

# An HTTP-Based Versioning Mechanism for Linked Data



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

Memento is partially funded by the  
Library of Congress

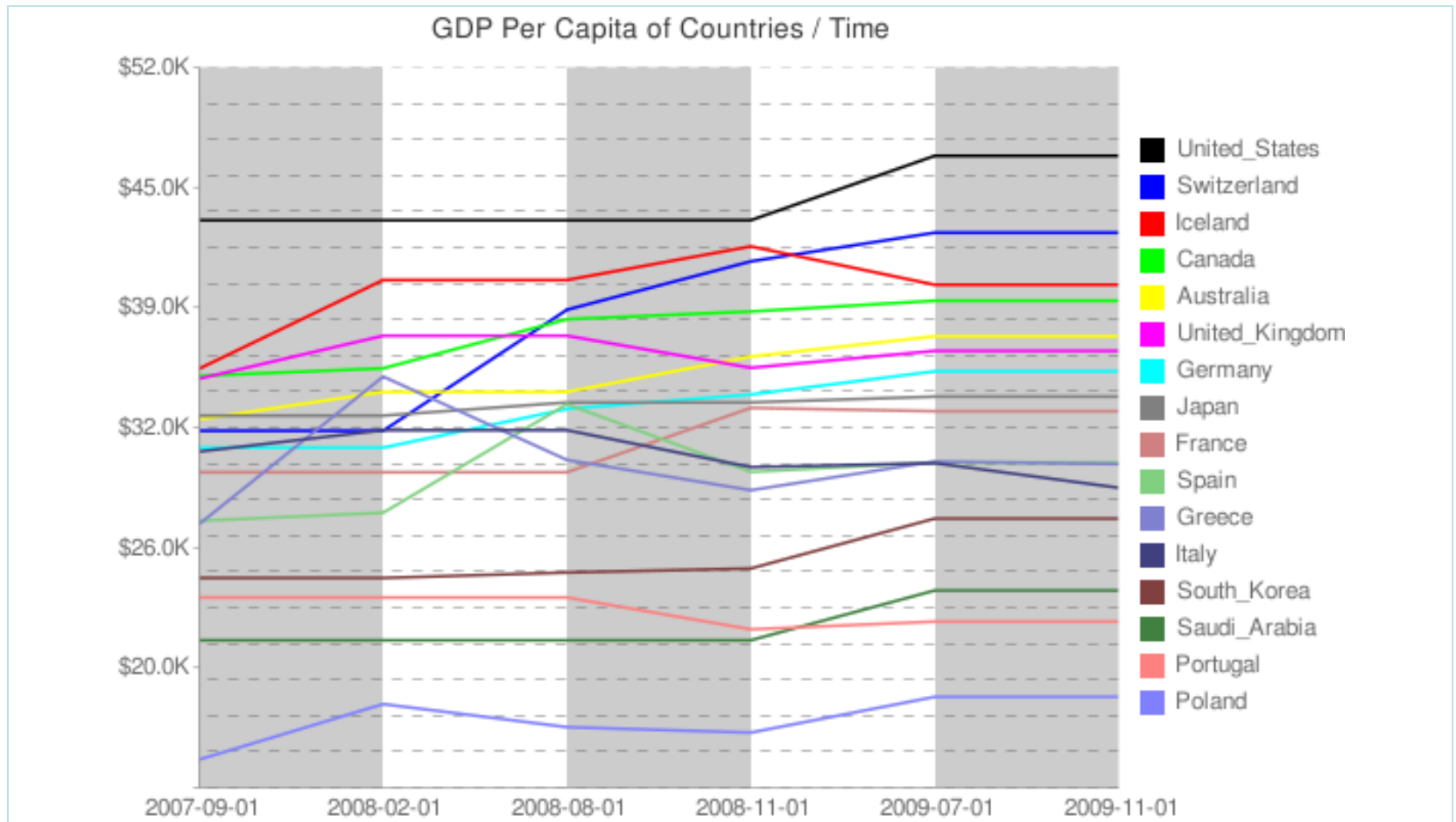
Presentation at <http://bit.ly/ac9GhH>



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Time-Series Analysis across DBpedia Versions



Data collected through HTTP Navigation



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Outline

- Memento - Time Travel for the Web
- Resource Versioning suggested by Memento
- Resource Versioning for Linked Data
- DBpedia Demonstrator



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Outline

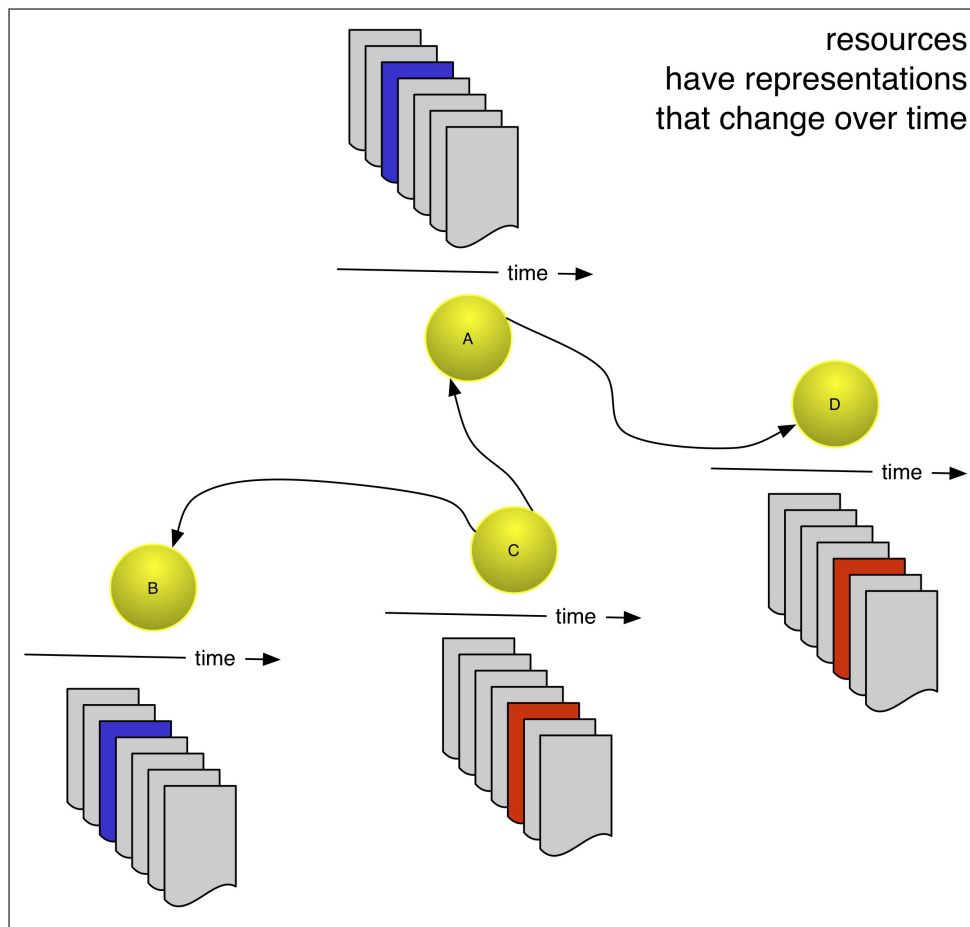
- Memento - Time Travel for the Web
- Resource Versioning suggested by Memento
- Resource Versioning for Linked Data
- DBpedia Demonstrator



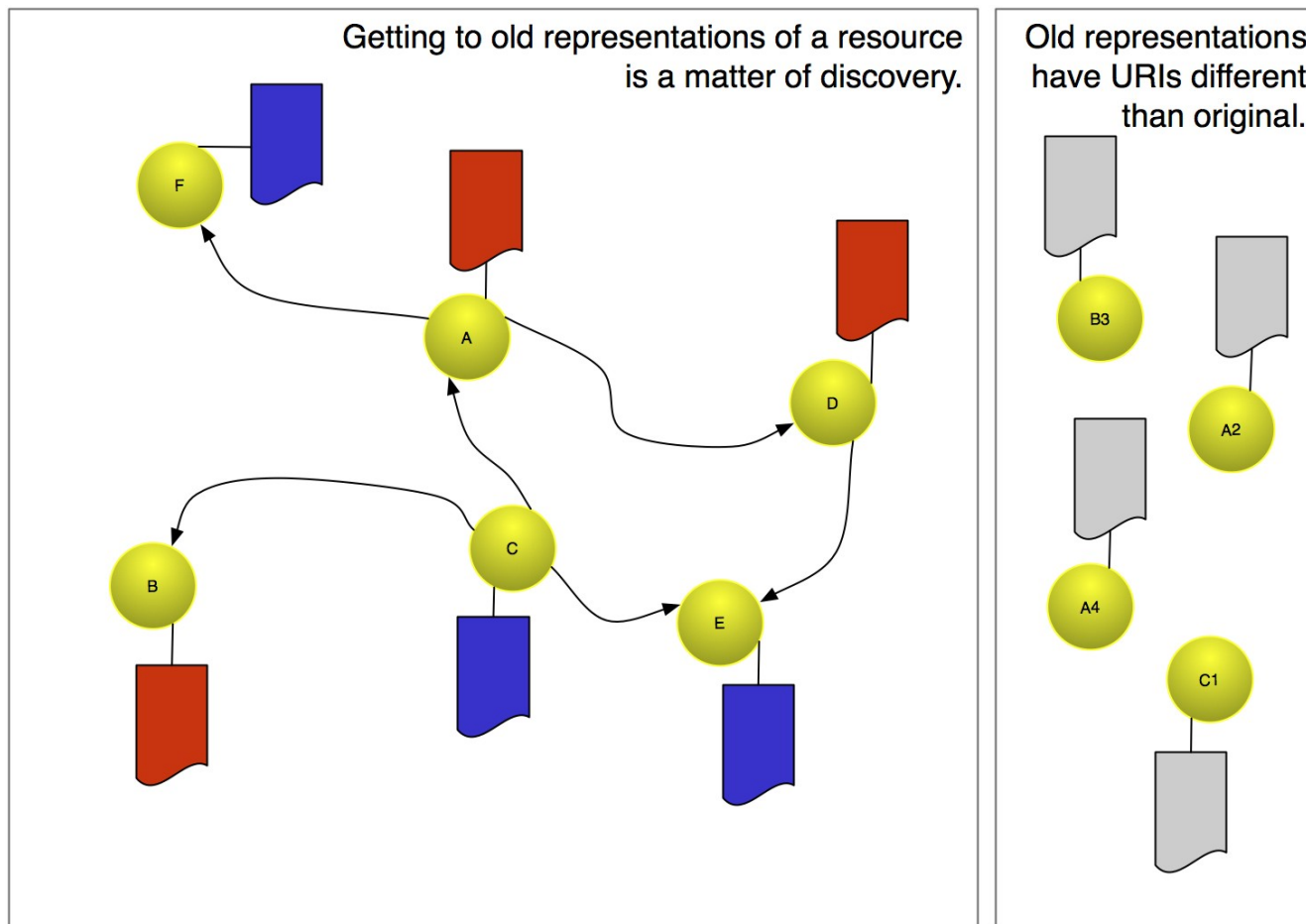
An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Resources have Representations that Change over Time



# Archived Resources serve prior Representations



Sep 11 2001, 20:36:10 UTC

Dec 20 2001, 4:51:00 UTC

# Archived Resources

The screenshot shows the CNN.com website with a 'SPECIAL REPORT' banner. The main headline is 'AMERICA UNDER ATTACK' with a sub-headline 'CNN EXCLUSIVE'. A large image shows the World Trade Center towers on fire. To the right, under 'COMPLETE COVERAGE', is a list of links: 'World Trade Center towers collapse after hit', 'Crash destroys part of Pentagon', 'White House, Capitol evacuated', 'American, United confirm losing planes', 'Bush: 'We'll hunt them down' | Statement', 'U.S. officials: More attacks can't be ruled out', and 'Attacks strike financial markets'. Below this is the 'U.S. SCENE' section with links like 'FAA grounds all U.S. flights until noon Wednesday' and 'U.S. military on 'high alert''. At the bottom, there are sections for 'VIDEO', 'PHOTO GALLERY', 'CHRONOLOGY', and 'EXTRA INFO'.

The screenshot shows the Wikipedia article 'September 11 attacks'. At the top, it says 'From Wikipedia, the free encyclopedia'. A red box contains a warning: 'This is an old revision of this page, as edited by The Cunctator (talk | contribs) at 04:51, 20 December 2001. It may differ significantly from the current revision.' Below this is a 'In Memoriam, September 11, 2001' section. The main text begins: 'On the morning of September 11, 2001, what might well be the most devastating terrorist attack in the history 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 ranged as high as 6500 people), including hundreds of firefighters who had rushed in, at the World Trade Center; and 125 at the Pentagon.' The article continues with details about the attacks and the aftermath.

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

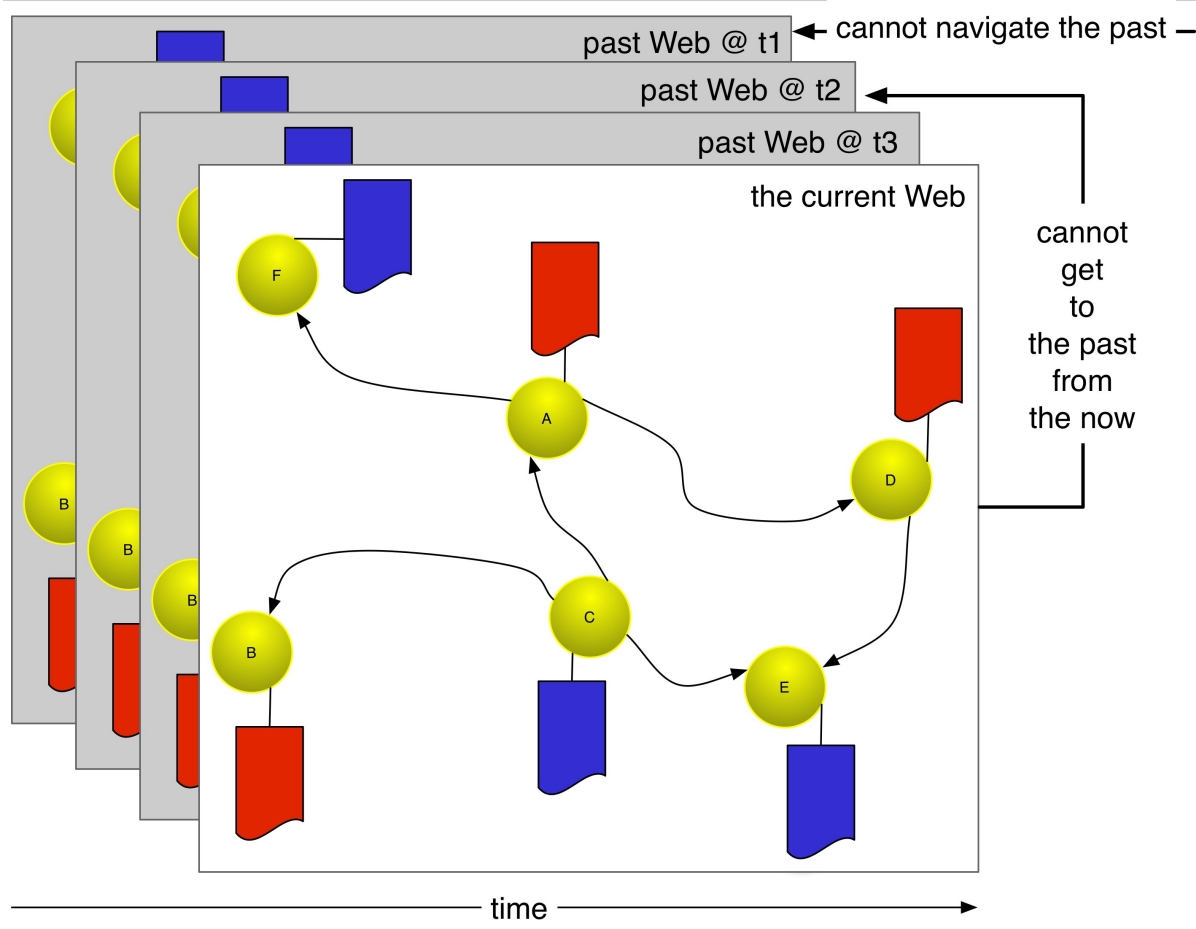
[http://en.wikipedia.org/w/index.php?title=September\\_11\\_attacks&oldid=282333](http://en.wikipedia.org/w/index.php?title=September_11_attacks&oldid=282333)  
archived resource for  
[http://en.wikipedia.org/wiki/September\\_11\\_attacks](http://en.wikipedia.org/wiki/September_11_attacks)



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# 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 the Current And Past Web.





# Vision: Navigate the Web of the Past

[http://en.wikipedia.org/wiki/  
Robots\\_exclusion\\_protocol](http://en.wikipedia.org/wiki/Robots_exclusion_protocol)



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Vision: Navigate the Web of the Past

[http://en.wikipedea.org/wiki/  
Robots\\_exclusion\\_protocol](http://en.wikipedia.org/wiki/Robots_exclusion_protocol)

Oct 11 2009, 05:30:33 UTC



Set browser time dial to ...



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Vision: Navigate the Web of the Past

[http://en.wikipedia.org/wiki/Robots\\_exclusion\\_protocol](http://en.wikipedia.org/wiki/Robots_exclusion_protocol)

Oct 11 2009, 05:30:33 UTC

You are viewing a Memento of [http://en.wikipedia.org/wiki/Robots\\_exclusion\\_protocol](http://en.wikipedia.org/wiki/Robots_exclusion_protocol) as at Sun, 11 Oct 2009 05:30:43 GMT, archived at: <http://en.wikipedia.org/w/index.php?oldid=314187708> [hide this]

## Robots exclusion standard

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by 128.189.120.54 (talk) at 20:49, 15 September 2009. It may differ significantly from the current revision.  
(diff) ← Previous revision | Current revision (diff) | Newer revision → (diff)

For restricting Wikipedia bots, see [Template:Bots](#).

The **Robot Exclusion Standard**, also known as the **Robots Exclusion Protocol** or **robots.txt protocol**, is a convention to prevent cooperating [web spiders](#) and other [web robots](#) from accessing all or part of a [website](#) which is otherwise publicly viewable. Robots are often used by [search engines](#) to categorize and archive web sites, or by webmasters to proofread source code. The standard is unrelated to, but can be used in conjunction with, [sitemaps](#), a robot *inclusion* standard for websites.

### Contents

- History
- About the standard
- Disadvantages
- Automated Content Access Protocol
- Examples
- Nonstandard extensions
  - 6.1 Crawl-delay directive
  - 6.2 Allow directive
  - 6.3 Sitemap
- Extended standard
- See also
- References
- External links

### History

robots.txt was popularized with the advent of [AltaVista](#), the first popular search engine.

### About the standard

If a site owner wishes to give instructions to web robots he must place a text file called `robots.txt` to the root of the web site hierarchy (e.g. `www.example.com/robots.txt`). This text file should contain the instructions in a specific format (see examples below). Robots that **wish** to follow the instructions try to fetch this file and read the instructions before fetching any other file from the web site. If this file doesn't exist web robots assume that the web owner wishes to provide no specific instructions.

A robots.txt file on a website will function as a request that specified robots ignore specified files or directories in their search. This might be, for example, out of a preference for privacy from search engine results, or the belief that the content of the selected directories might be misleading or irrelevant to the categorization of the site as a whole, or out of a desire that an application only operate on certain data.

For websites with multiple subdomains, each subdomain must have its own robots.txt file. If `example.com` had a robots.txt file but `a.example.com` did not, the rules that would apply for `example.com` would not apply to `a.example.com`.

### Disadvantages

From Wikipedia History: Version Sep 15 2009, 20:49:00 UTC



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC





# Vision: Navigate the Web of the Past

http://www.robotstxt.org/

Oct 11 2009, 05:30:33 UTC



Browser time dial still at ...



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Vision: Navigate the Web of the Past

http://www.robotstxt.org/

Oct 11 2009, 05:30:33 UTC

Browser time dial still at ...

You are viewing a Memento of <http://www.robotstxt.org/wc/exclusion.html> as at Sun, 11 Oct 2009 05:30:43 GMT, archived at: <http://web.archive.org/web/20071109062104/> [hide this]

Navigation

<META> tags

Frequently Asked

About robotstxt.org

Tools

/robots.txt checker

Robots

Web Robots (also known as Web Wanderers, Crawlers, or Spiders), are programs that traverse the Web automatically. Search engines such as Google use them to index the web content, spammer use them to scan for email addresses, and they have many other uses.

On this site you can learn more about web robots.

- [About /robots.txt](#) explains what /robots.txt is, and how to use it.
- The [FAQ](#) answers many frequently asked questions, such as [How do I stop robots visiting my site?](#) and [How can I get the best listing in search engines?](#)
- The [Other Sites](#) page links to external resources for robot writers and a webmasters.
- The [Robots Database](#) has a list of robots.
- The [/robots.txt checker](#) can check your site's /robots.txt file and meta tags.
- The [IP Lookup](#) can help find out more about what robots are visiting you.

Last updated: 08 Nov 2007 15:08:21

Advertisement

[About this site](#) | [Privacy and cookies policy](#) | [Contact us](#) | © 2007. All rights reserved

From Internet Archive: Version Nov 09 2007, 06:21:04 UTC



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# The Memento Approach

*HTTP navigation to an archived resource by leveraging:*

- *The original resource;*
- *HTTP datetime content negotiation.*



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



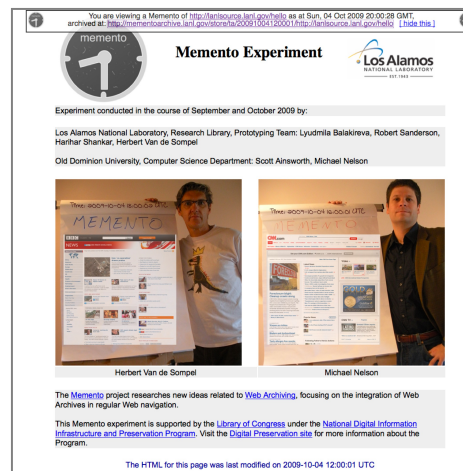




original  
resource



Mementos



← original server →

← archival server →



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC







← DT-conneg with URI-G to get URI-M →

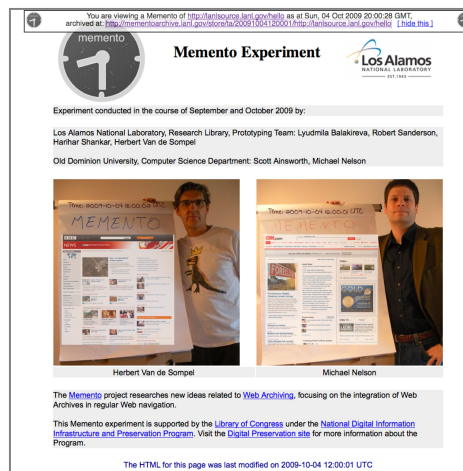
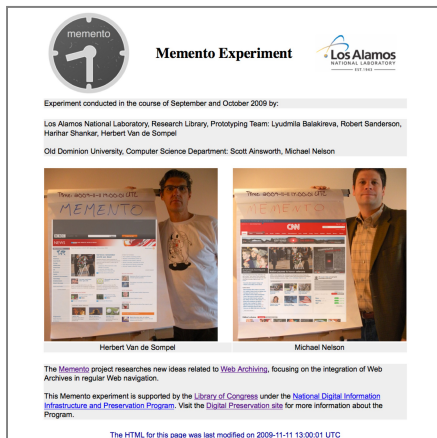
original  
resource

TimeGate

transparently  
negotiable  
resource

Mementos

variant  
resources



original server

archival server



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC





HTTP  
Link  
timegate



DT-conneg with URI-G to get URI-M



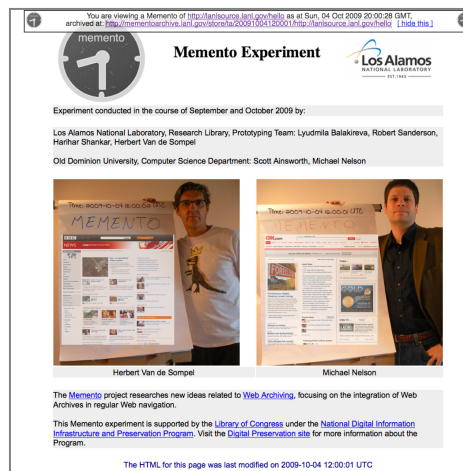
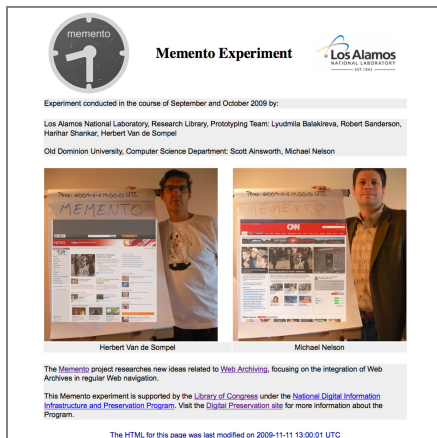
original  
resource

TimeGate

transparently  
negotiable  
resource

Mementos

variant  
resources



original server

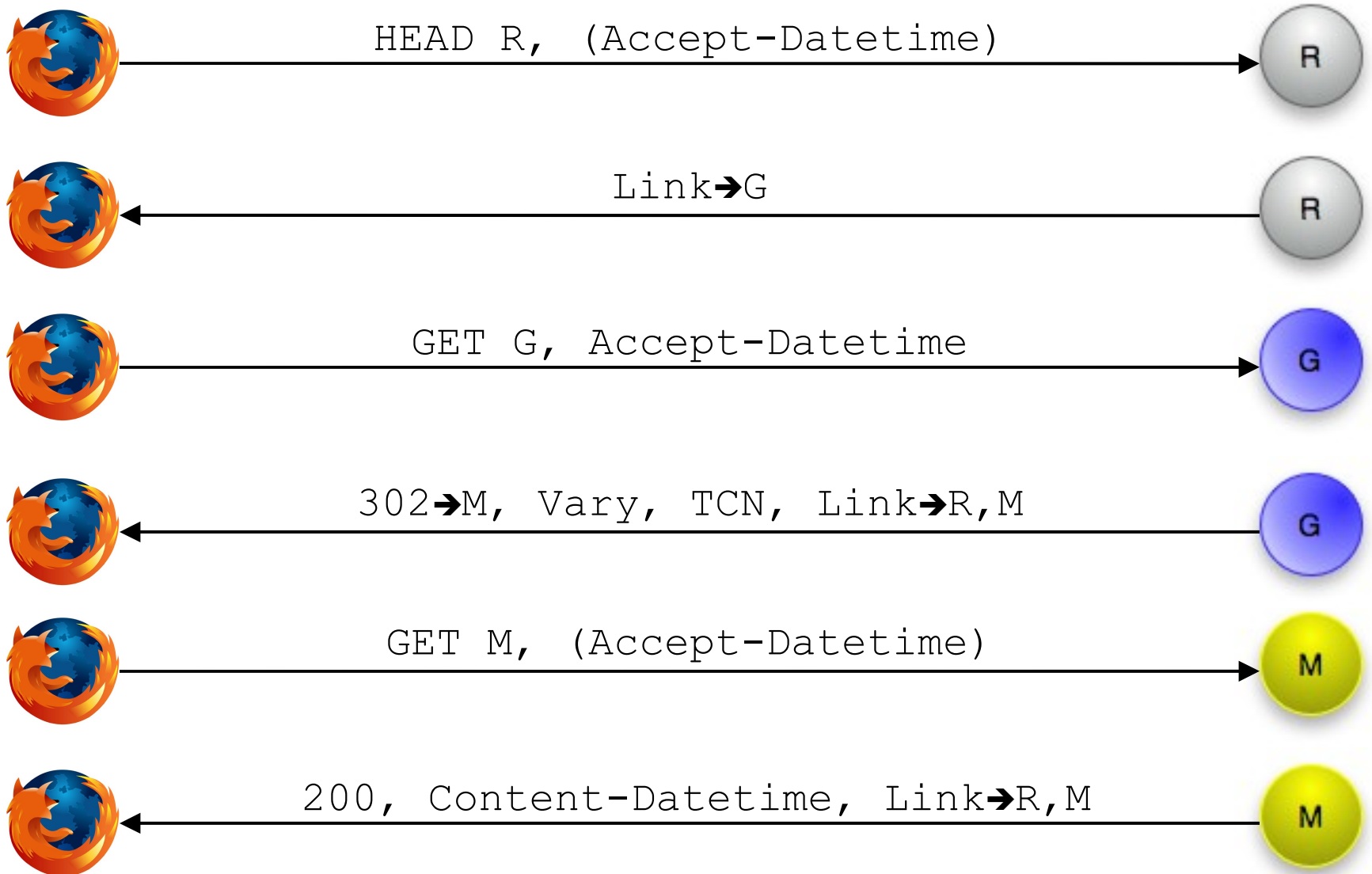
archival server



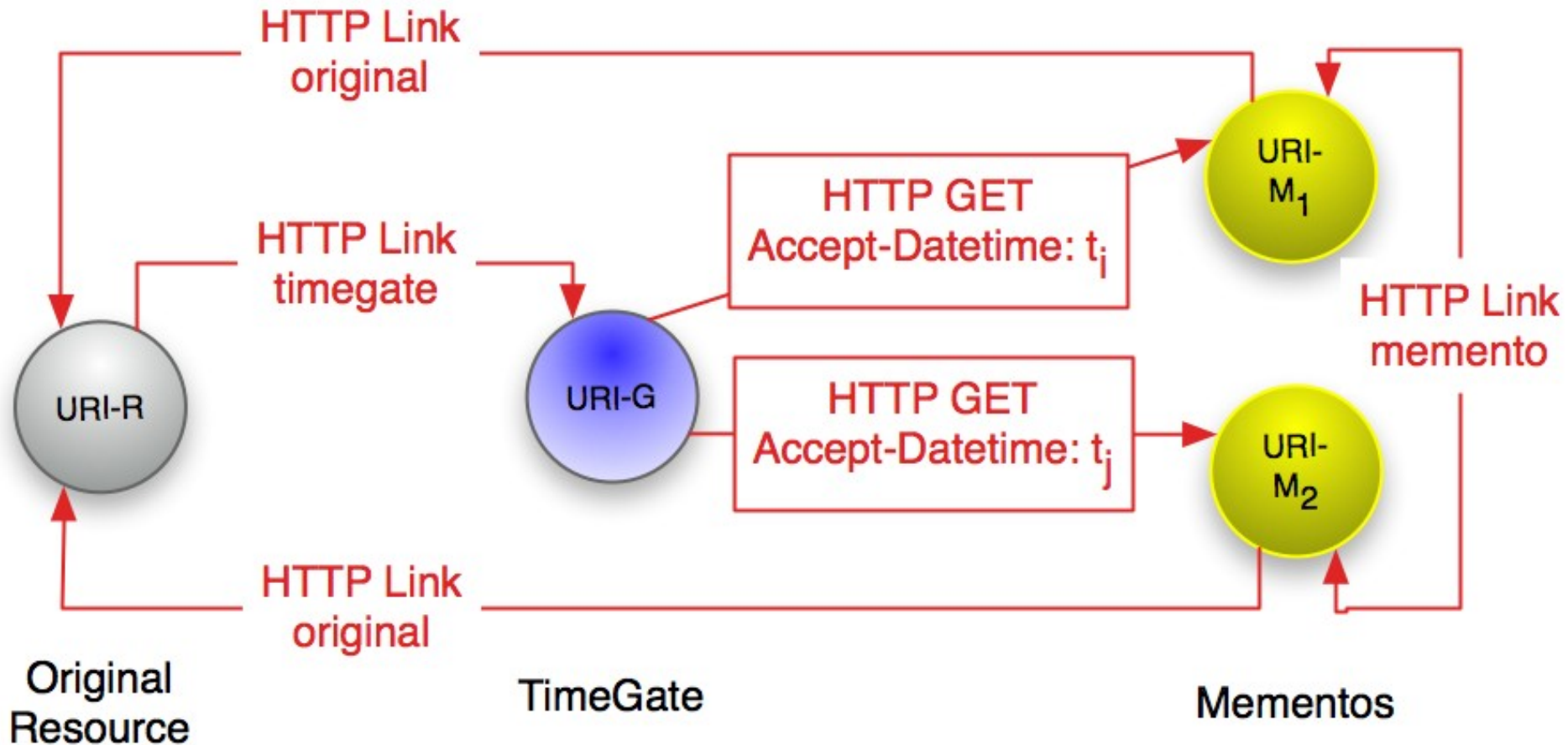
An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Memento HTTP Flow



# The Memento Framework



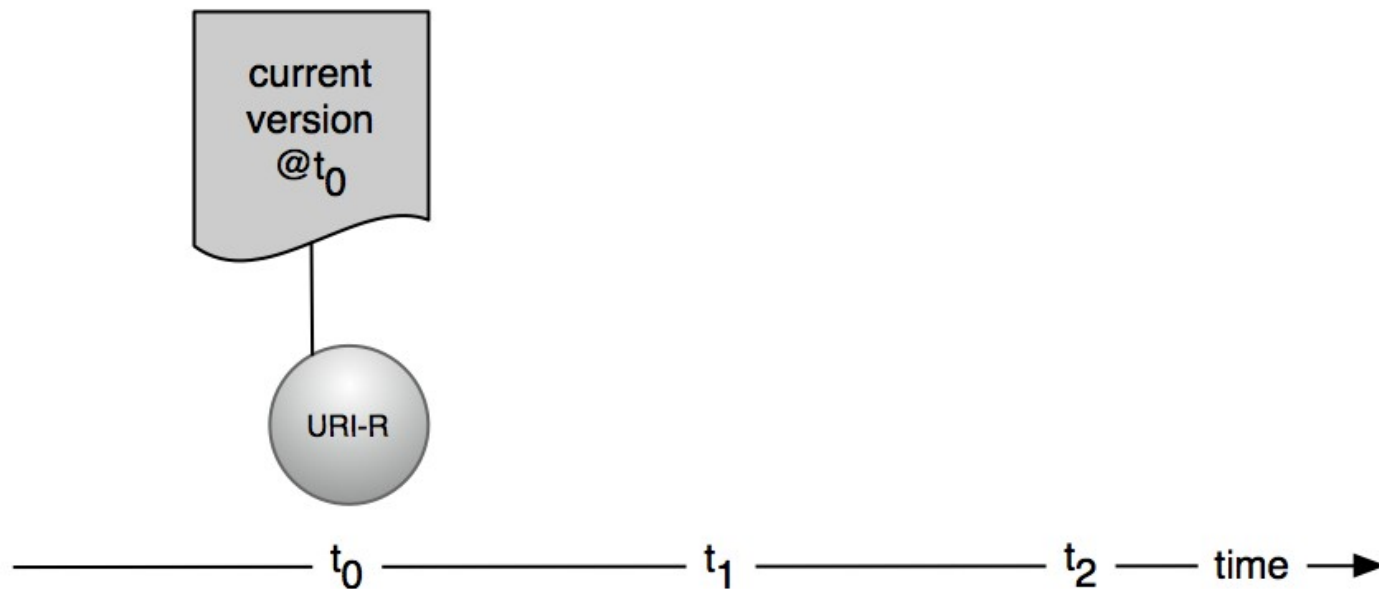
# Outline

- Memento - Time Travel for the Web
- Resource Versioning suggested by Memento
- Resource Versioning for Linked Data
- DBpedia Demonstrator

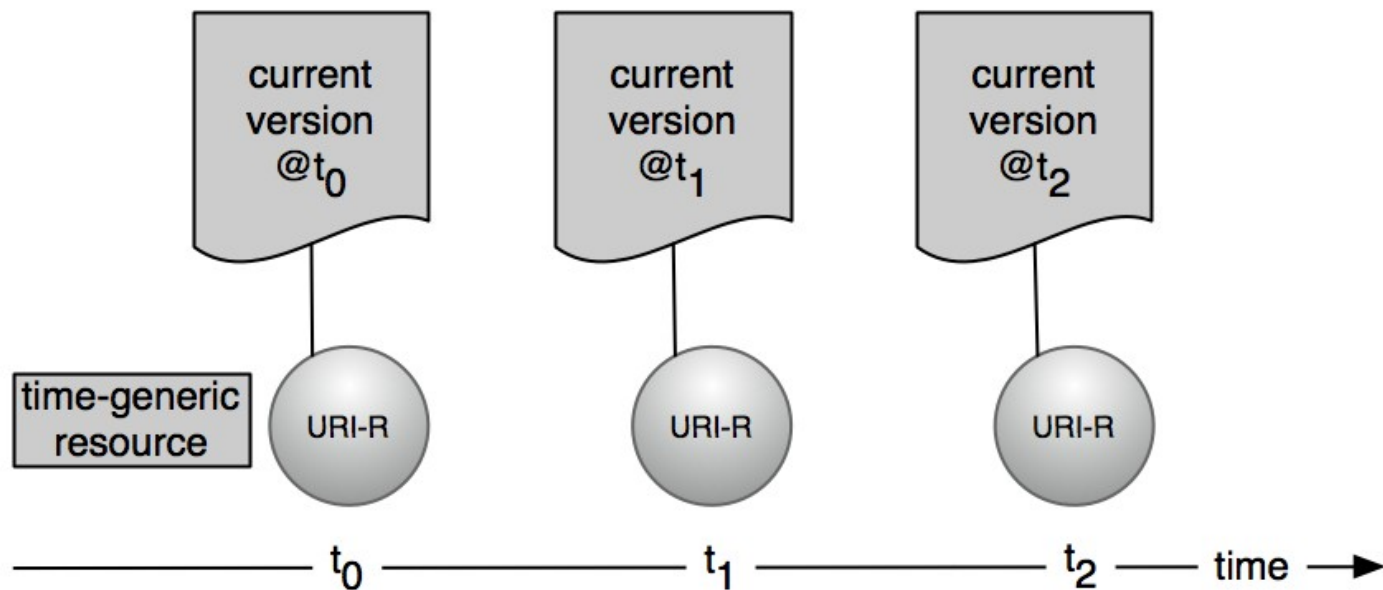


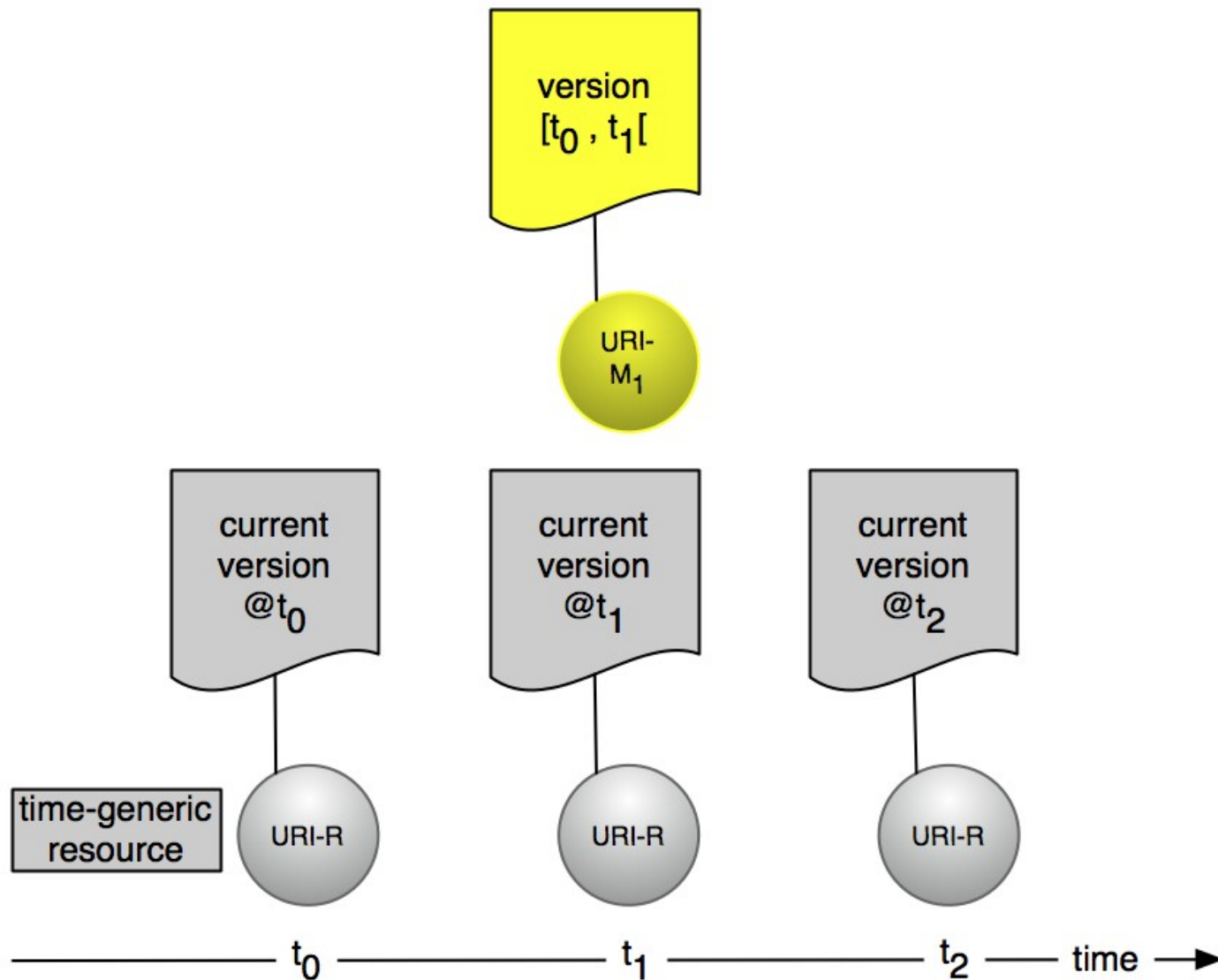
An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



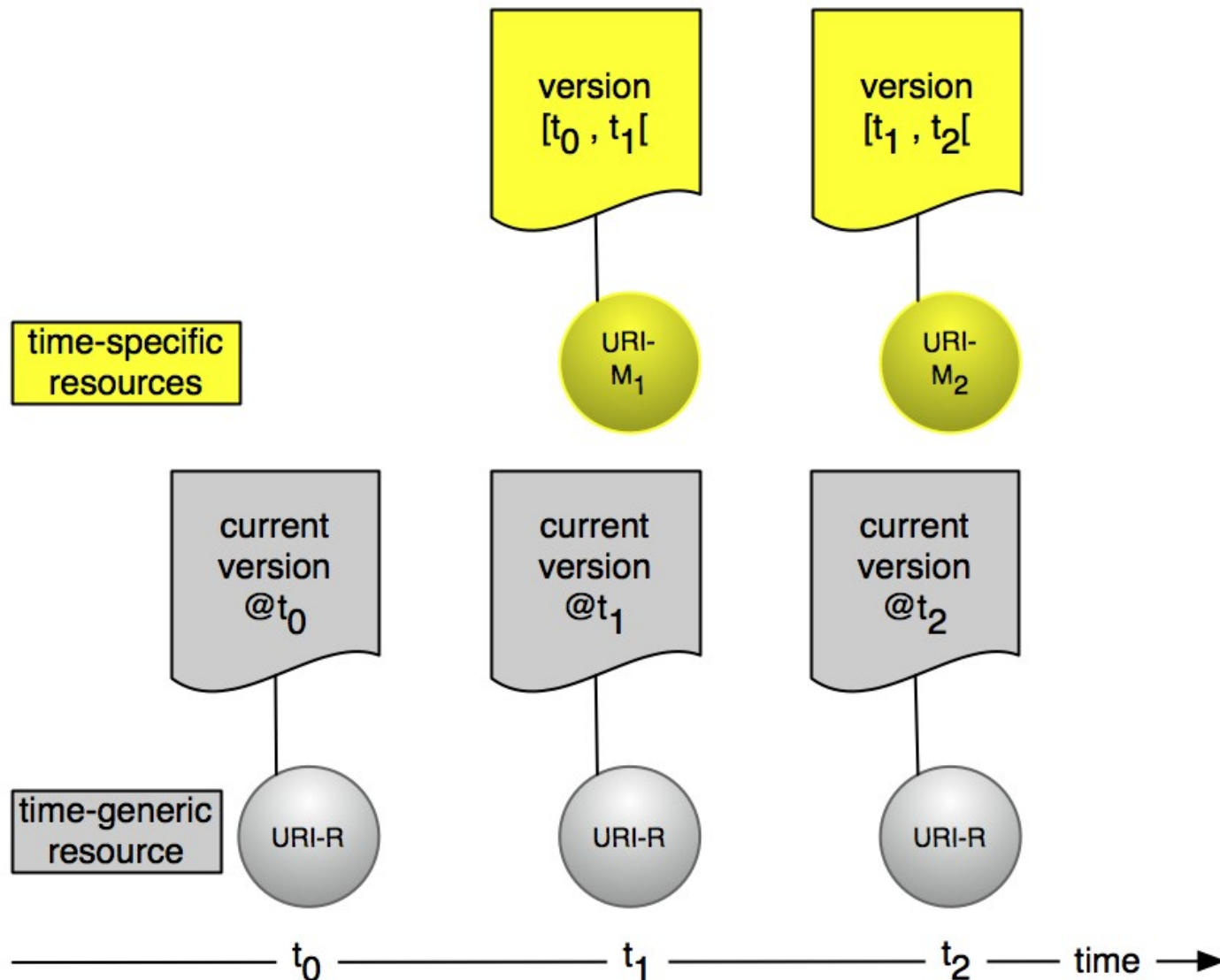


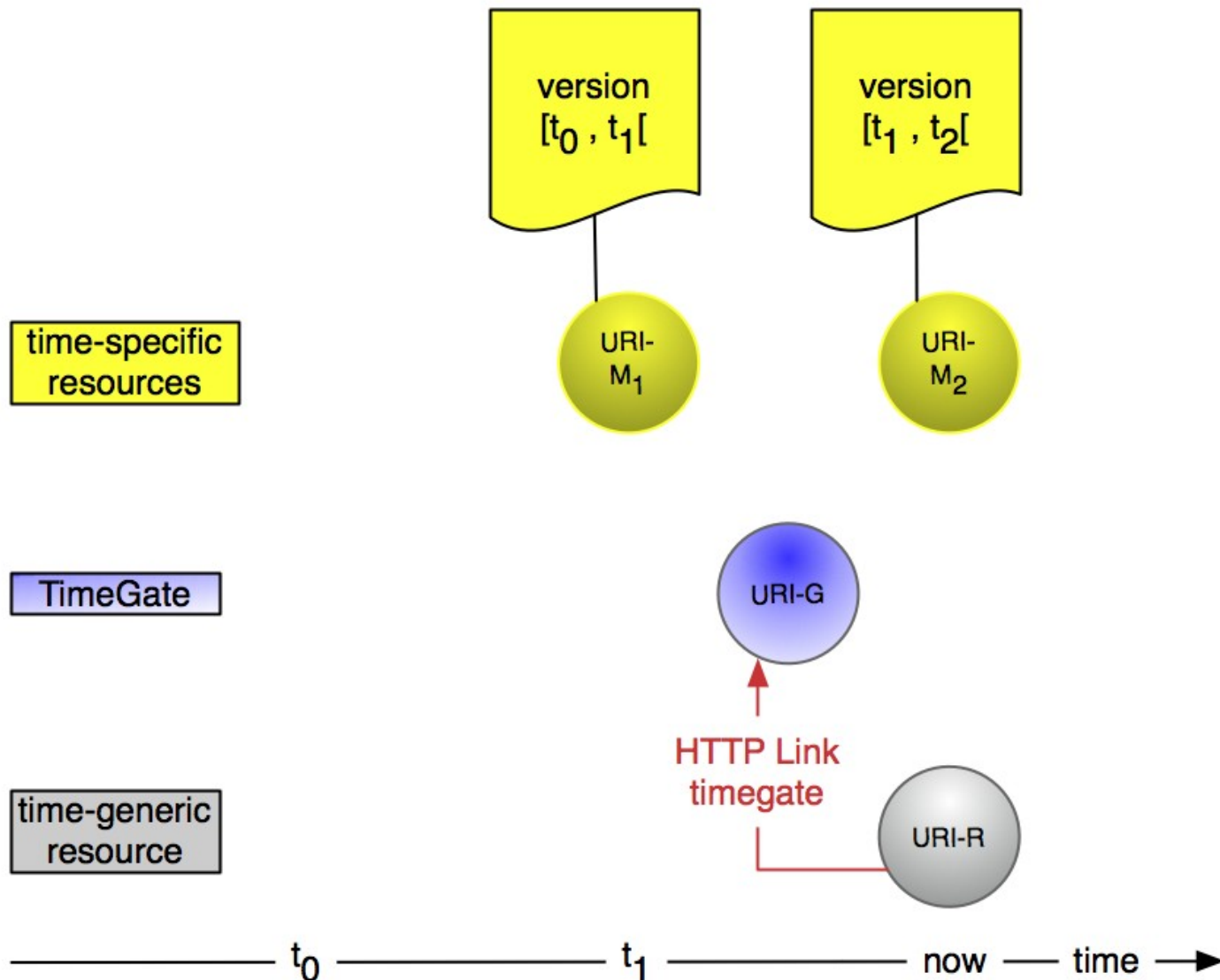


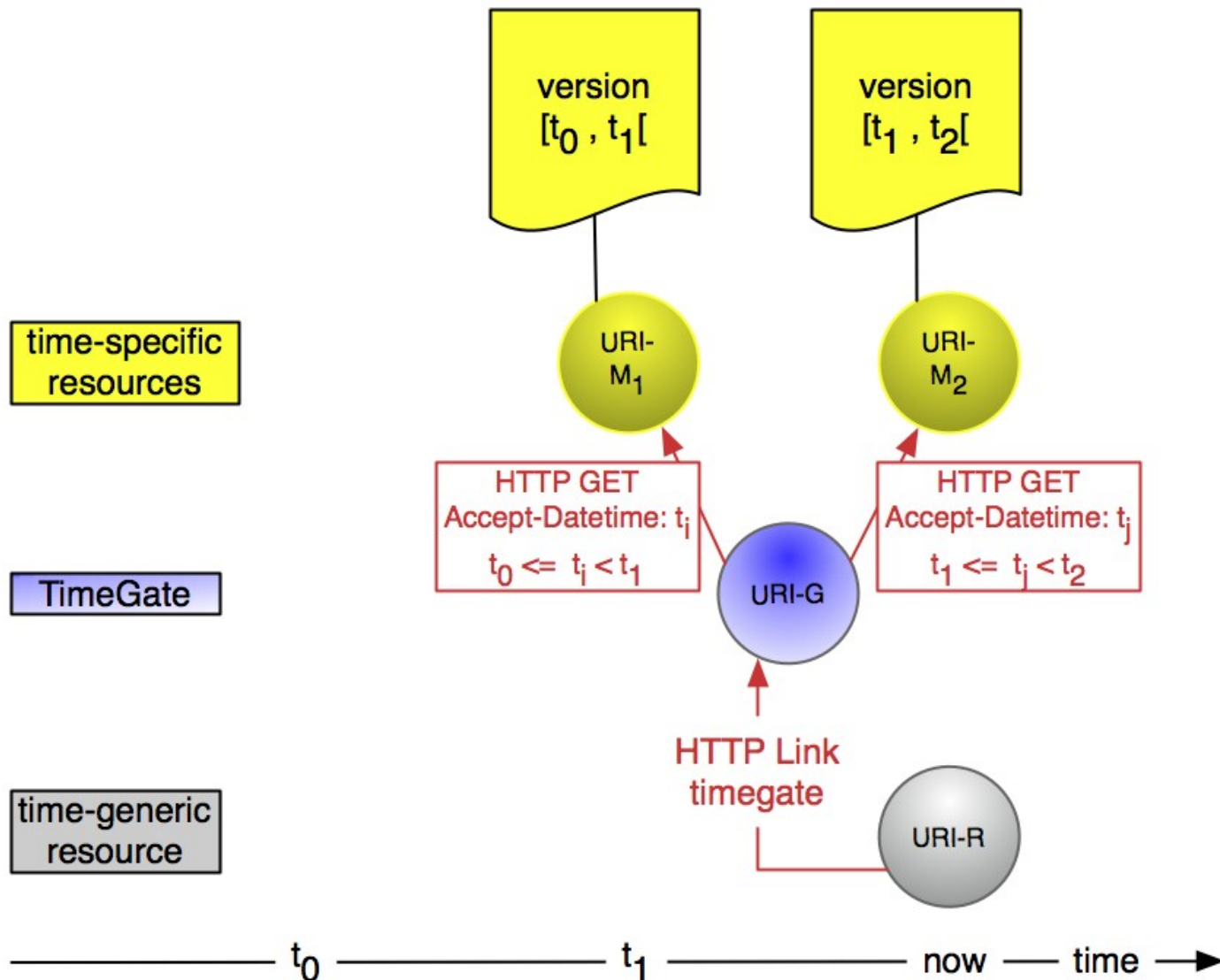


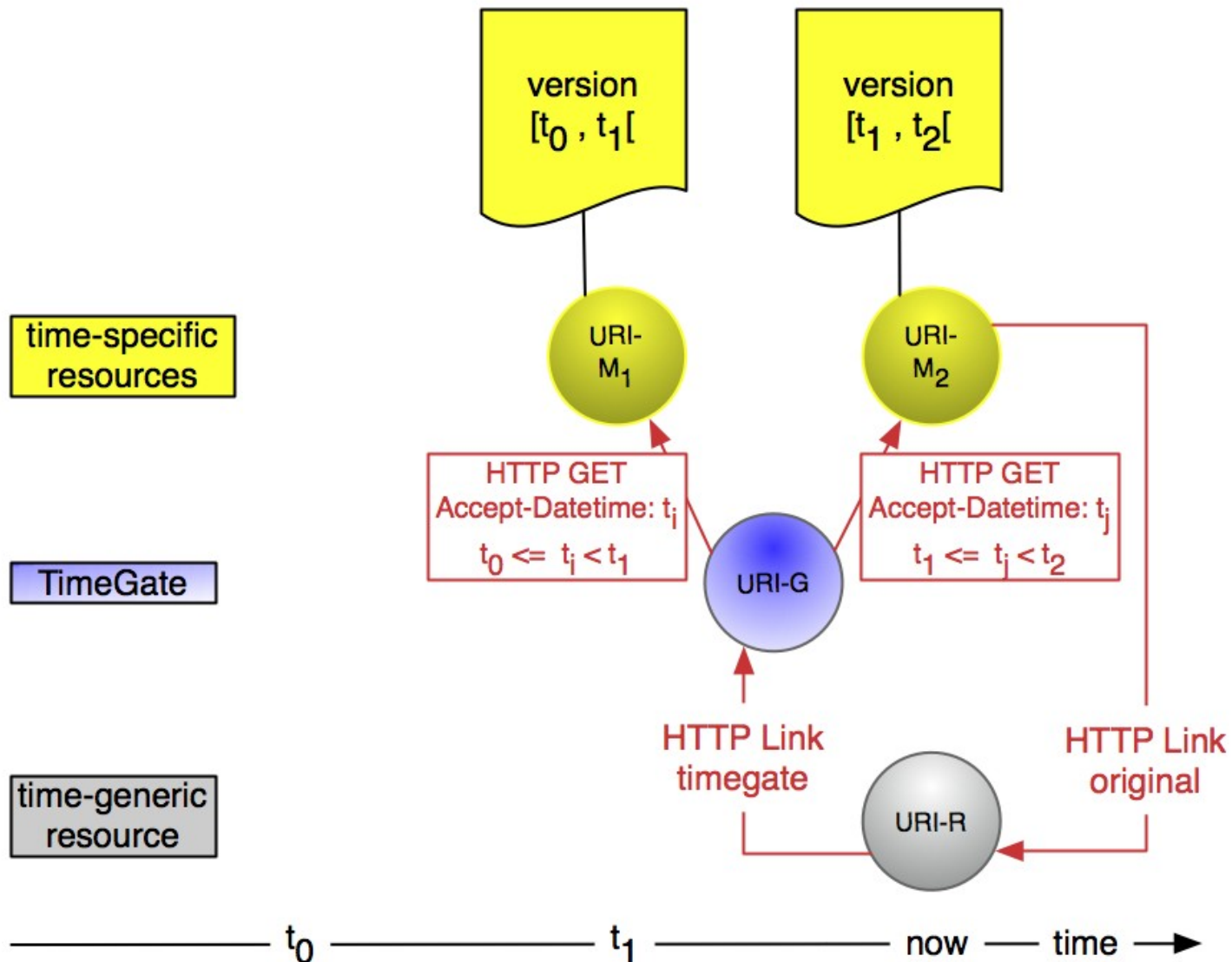


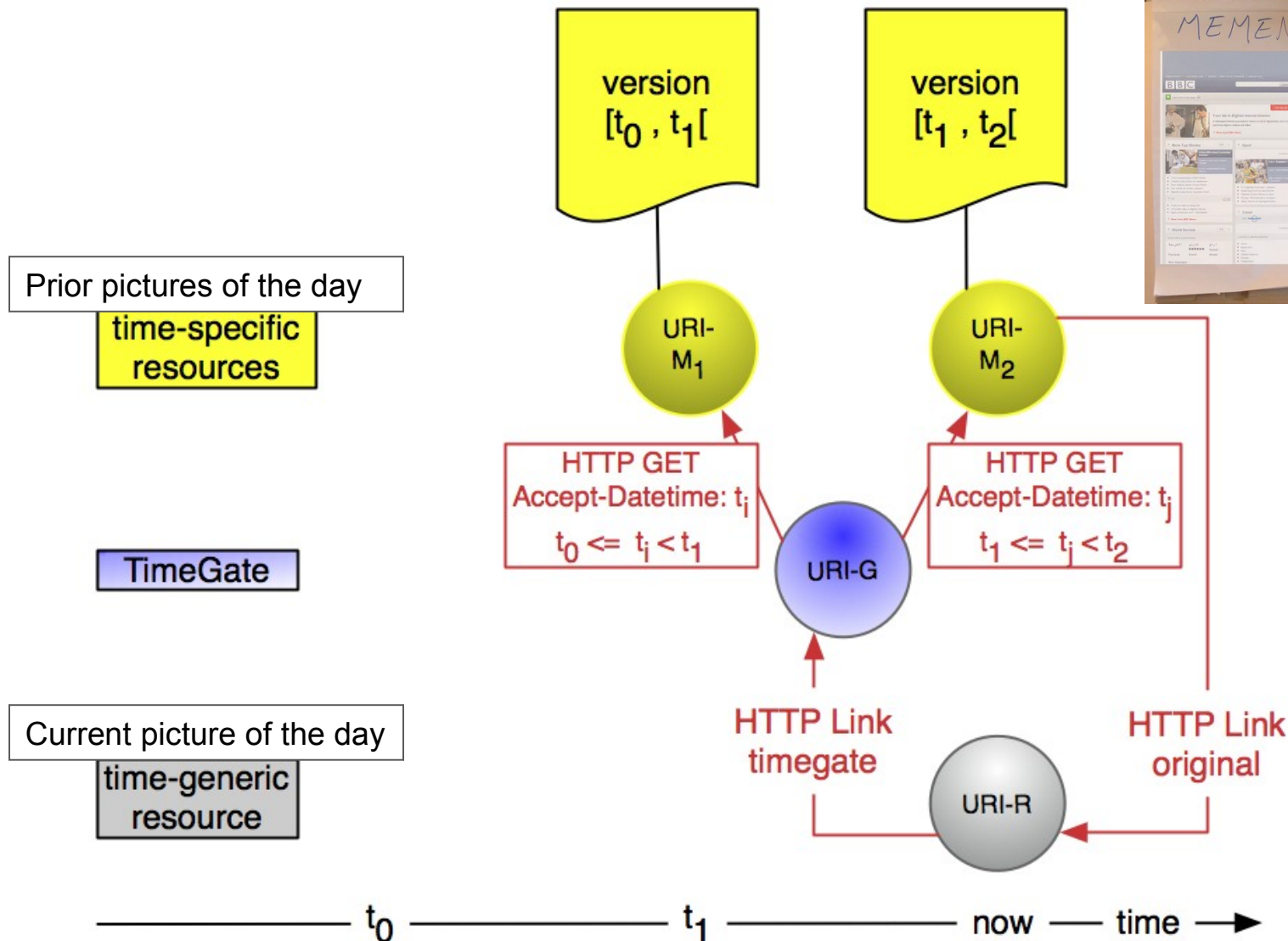












# Time Travel across Versions of a *Picture of the Day*



Data collected through HTTP Navigation



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Outline

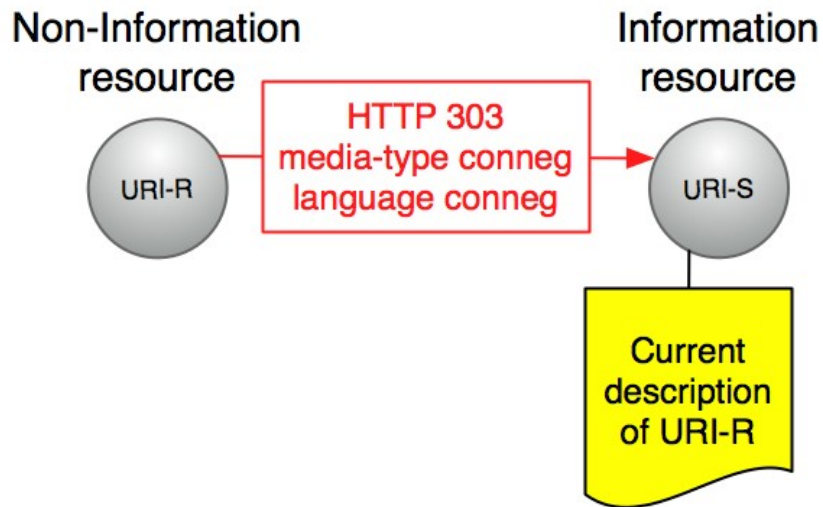
- Memento - Time Travel for the Web
- Resource Versioning suggested by Memento
- Resource Versioning for Linked Data
- DBpedia Demonstrator



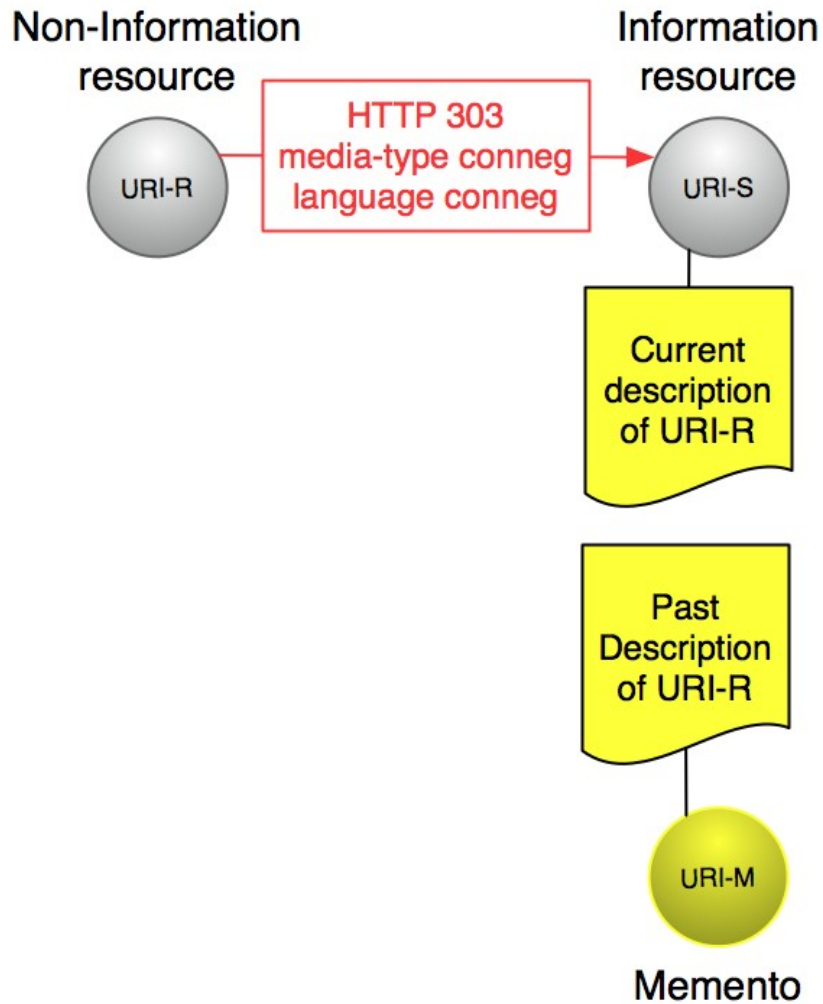
An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC

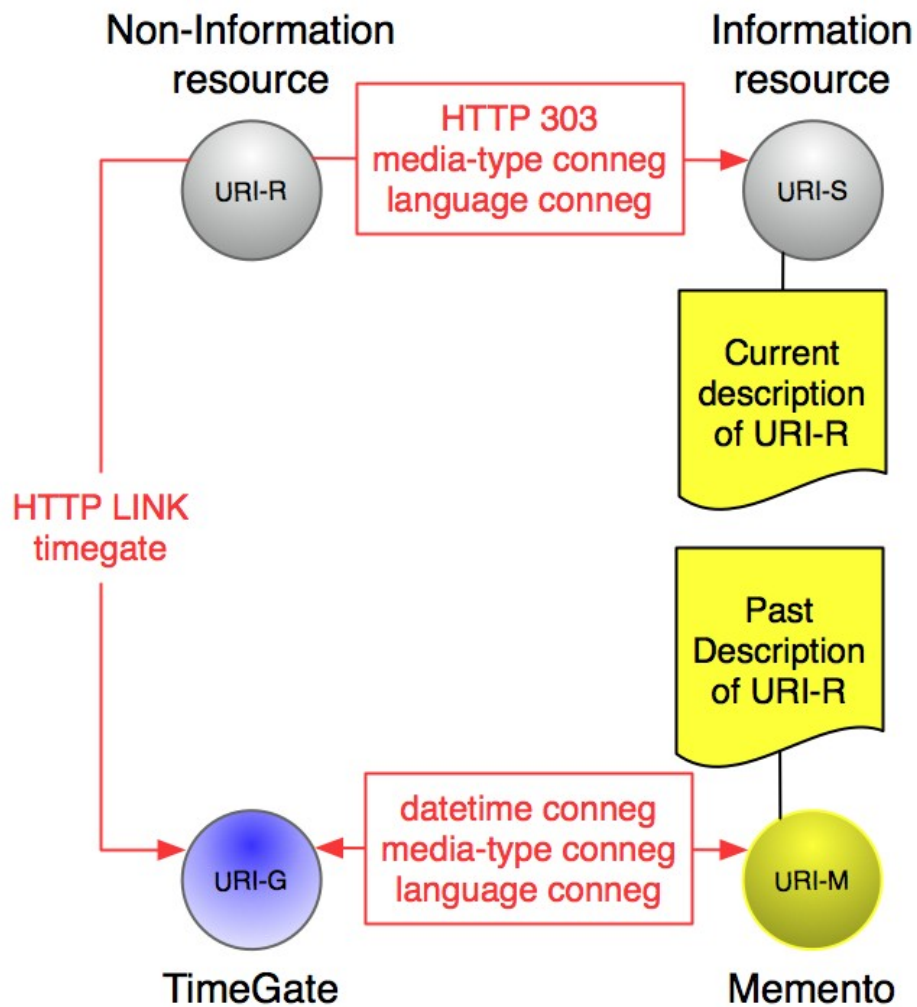


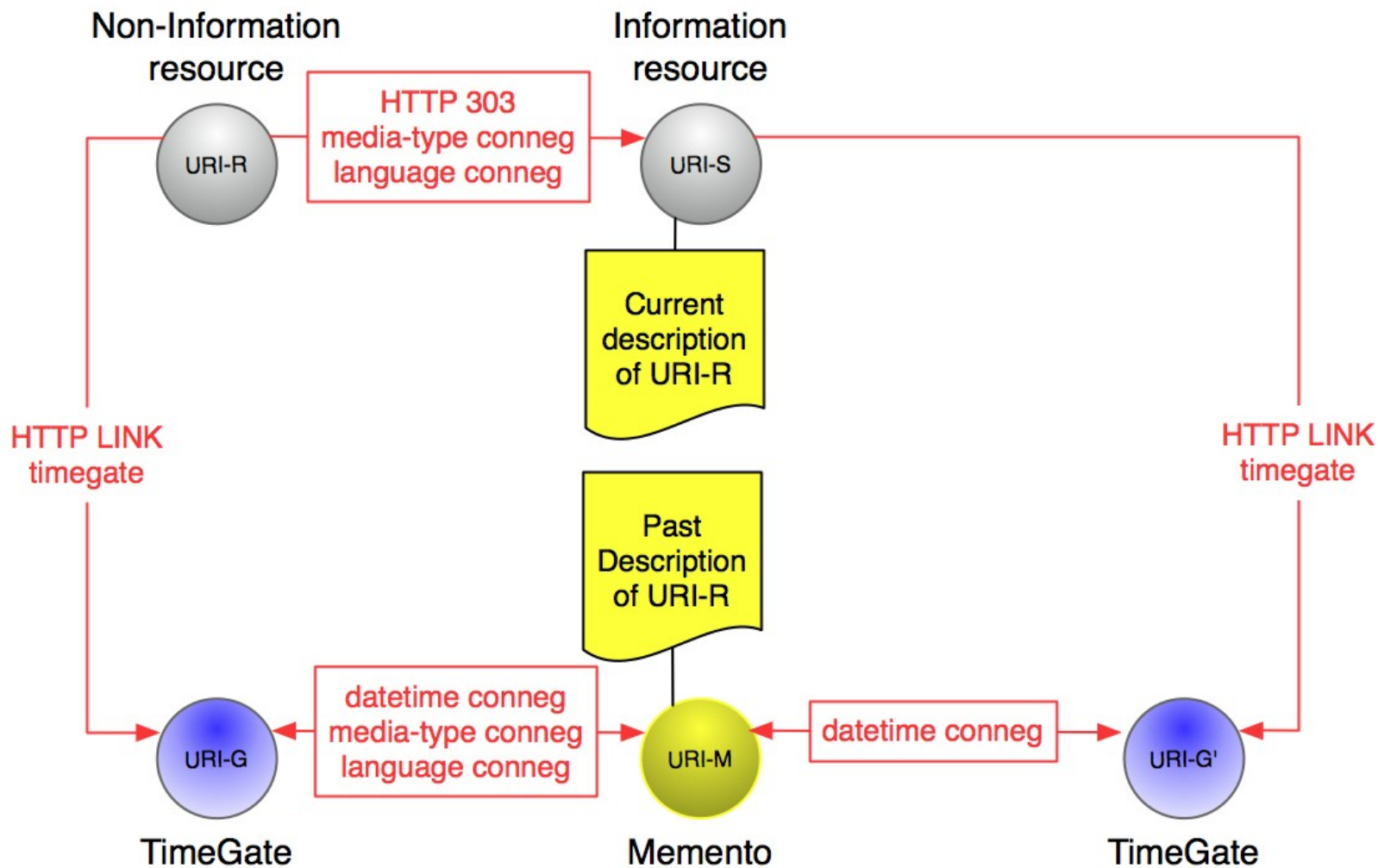












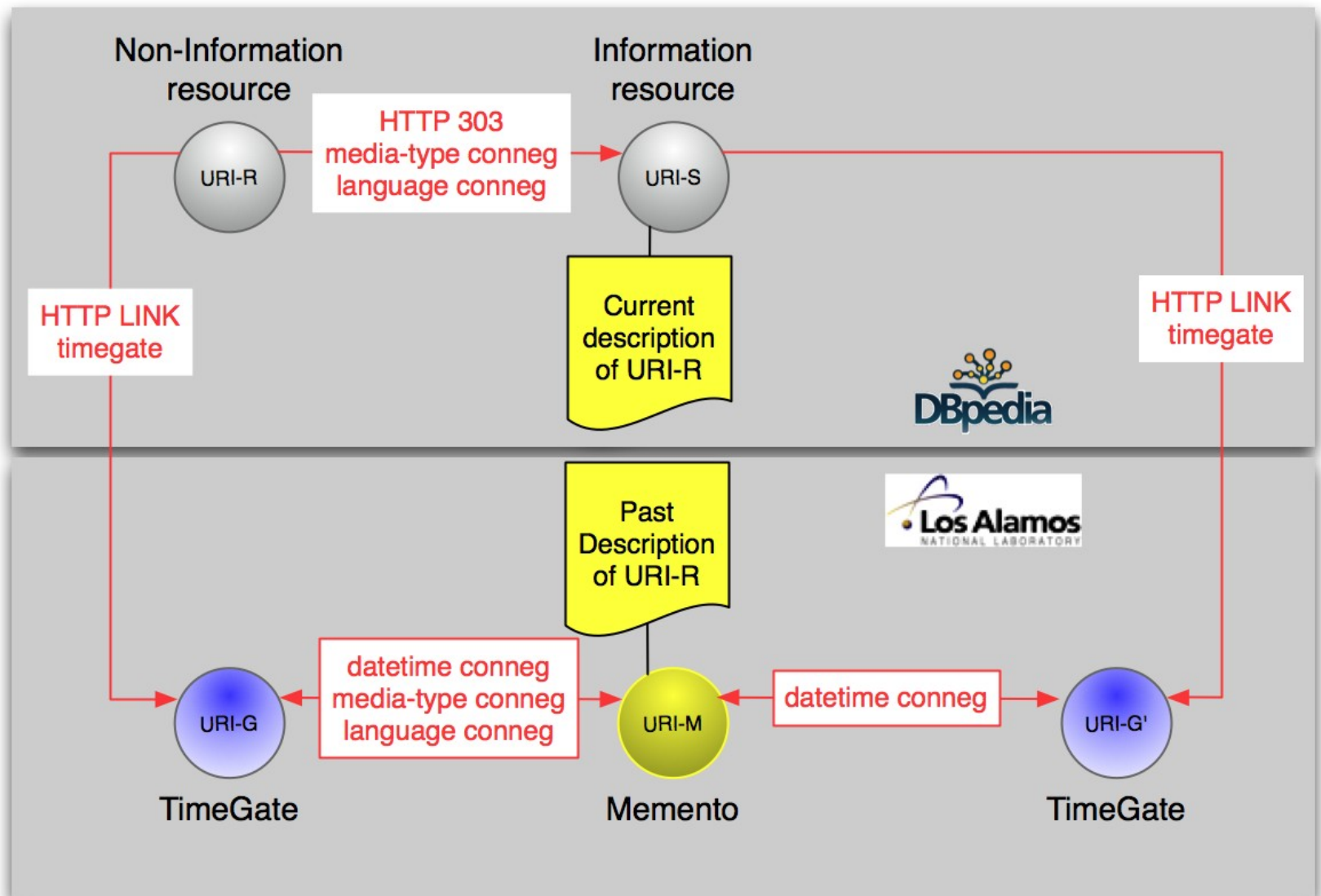
# Outline

- Memento - Time Travel for the Web
- Resource Versioning suggested by Memento
- Resource Versioning for Linked Data
- DBpedia Demonstrator

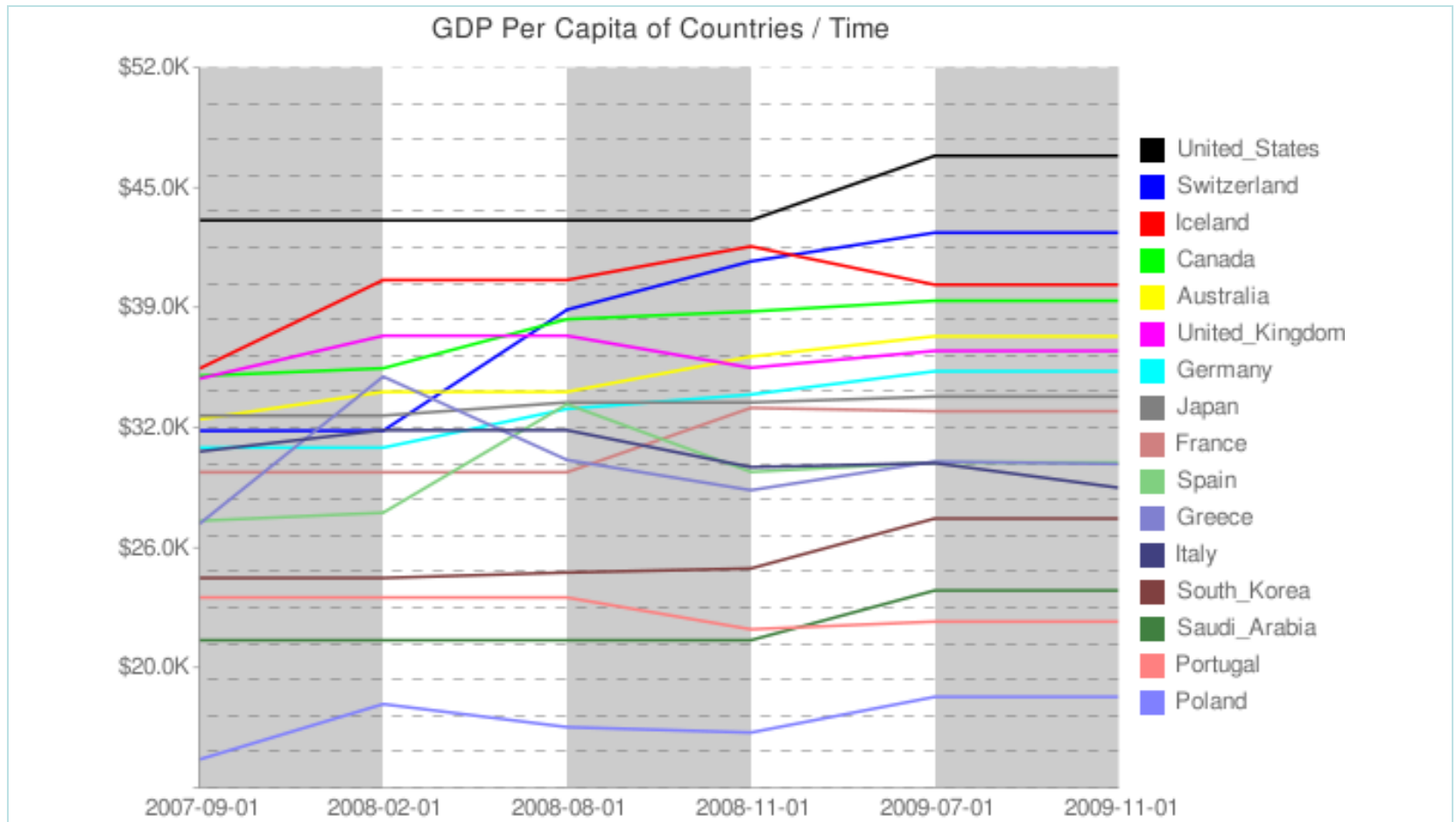


An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC





# Time-Series Analysis across DBpedia Versions



Data collected through HTTP Navigation

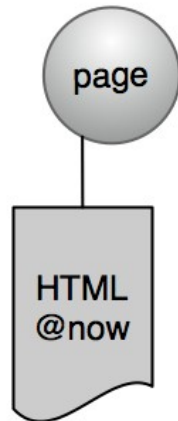


An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Conclusions

`http://weather.example.com/oxaca`



URI as access point to page

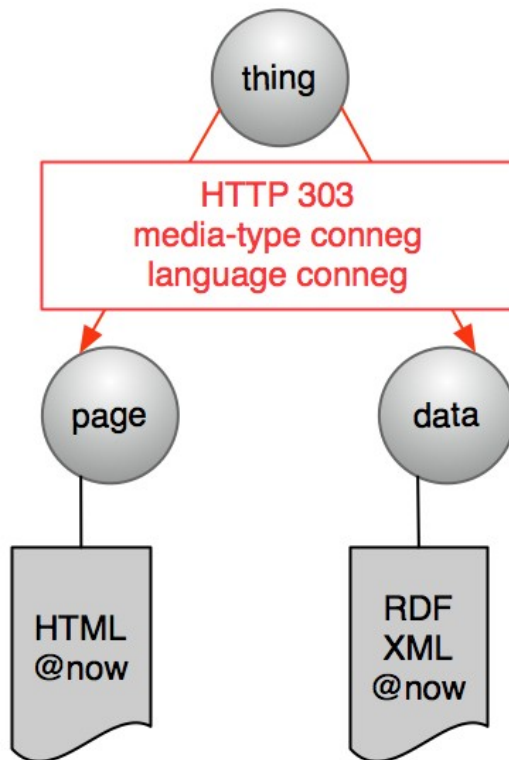


An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC



# Conclusions

<http://weather.example.com/oxaca>

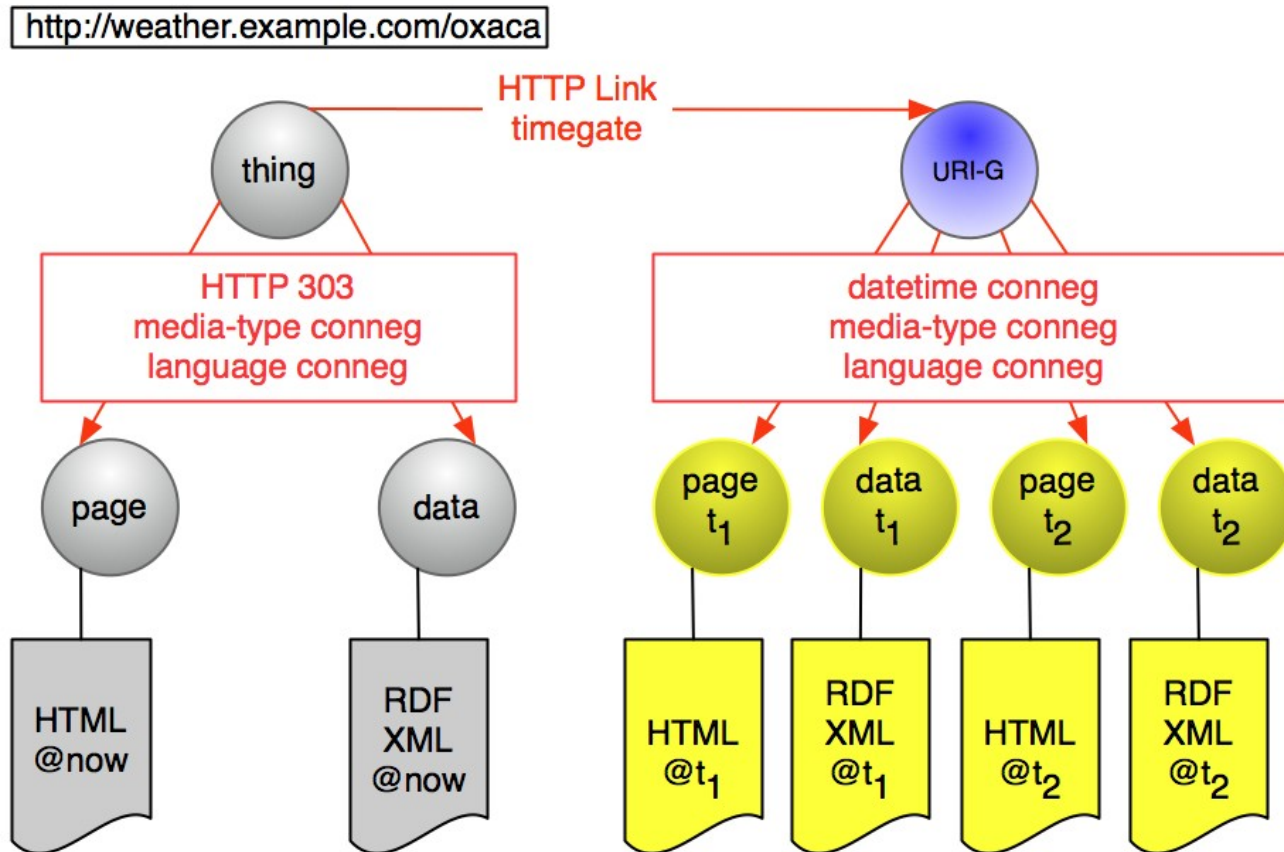


Increased value: URI as access point to page and data





# Conclusions



Increased value: URI as access point to current & historical page and data



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



<http://www.mementoweb.org>

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



An HTTP-Based Versioning Mechanism for Linked Data  
LDOW 2010, Raleigh, NC

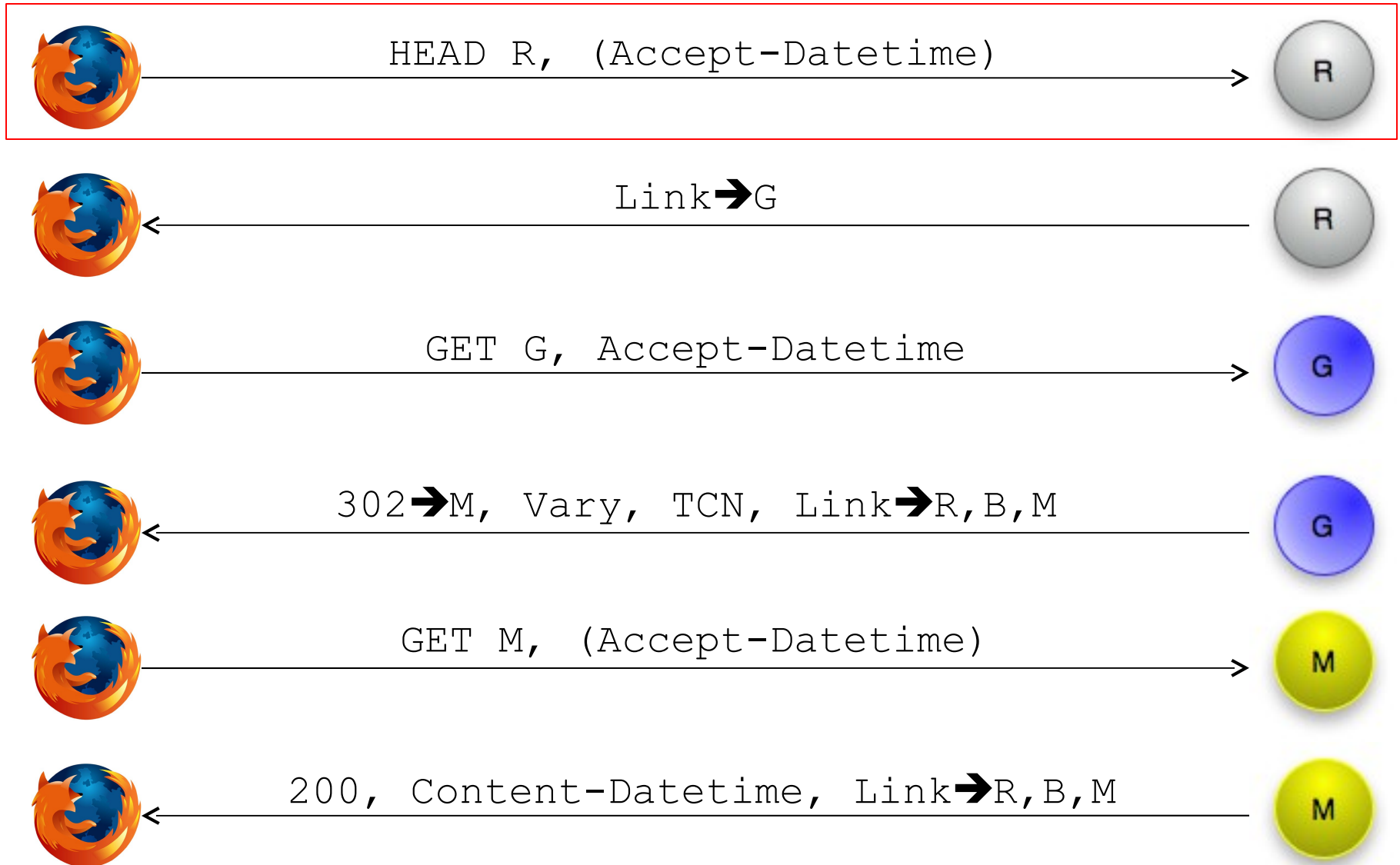


# References

- Tim Berners-Lee (1996,2000) Generic Resources.  
<http://www.w3.org/DesignIssues/Generic.html>
- Van de Sompel, H., Sanderson, R., Nelson, M.L., Balakireva, L., Ainsworth, S., Shankar, H. (2010) An HTTP-Based Versioning Mechanism for Linked Data. Proceedings of the 3rd Workshop on Linked Data on the Web.  
<http://arxiv.org/abs/1003.3661>
- Sanderson, R., and Van de Sompel, H. (2010) Making Web Annotations Persistent over Time. Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital libraries.  
<http://arxiv.org/abs/1003.2643>
- Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L., Ainsworth, S., Shankar, H. (2009) Memento: Time Travel for the Web.  
<http://arxiv.org/abs/0911.1112>



# Memento HTTP Flow



# Memento HTTP Flow: URI-R

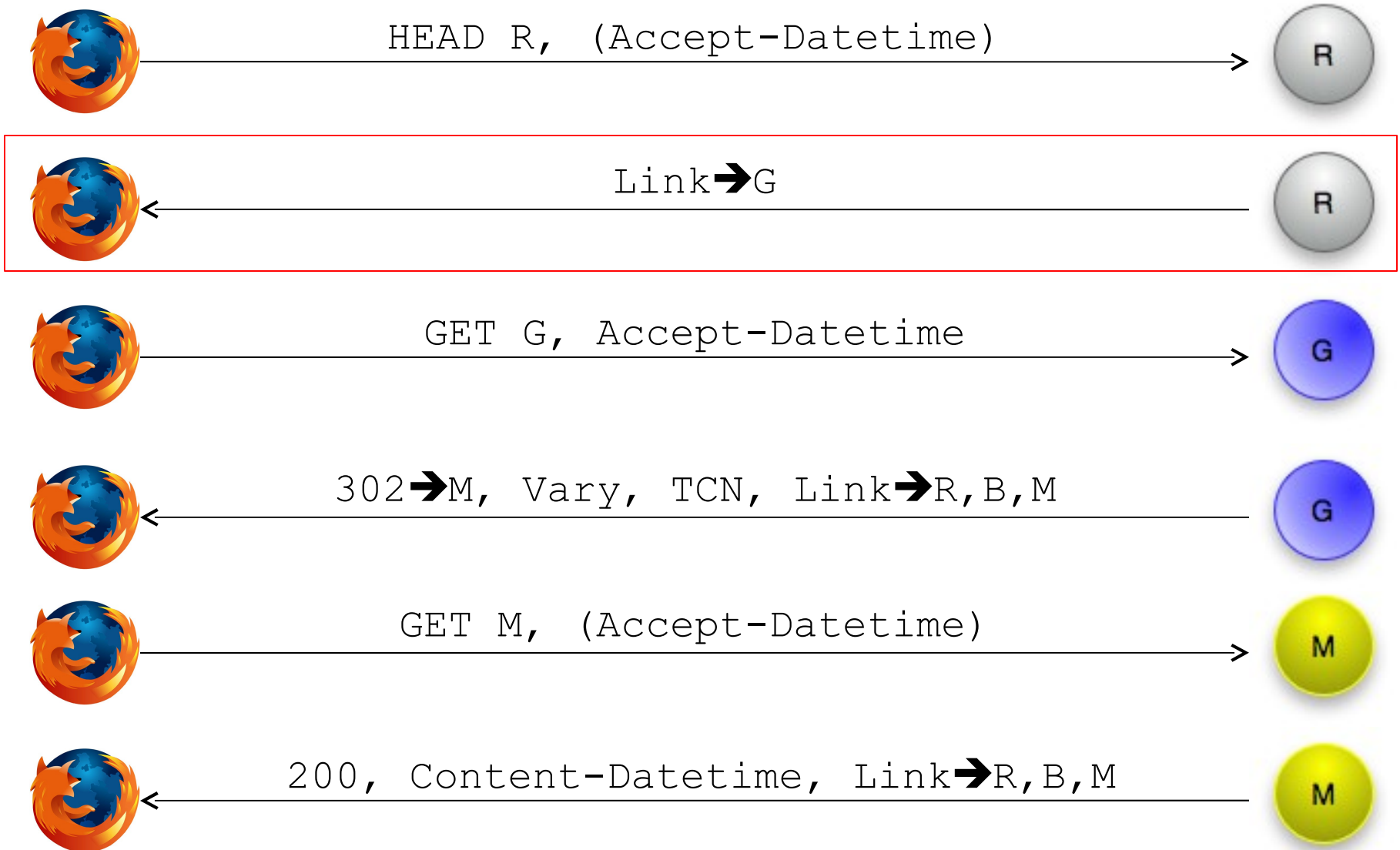


HEAD R, (Accept-Datetime)



```
HEAD /resource/France HTTP/1.1
Host: dbpedia.org
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
```

# Memento HTTP Flow



# Memento HTTP Flow: Success – URI-R



Link → G



HTTP/1.1 303 See Other

Server: Virtuoso/06.01.3127 (Solaris) x86\_64-sun-solaris2.10-64 VDB

Connection: close

Content-Type: text/html; charset=UTF-8

Date: Tue, 20 Apr 2010 16:48:51 GMT

Accept-Ranges: bytes

Location: <http://dbpedia.org/page/France>

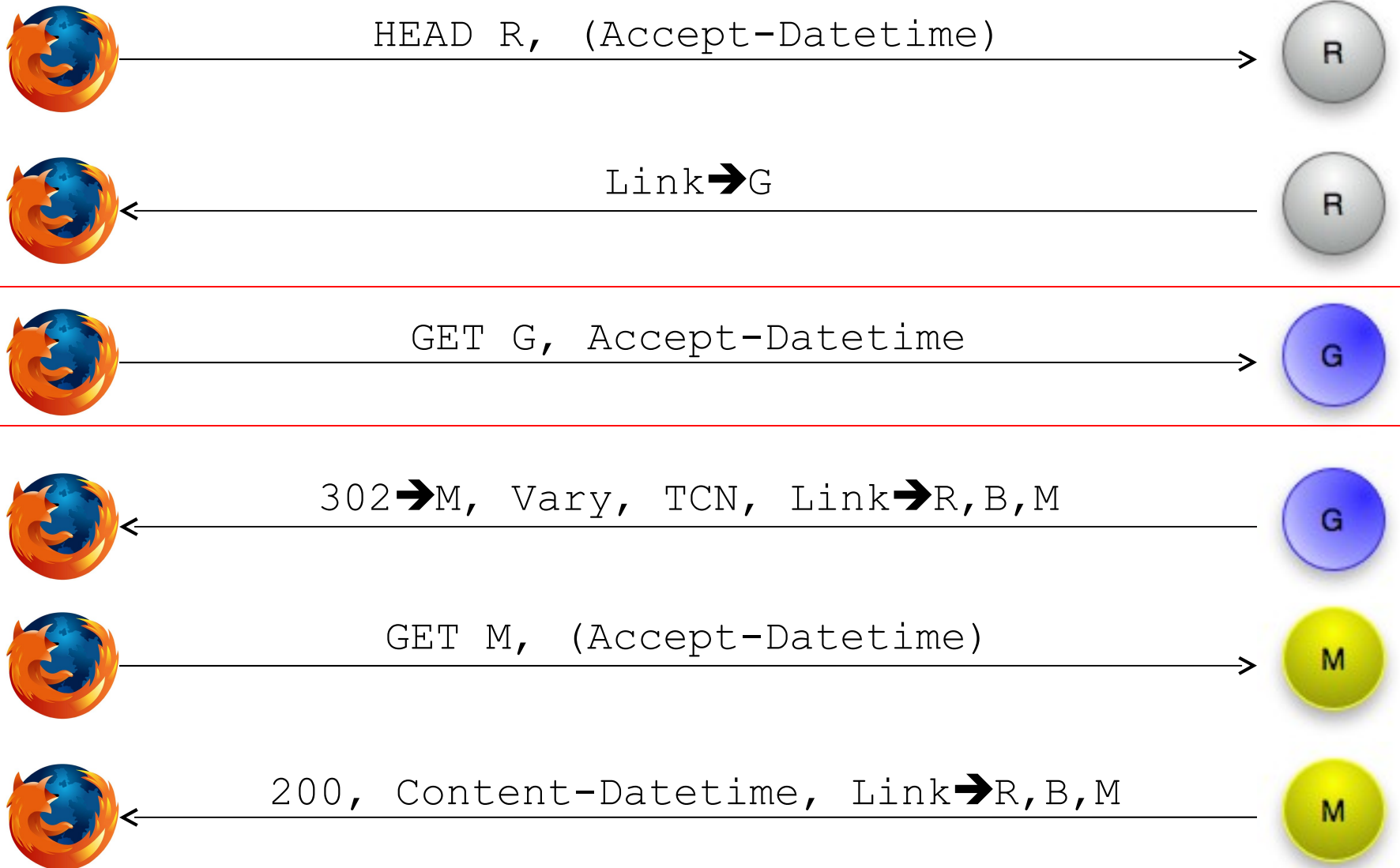
Transfer-Encoding: chunked

Content-Encoding: gzip

Link:

<<http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/resource/France>>  
; rel="timegate"

# Memento HTTP Flow





# Memento HTTP Flow: URI-G

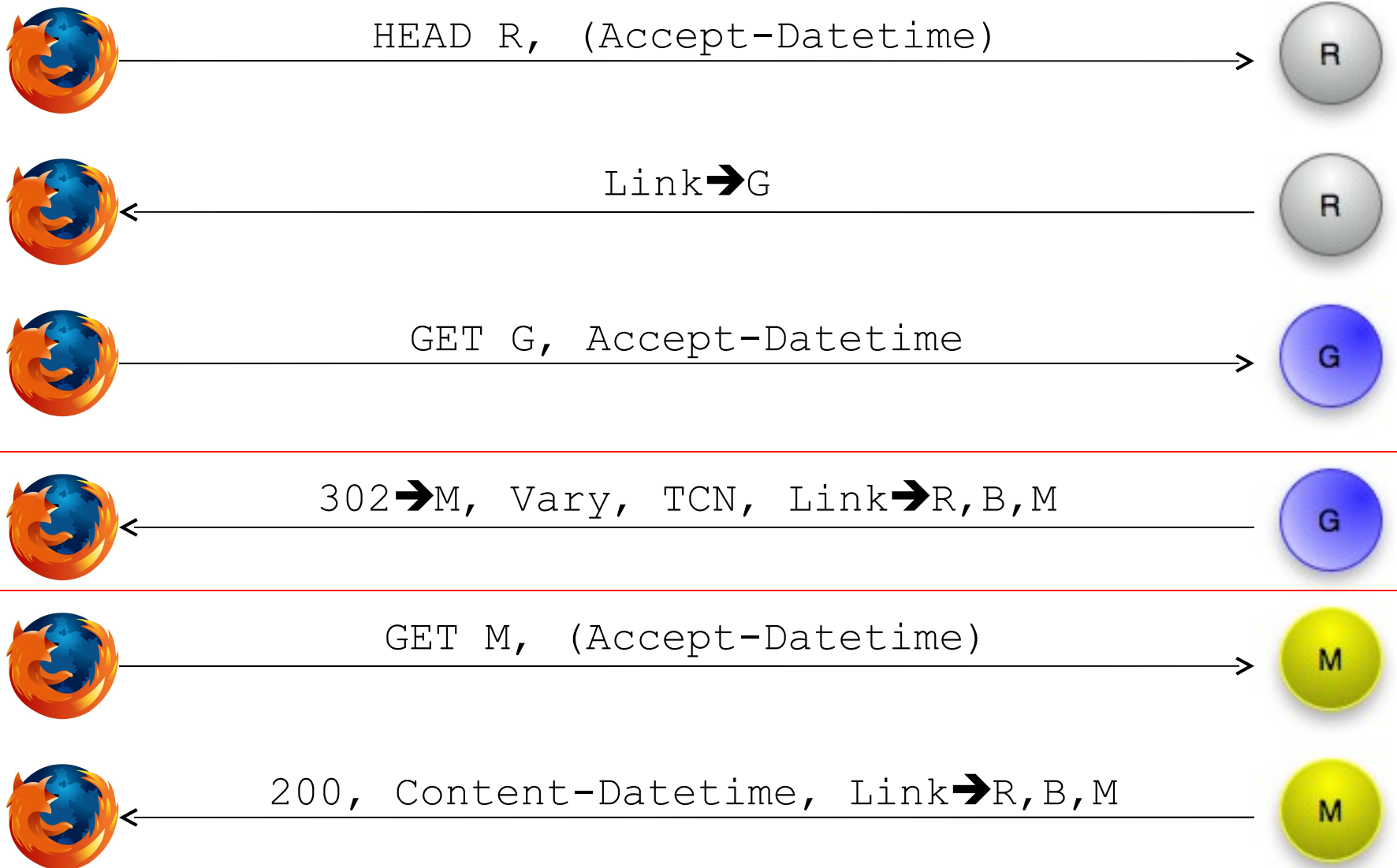


GET G, Accept-Datetime



```
GET /dbpedia/timeline/http://dbpedia.org/resource/France HTTP/1.1
Host: mementoarchive.lanl.gov
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Accept-Datetime: Wed, 08 Jul 2009 06:00:00 GMT
```

# Memento HTTP Flow



# Memento HTTP Flow: Success – URI-G

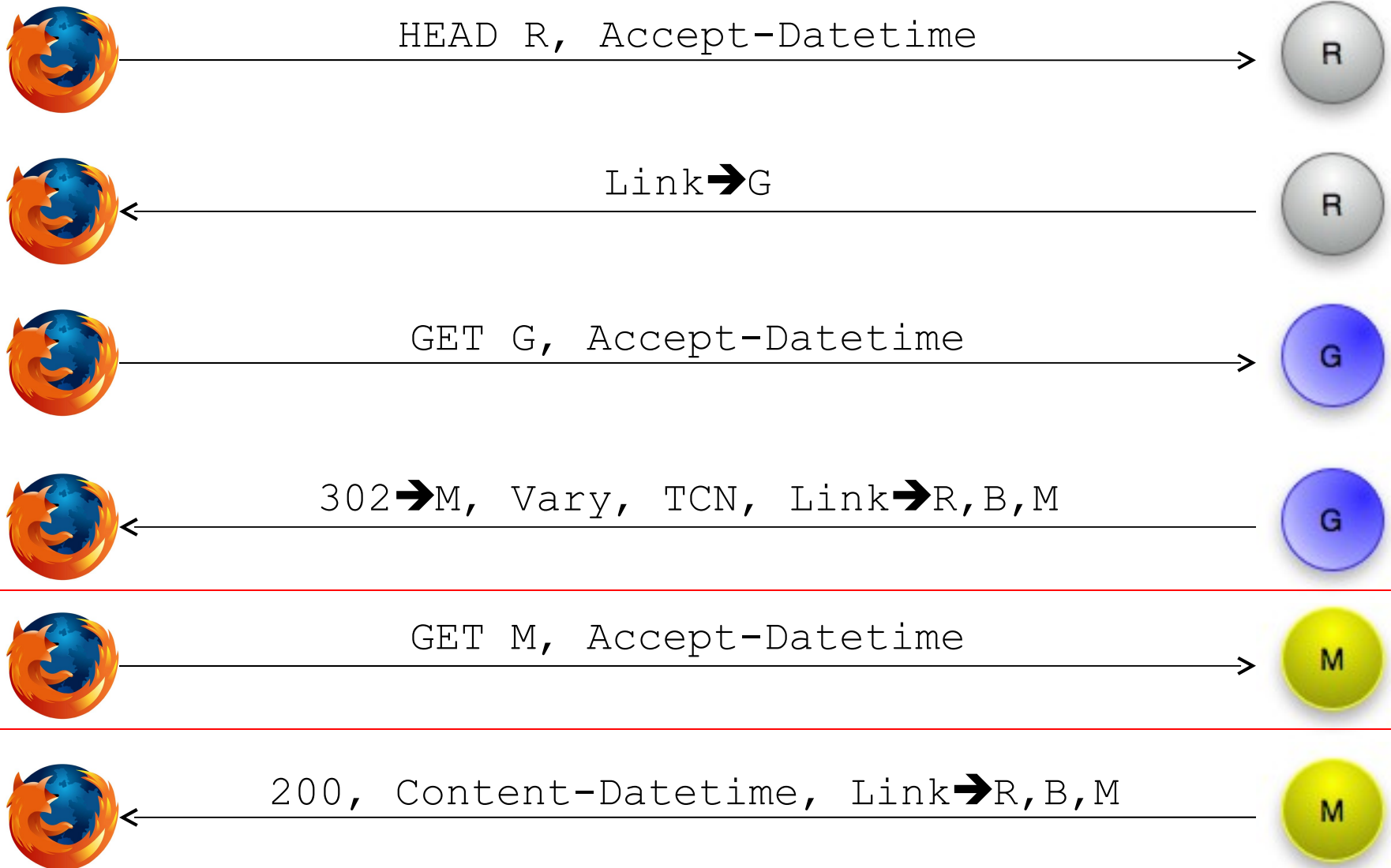


302 → M, Vary, TCN, Link → R, B, M



```
HTTP/1.1 302 Found
Date: Tue, 20 Apr 2010 17:51:00 GMT
Server: Apache
TCN: choice
Vary: negotiate, Accept-Datetime
Location:
http://mementoarchive.lanl.gov/dbpedia/memento/20090701/http://dbpedia.org/page/France.html
Link:
<http://dbpedia.org/resource/France>;rel="original",
<http://mementoarchive.lanl.gov/dbpedia/memento/20070901/http://dbpedia.org/page/France.html>;rel="first-memento";datetime="Sat, 01 Sep 2007 00:00:00 GMT",
<http://mementoarchive.lanl.gov/dbpedia/memento/20091101/http://dbpedia.org/page/France.html>;rel="last-memento next-memento";datetime="Sun, 01 Nov 2009 00:00:00 GMT",
<http://mementoarchive.lanl.gov/dbpedia/memento/20081101/http://dbpedia.org/page/France.html>;rel="prev-memento";datetime="Sat, 01 Nov 2008 00:00:00 GMT",
<http://mementoarchive.lanl.gov/dbpedia/memento/20090701/http://dbpedia.org/page/France.html>;rel="memento";datetime="Wed, 01 Jul 2009 00:00:00 GMT",
<http://mementoarchive.lanl.gov/dbpedia/timebundle/http://dbpedia.org/page/France>;rel="timebundle"
Transfer-Encoding: chunked
```

# Memento HTTP Flow



# Memento HTTP Flow: URI-M

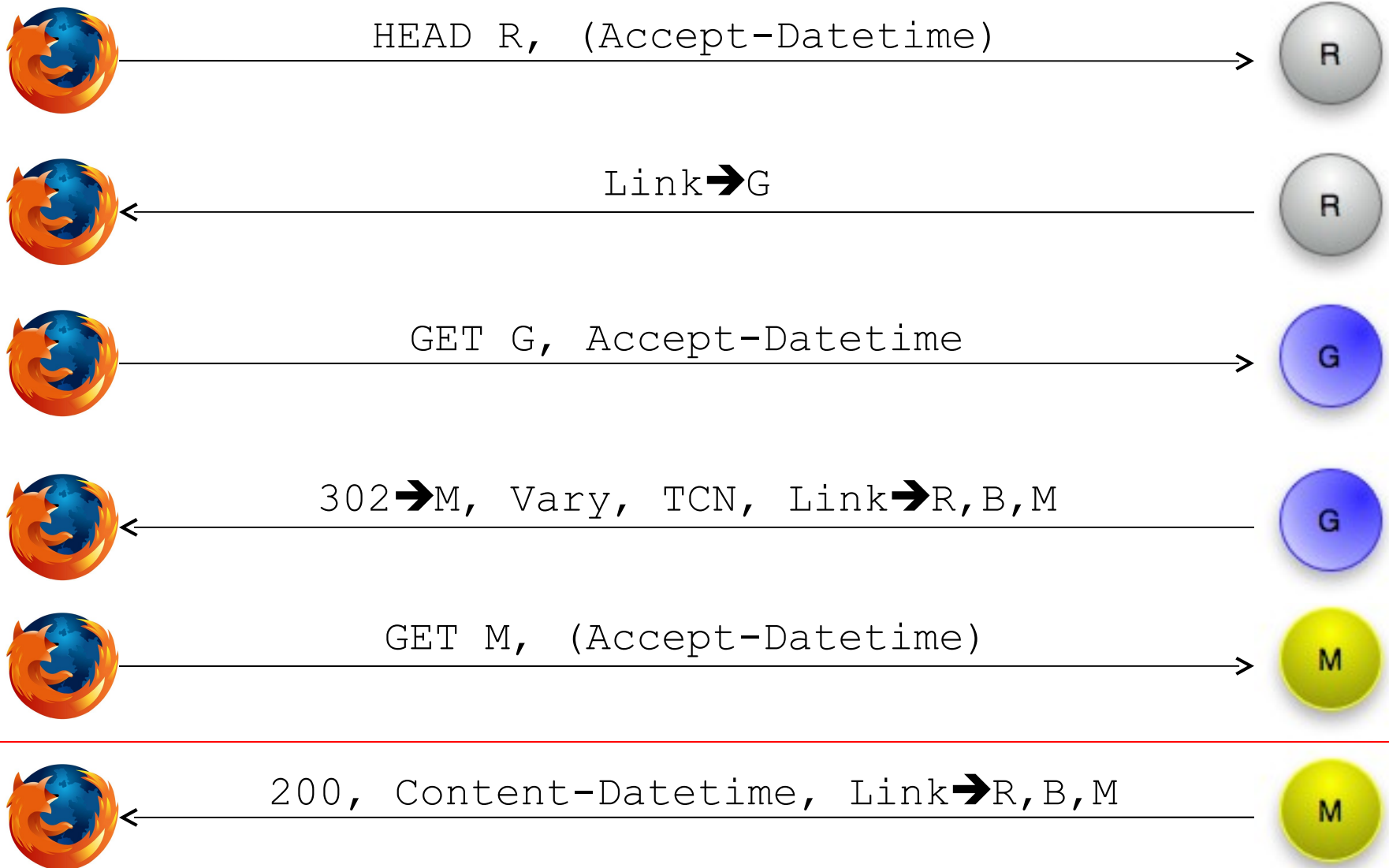


GET M, (Accept-Datetime)



```
GET /dbpedia/memento/20090701/http://dbpedia.org/page/France.html HTTP/1.1
Host: mementoarchive.lanl.gov
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Connection: close
```

# Memento HTTP Flow



# Memento HTTP Flow: Success – URI-M



200, Content-Datetime, Link → R, B, M



```
HTTP/1.1 200 OK
Date: Tue, 20 Apr 2010 17:51:00 GMT
Server: Apache
Content-Datetime: Wed, 01 Jul 2009 00:00:00 GMT
Link:
<http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/page/
  France.html>;rel="timegate",
<http://dbpedia.org/page/France.html>;rel="original",
<http://mementoarchive.lanl.gov/dbpedia/memento/20070901/http://dbpedia.org/page/
  France.html>;rel="first-memento";datetime="Sat, 01 Sep 2007 00:00:00 GMT",
<http://mementoarchive.lanl.gov/dbpedia/memento/20091101/http://dbpedia.org/page/
  France.html>;rel="last-memento next-memento";datetime="Sun, 01 Nov 2009 00:00:00
  GMT",
<http://mementoarchive.lanl.gov/dbpedia/memento/20081101/http://dbpedia.org/page/
  France.html>;rel="prev-memento";datetime="Sat, 01 Nov 2008 00:00:00
<http://mementoarchive.lanl.gov/dbpedia/timebundle/http://dbpedia.org/page/
  France.html>;rel="timebundle"
Connection: close
Transfer-Encoding: chunked
Content-Type: text/html; charset=UTF-8
```