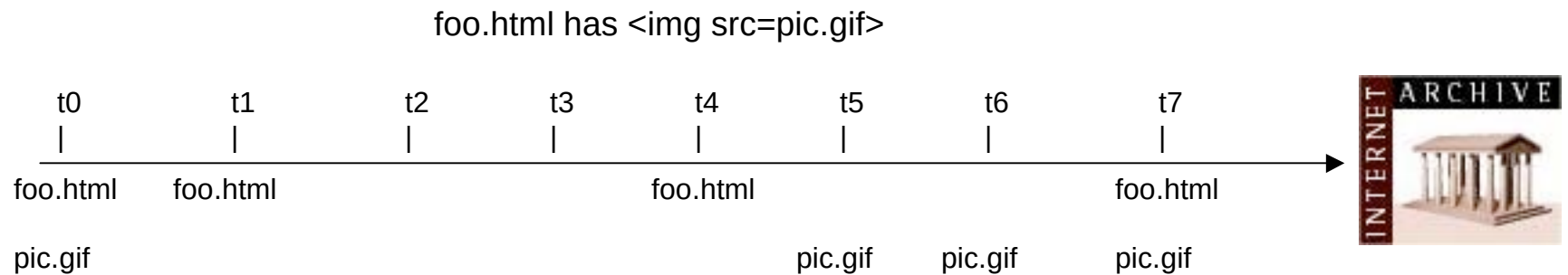
```
GET /pic.gif
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t0
```

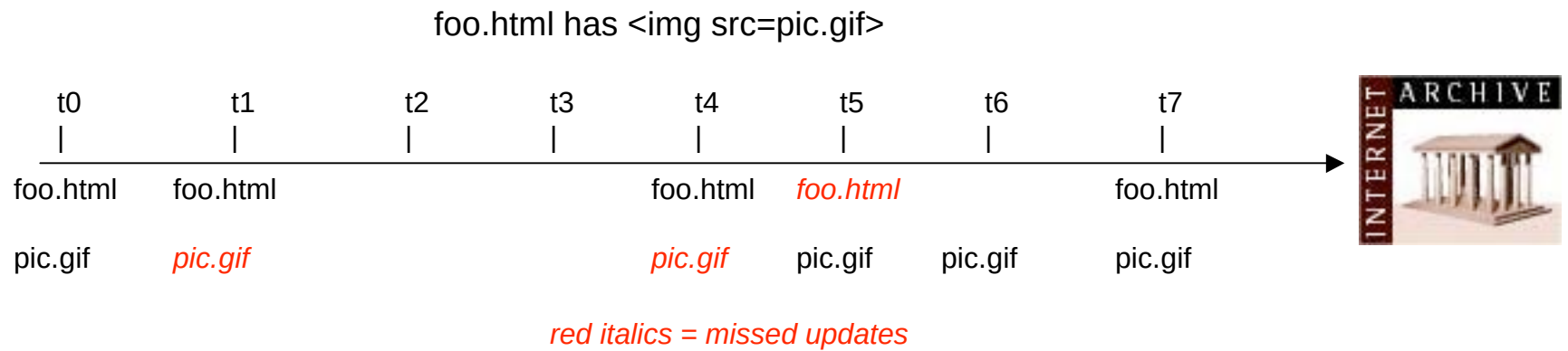
pic.gif correct



Uncertainty in Third-Party Archives

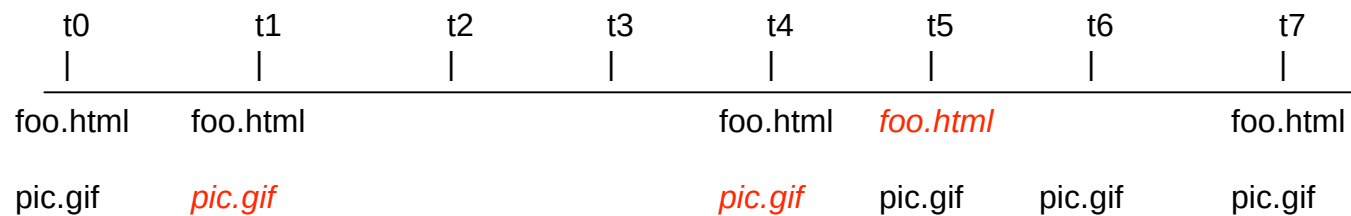


Missed Updates



foo.html @ t4

foo.html has



```
GET /foo.html
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t4
```

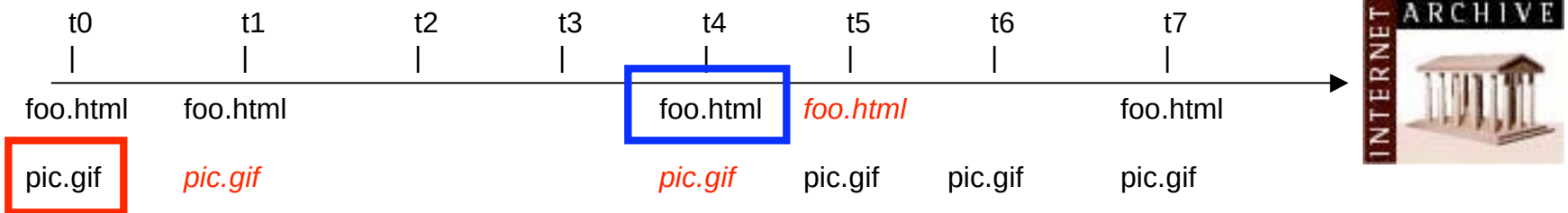
```
GET /pic.gif
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t0
```



foo.html @ t4

foo.html has



```
GET /foo.html
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t4
```

foo.html correct

```
GET /pic.gif
Accept-Datetime: t4

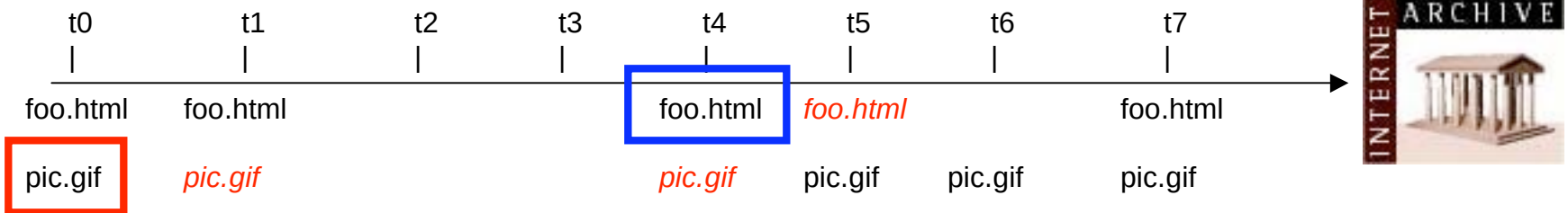
HTTP/1.1 200 OK
Memento-Datetime: t0
```

pic.gif incorrect
(should be t4)



foo.html @ t4

foo.html has



```
GET /foo.html
Accept-Datetime: t4

HTTP/1.1 200 OK
Memento-Datetime: t4
```

foo.html correct

```
GET /pic.gif
Accept-Datetime: t4

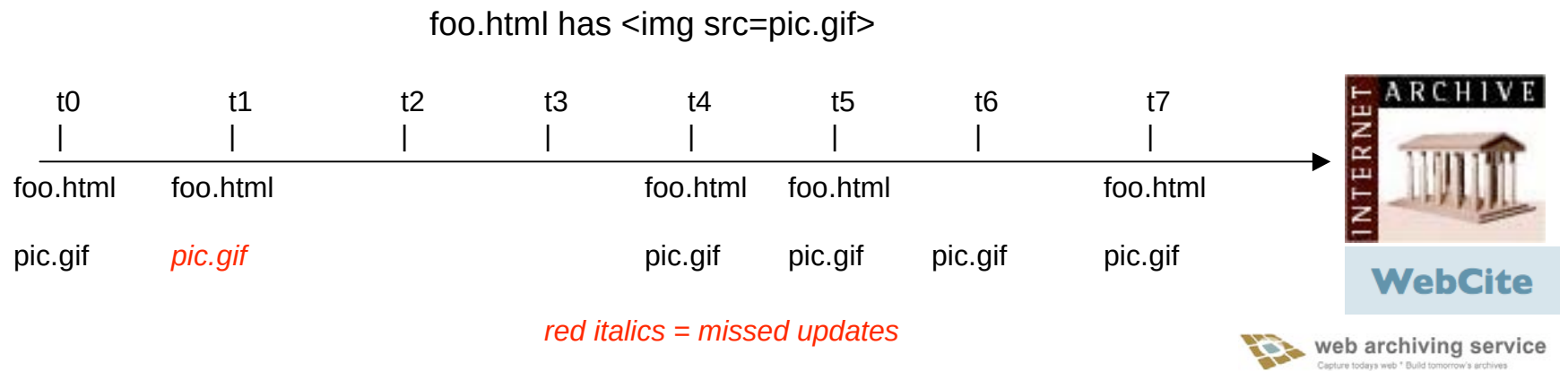
HTTP/1.1 200 OK
Memento-Datetime: t0
```

pic.gif incorrect
(should be t4)

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



Decrease Uncertainty With More Observations?

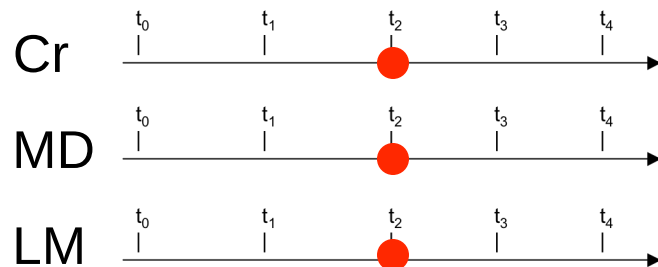


Three Notions of Time: Cr, LM, MD

- 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



Internet Archive Wayback Machine

http://wayback.archive-it.org/1029/*/http://www.nyu.edu/fas/pr

Requested Date: 1996 2010 07/19/2010

Internet Archive Wayback Machine

ARCHIVE-IT

ns11mm Web Archive (National September 11 Memorial Museum)

INTERNET ARCHIVE waybackmachine

Enter Web Address: http:// All Take Me Back Compare Archive Pages

Searched for http://www.nyu.edu/fas/projects/vcb/case_911_FLASHcontent.html 1 Result [RSS](#)

[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

* denotes when page was updated

Search Results for Jan 1, 2005 - Dec 31, 2010					
2005	2006	2007	2008	2009	2010
0 pages	0 pages	0 pages	1 page	0 pages	0 pages
			Mar 6, 2008 *		

[Home](#) | [Copyright](#) | [Privacy Policy](#)

Find: Next Previous Highlight all Match case

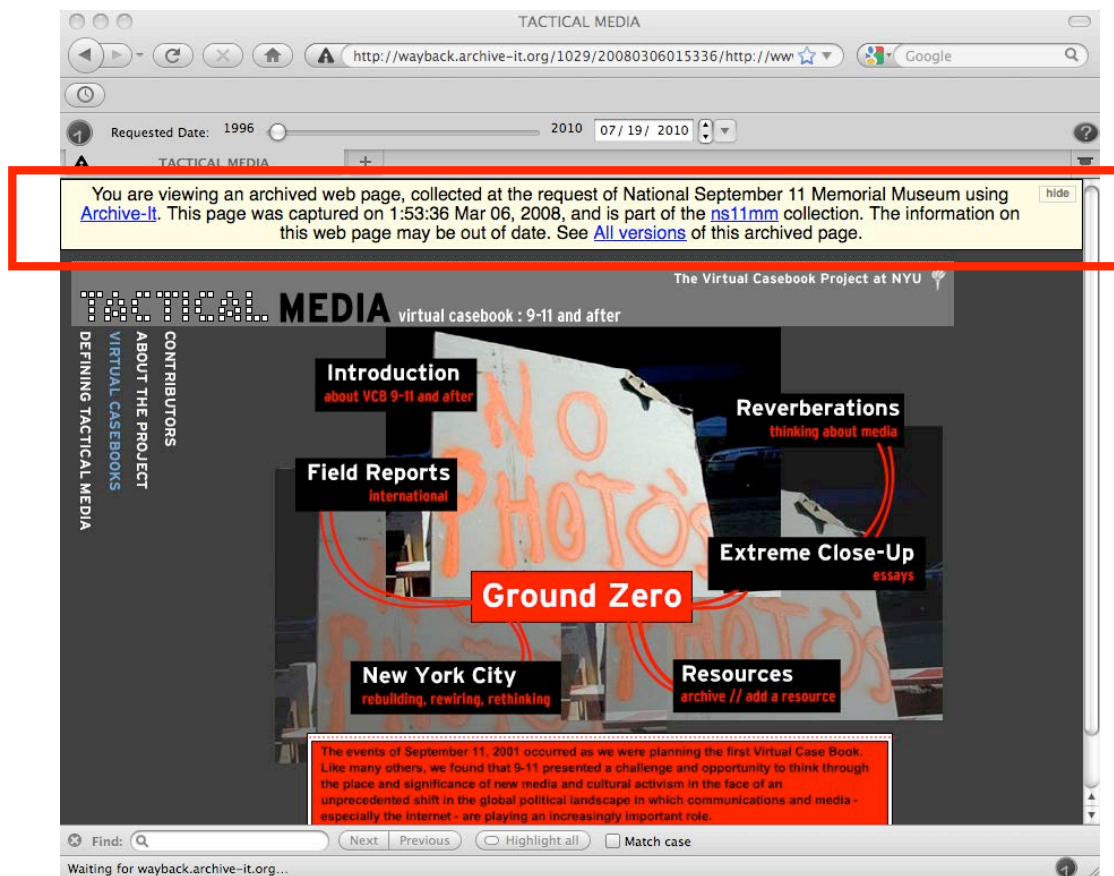
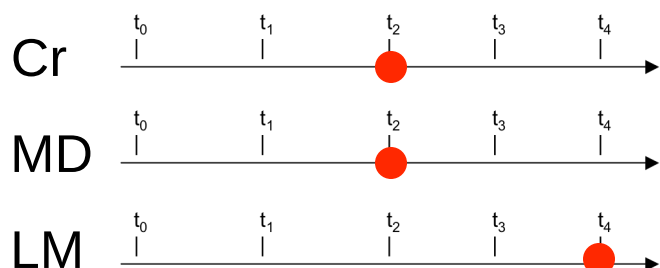
Done



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



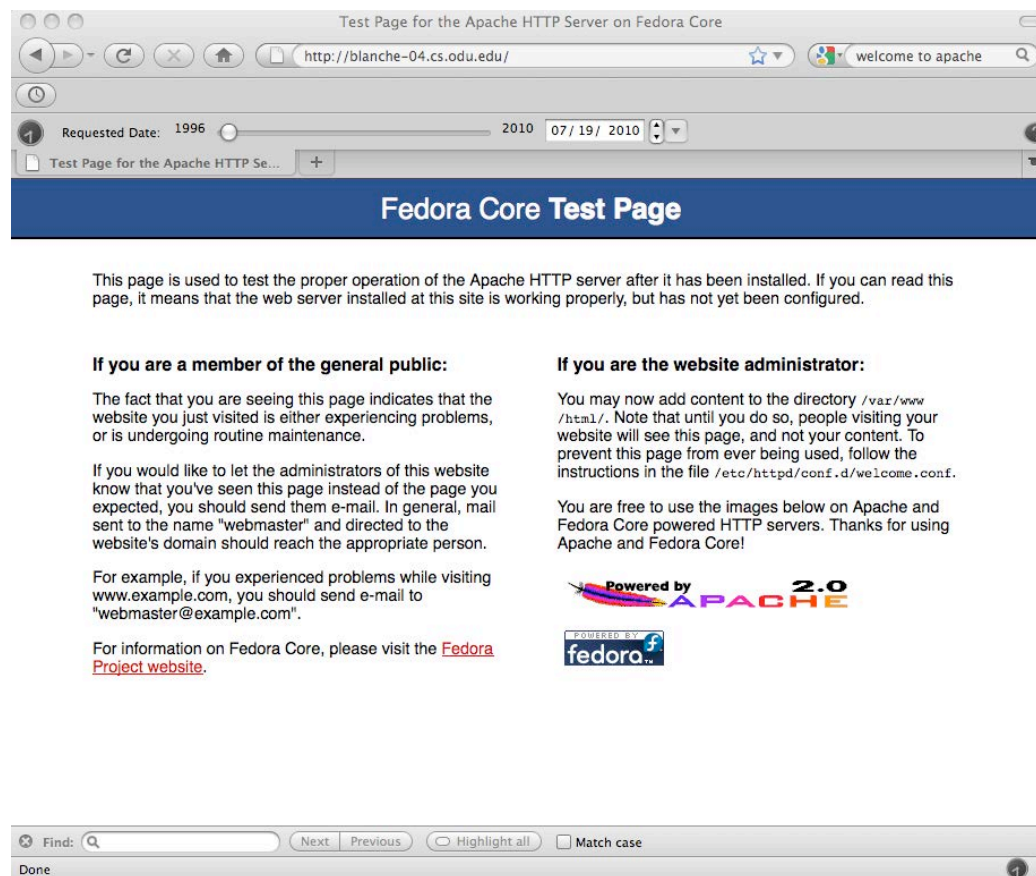
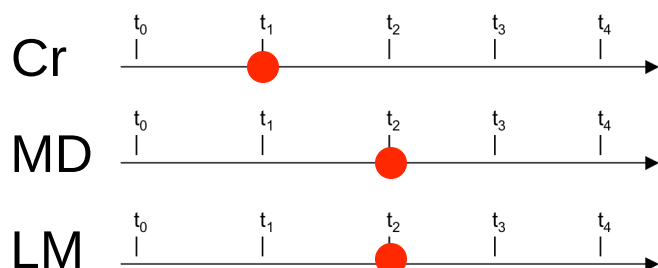
$$Cr == MD < LM$$



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



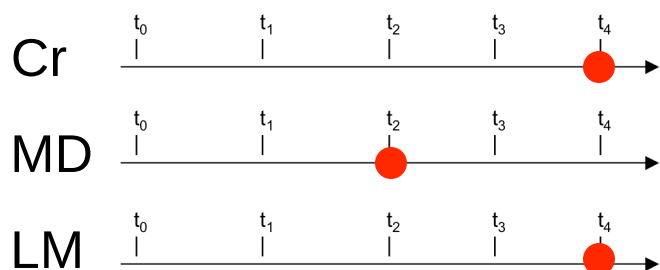
$$Cr < MD \leq LM$$



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



$$MD < Cr \leq LM$$



The screenshot shows the Internet Archive Wayback Machine interface. The browser address bar displays `http://web.archive.bibalex.org/web/*/http://www.cs.odu.edu/~mln/`. The "Requested Date" is set to 1996, and the "Take Me Back" button is visible. The search results for `http://www.cs.odu.edu/~mln/` are displayed, showing 35 results. The results are organized by year, with a table for the years 1996 through 2010. The table shows the number of pages archived for each year and the dates of the captures.

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	5 pages	6 pages	8 pages	7 pages	5 pages	1 pages	0 pages	0 pages	0 pages
						May 26, 2002 * May 28, 2002 * Jul 24, 2002 * Oct 05, 2002 * Nov 22, 2002 *	Apr 26, 2003 * Jun 25, 2003 * Aug 03, 2003 * Oct 20, 2003 * Oct 24, 2003 * Dec 06, 2003	May 13, 2004 Jun 03, 2004 Jun 09, 2004 Jul 01, 2004 Aug 26, 2004 Aug 28, 2004 Sep 30, 2004 Dec 02, 2004	Feb 05, 2005 Feb 06, 2005 Feb 13, 2005 Mar 10, 2005 Apr 03, 2005 Aug 29, 2005 * Dec 05, 2005	Feb 05, 2006 Feb 11, 2006 Mar 03, 2006 Apr 17, 2006 Apr 27, 2006	Dec 15, 2007 *			

Home | FAQs | Terms of Use | Privacy Policy



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



How does Memento do This?

There are two components to the Memento Solution:

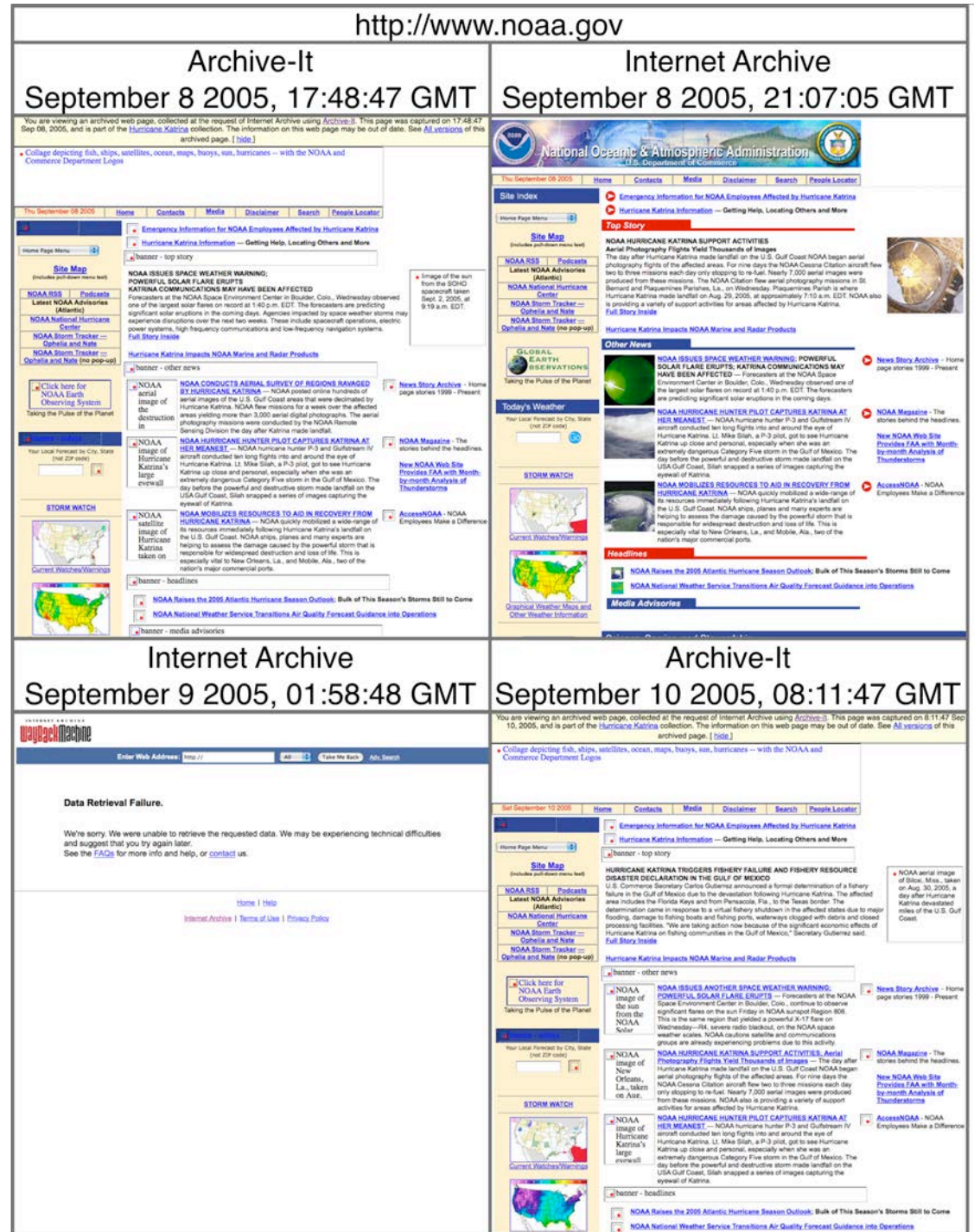
- *Component 1: Navigation towards an archived resource via its original resource, by leveraging content negotiation.*
- *Component 2: A discovery API for archives that allows requesting a list of all archived versions it holds for a resource with a given URI.*



- Mementos for any given URI-R are distributed across archives.

- In order to get a correct perspective of available Mementos, different archives need to be consulted.

- Can do so in distributed consultation mode (sloooow), or by consulting an aggregator.

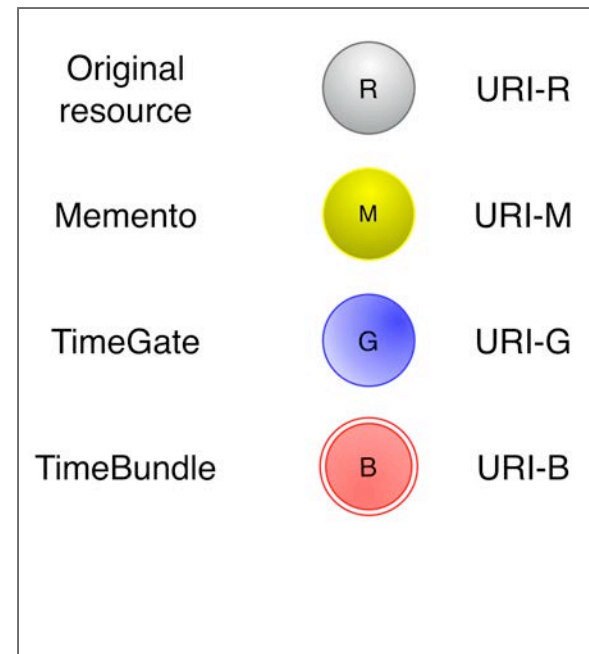


Terminology Intermission

*We introduce the term **TimeBundle** to refer to a resource via which an overview of all Mementos for an original resource **URI-R** is available.*

A TimeBundle for a resource URI-R, is a resource URI-B[URI-R] that is an aggregation of:

- (a) All Mementos URI-Mi [URI-R@t_i] available from an archive,
- (b) The archive's TimeGate URI-G for URI-R,
- (c) The original resource URI-R itself.



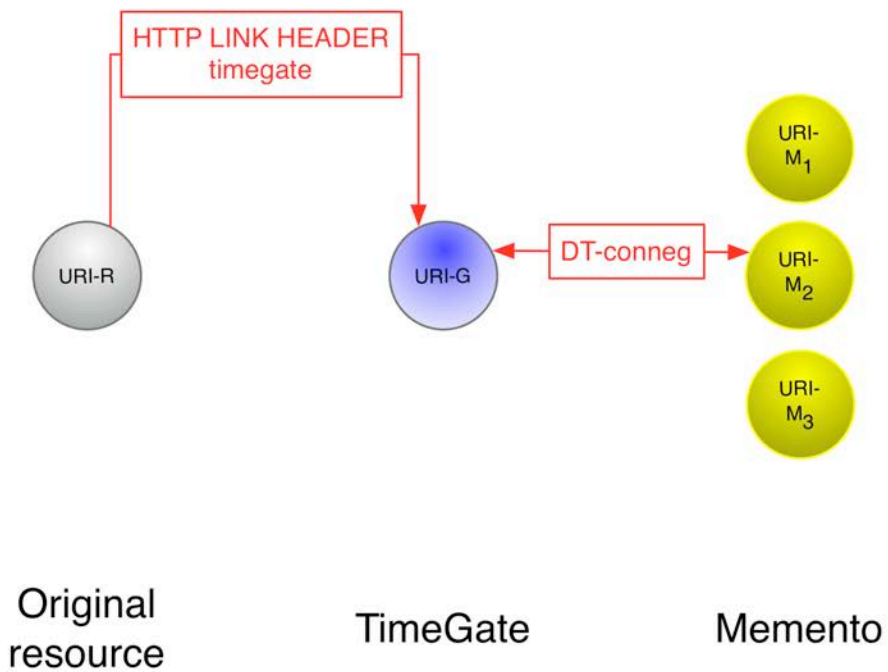


Original
resource



Memento

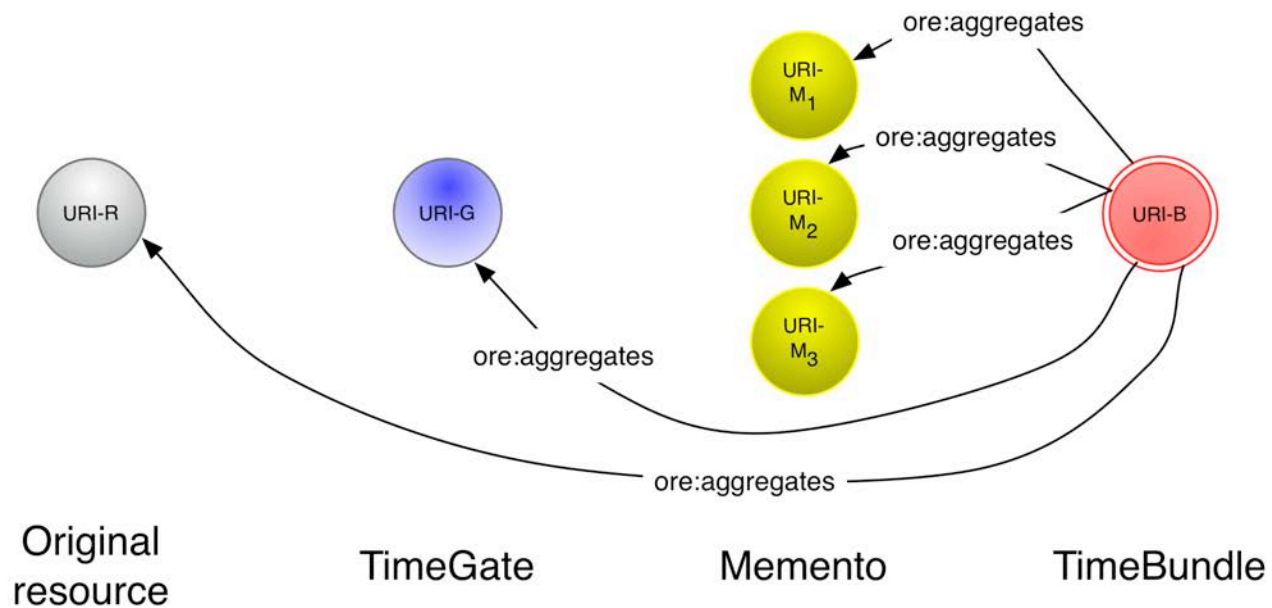




← Memento DT-conneg component →

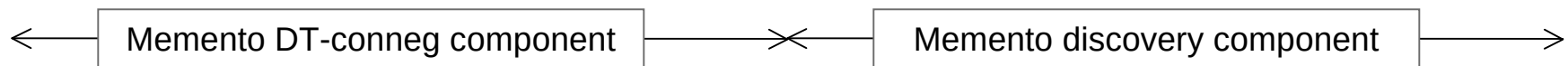
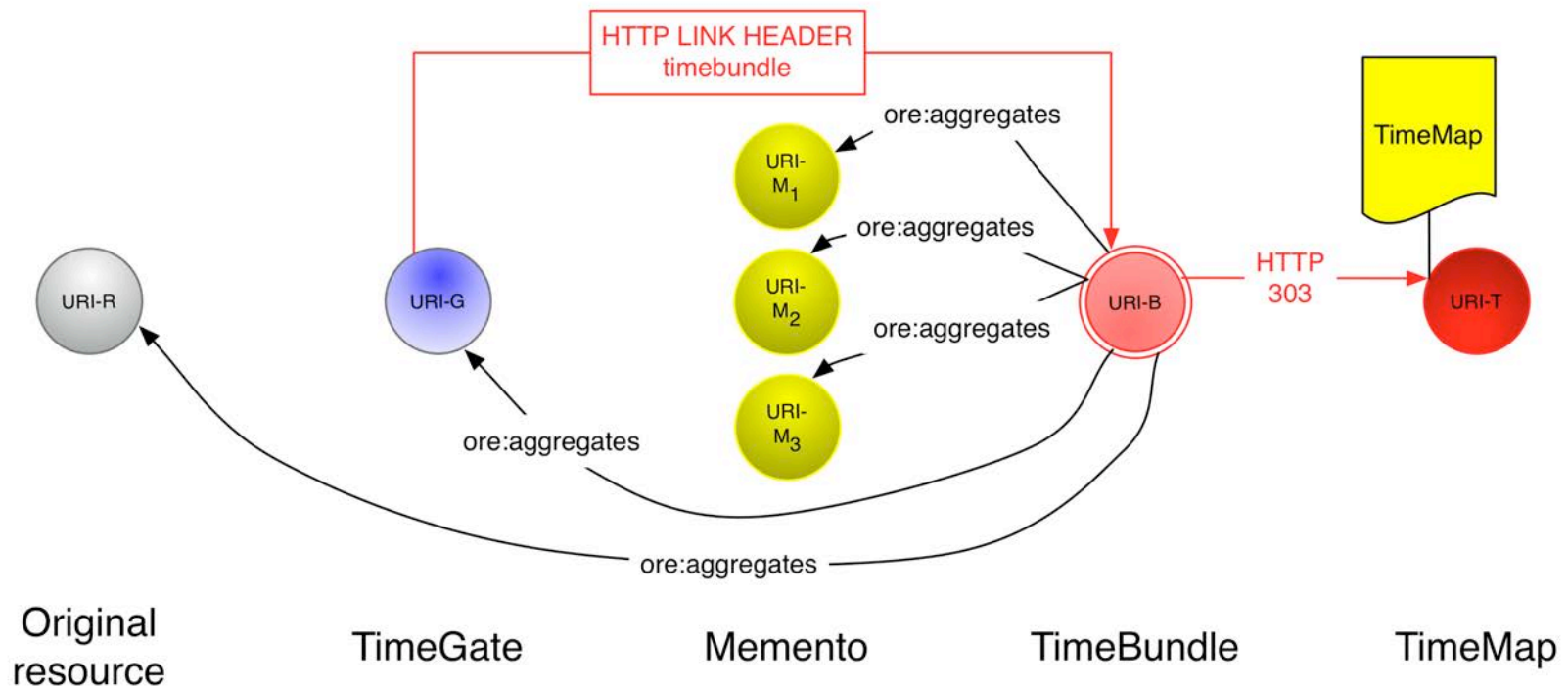


See OAI-ORE: <http://www.openarchives.org/ore/1.0/toc/>



← Memento DT-conneg component →





Examining the URI-G Response...



302→M, Vary, Link→R,B,M



```
HTTP/1.1 302 Found
Date: Thu, 21 Jan 2010 00:06:50 GMT
Server: Apache
TCN: choice
Vary: negotiate, accept-datetime
Location: http://wayback.archive-it.org/1610/20090928171405/http://
  www.digitalpreservation.gov/
Link: <http://www.digitalpreservation.gov/>; rel="original",
  <http://mementoproxy.lanl.gov/aggr/timebundle/http://www.digitalpreservation.gov/>;
    rel="timebundle",
  <http://wayback.archive-it.org/256/20051108162921/http://www.digitalpreservation.gov/>;
    rel="first-memento"; datetime="Tue, 08 Nov 2005 00:00:00 GMT",
  <http://webcitation.org/query?id=1257028234035091>;
    rel="next-memento"; datetime="Sat, 31 Oct 2009 18:30:35 GMT",
  <http://webcitation.org/query?id=1213058061345794>;
    rel="prev-memento"; datetime="Mon, 09 Jun 2008 20:34:23 GMT",
  <http://wayback.archive-it.org/256/20100120102000/http://www.digitalpreservation.gov/>;
    rel="last-memento"; datetime="Wed, 20 Jan 2010 10:20:00 GMT"
Content-Length: 0
Connection: close
```

Dereferencing URI-B

```
% telnet mementoproxy.lanl.gov 80
Trying 204.121.6.37...
Connected to ttt.lanl.gov.
Escape character is '^]'.
HEAD /aggr/timebundle/http://www.digitalpreservation.gov/ HTTP/1.1
Host: mementoproxy.lanl.gov
Connection: close
```

```
HTTP/1.1 303 See Other
Date: Wed, 21 Jul 2010 03:09:46 GMT
Server: Apache
Location:
  http://mementoproxy.lanl.gov/aggr/timemap/rdf/http://www.digitalpreservation.gov/
Vary: Accept
Connection: close
Content-Type: text/plain; charset=UTF-8
```

Connection closed by foreign host.



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



RDF?! Yuck!

```
% telnet mementoproxy.lanl.gov 80
Trying 204.121.6.37...
Connected to ttt.lanl.gov.
Escape character is '^]'.
HEAD /aggr/timebundle/http://www.digitalpreservation.gov/ HTTP/1.1
Accept: application/rdf+xml; q=0.0
Host: mementoproxy.lanl.gov
Connection: close

HTTP/1.1 303 See Other
Date: Wed, 21 Jul 2010 03:12:42 GMT
Server: Apache
Location:
  http://mementoproxy.lanl.gov/aggr/timemap/link/http://www.digitalpreservation.gov/
Vary: Accept
Connection: close
Content-Type: text/plain; charset=UTF-8

Connection closed by foreign host.
```



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



TimeMap: URI-T

<http://mementoproxy.lanl.gov/aggr/timemap/rdf/http://www.digitalpreservation.gov/>

<http://mementoproxy.lanl.gov/aggr/timemap/link/http://www.digitalpreservation.gov/>

```
<http://mementoproxy.lanl.gov/aggr/timebundle/http://www.digitalpreservation.gov/>;rel="timebundle",  
<http://www.digitalpreservation.gov/>;rel="original",  
<http://web.archive.org/web/20020802022406/www.digitalpreservation.gov/>;rel="first-memento";datetime="Fri, 02 Aug 2002 02:24:06 GMT",  
<http://web.archive.org/web/20020921111830/www.digitalpreservation.gov/>;rel="memento";datetime="Sat, 21 Sep 2002 11:18:30 GMT",  
<http://web.archive.org/web/20020924113650/www.digitalpreservation.gov/>;rel="memento";datetime="Tue, 24 Sep 2002 11:36:50 GMT",  
<http://web.archive.org/web/20020927005417/www.digitalpreservation.gov/>;rel="memento";datetime="Fri, 27 Sep 2002 00:54:17 GMT",  
...[deletia]...  
<http://web.archive.nationalarchives.gov.uk/20080911010610/http://www.digitalpreservation.gov/>;rel="memento";datetime="Thu, 11 Sep 2008 00:00:00 GMT",  
<http://web.archive.org/web/20090516160321/www.digitalpreservation.gov/>;rel="memento";datetime="Sat, 16 May 2009 16:03:21 GMT",  
<http://web.archive.org/web/20090616162603/www.digitalpreservation.gov/>;rel="memento";datetime="Tue, 16 Jun 2009 16:26:03 GMT",  
<http://web.archive.org/web/20090716162514/www.digitalpreservation.gov/>;rel="memento";datetime="Thu, 16 Jul 2009 16:25:14 GMT",  
<http://web.archive.org/web/20090816181051/www.digitalpreservation.gov/>;rel="memento";datetime="Sun, 16 Aug 2009 18:10:51 GMT",  
<http://web.archive.org/web/20090916193533/www.digitalpreservation.gov/>;rel="memento";datetime="Wed, 16 Sep 2009 19:35:33 GMT",  
<http://wayback.archive-it.org/1610/20090928171405/http://www.digitalpreservation.gov/>;rel="memento";datetime="Mon, 28 Sep 2009 00:00:00 GMT",  
<http://web.archive.org/web/20091016235112/www.digitalpreservation.gov/>;rel="memento";datetime="Fri, 16 Oct 2009 23:51:12 GMT",  
<http://webcitation.org/query?id=1257028234035091>;rel="memento";datetime="Sat, 31 Oct 2009 18:30:35 GMT",  
<http://web.archive.org/web/20091116214743/www.digitalpreservation.gov/>;rel="memento";datetime="Mon, 16 Nov 2009 21:47:43 GMT",  
<http://web.archive.org/web/20091216192113/www.digitalpreservation.gov/>;rel="memento";datetime="Wed, 16 Dec 2009 19:21:13 GMT",  
<http://web.archive.org/web/20100116192640/www.digitalpreservation.gov/>;rel="memento";datetime="Sat, 16 Jan 2010 19:26:40 GMT",  
<http://web.archive.org/web/20100216193825/www.digitalpreservation.gov/>;rel="memento";datetime="Tue, 16 Feb 2010 19:38:25 GMT",  
<http://web.archive.org/web/20100316200421/www.digitalpreservation.gov/>;rel="memento";datetime="Tue, 16 Mar 2010 20:04:21 GMT",  
<http://web.archive.org/web/20100416195253/www.digitalpreservation.gov/>;rel="memento";datetime="Fri, 16 Apr 2010 19:52:53 GMT",  
<http://web.archive.org/web/20100516200754/www.digitalpreservation.gov/>;rel="last-memento";datetime="Sun, 16 May 2010 20:07:54 GMT"
```



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



Aggregating TimeMaps

<http://mementoproxy.cs.odu.edu/aggr/timemap/link/http://digitalpreservation.gov/>

is the union of:



<http://mementoproxy.cs.odu.edu/ia/timemap/link/http://www.digitalpreservation.gov/>



<http://mementoproxy.cs.odu.edu/ait/timemap/link/http://www.digitalpreservation.gov/>

WebCite

<http://mementoproxy.cs.odu.edu/web/timemap/link/http://www.digitalpreservation.gov/>

etc.

(proxied at lanl.gov & cs.odu.edu while waiting for native support)



Review of Web Archiving: Michael L. Nelson
Web Archiving Cooperative, Stanford, Sep 09 2010



TimeBundle API: For Discovery, Cross-Archive Services

- Archive uses common approaches to make TimeBundles/TimeMaps discoverable:
 - SiteMaps,
 - Atom Feeds,
 - OAI-PMH.
- Aggregator harvests and merges TimeMaps. Based on this information, the Aggregator exposes its own TimeGates.
 - Cross-archive
 - Finer datetime granularity
 - Better chances of matching a client's datetime preference.
 - Can become a shared target for redirection for many web servers.

