Memento: Time Travel for the Web

Herbert Van de Sompel
Robert Sanderson
Michael L. Nelson

http://mementoweb.org/

Memento is funded by
The Library of Congress

Updated Technical Details (May 2011)

Memento wants to make it Easy

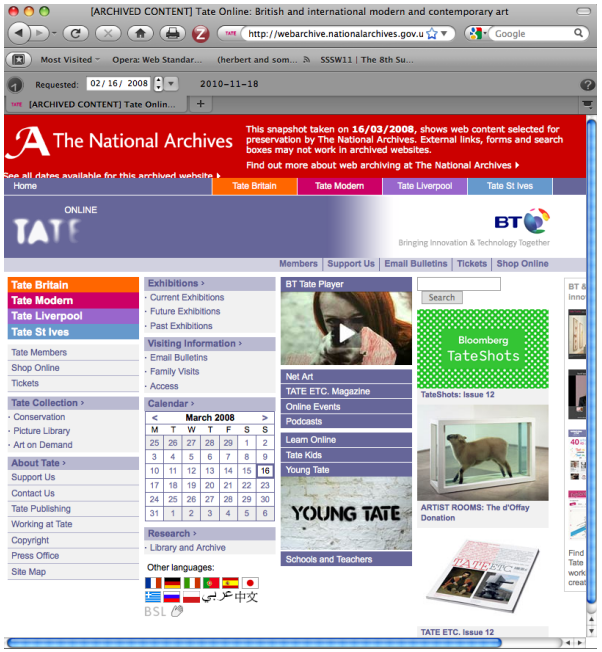
to navigate the Web of the Past

Technical Specification
https://datatracker.ietf.org/doc/draft-vandesompel-memento/

Los Alamos NATIONAL LABORATORY

Old Dominion University

Tate Online Today → Select Date March 16 2008 → Tate Online March 16 2008



From National Archives

Memento achieves this by introducing

a uniform version access capability to

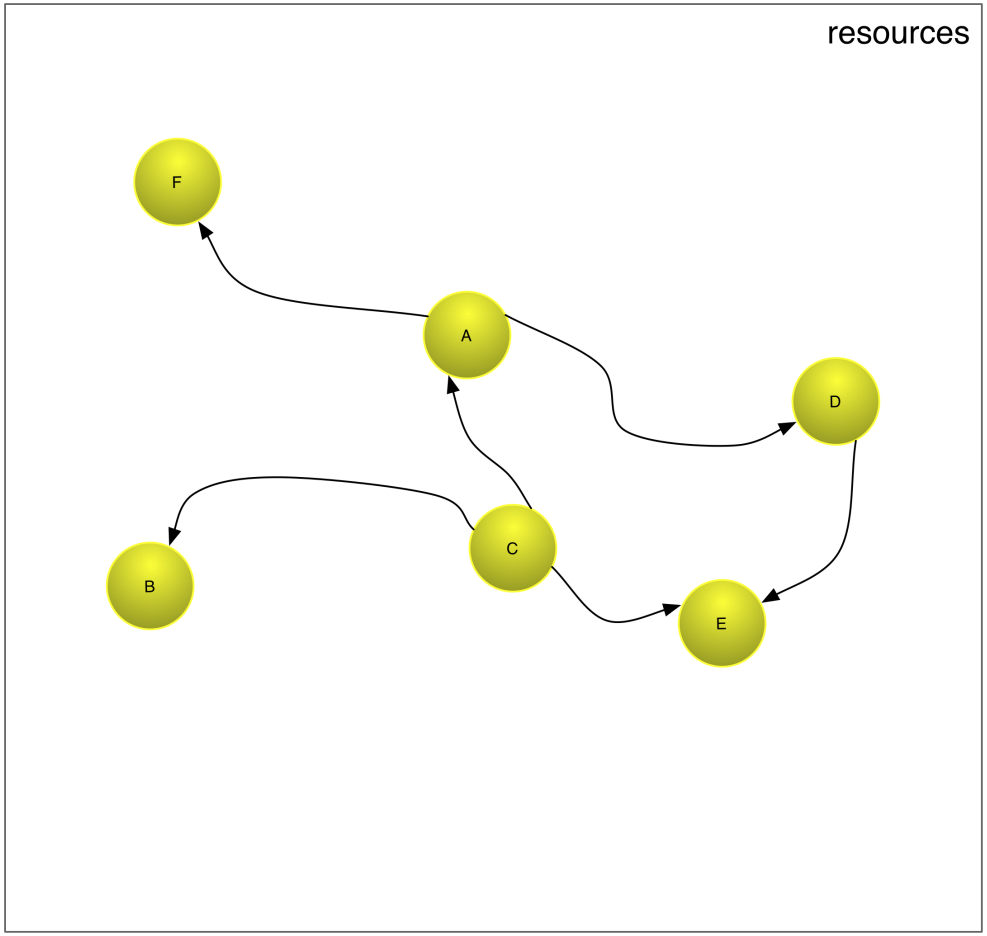
integrate the past and current Web

CC BY NC SA

Los Alamos NATIONAL LABORATORY
Old Dominion University

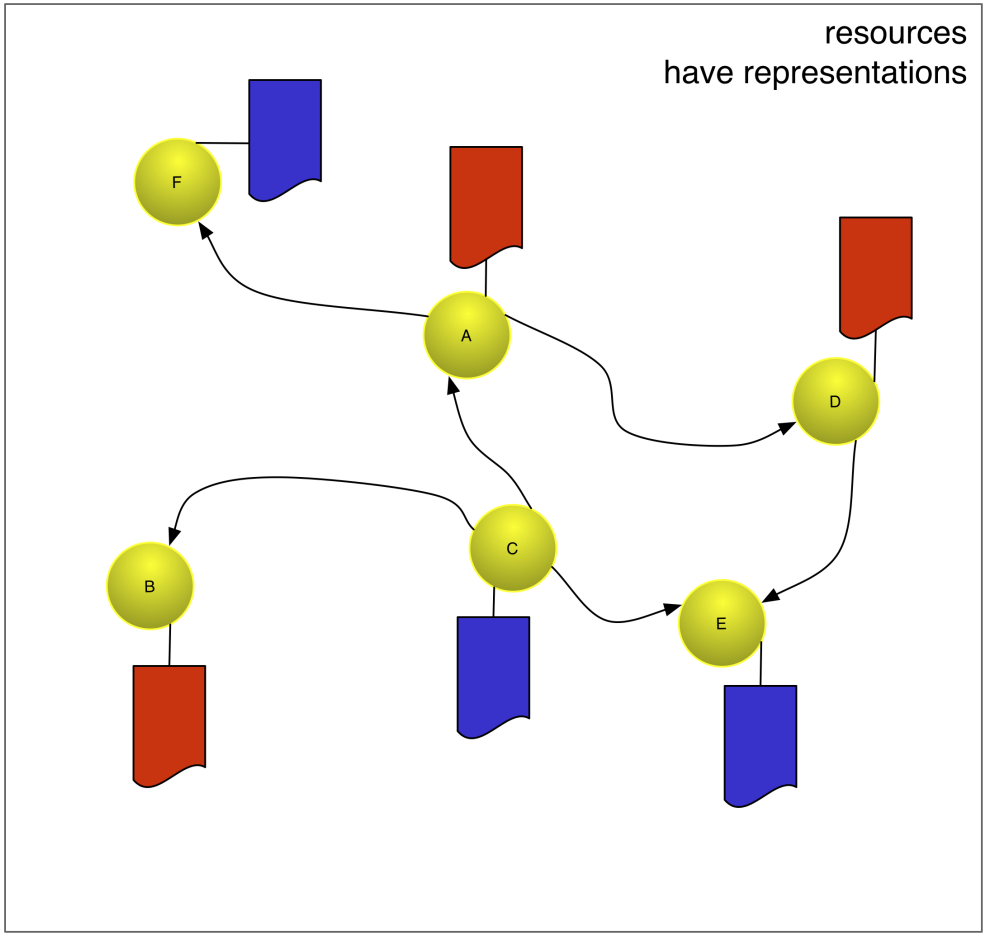
Problem Statement …

Resources

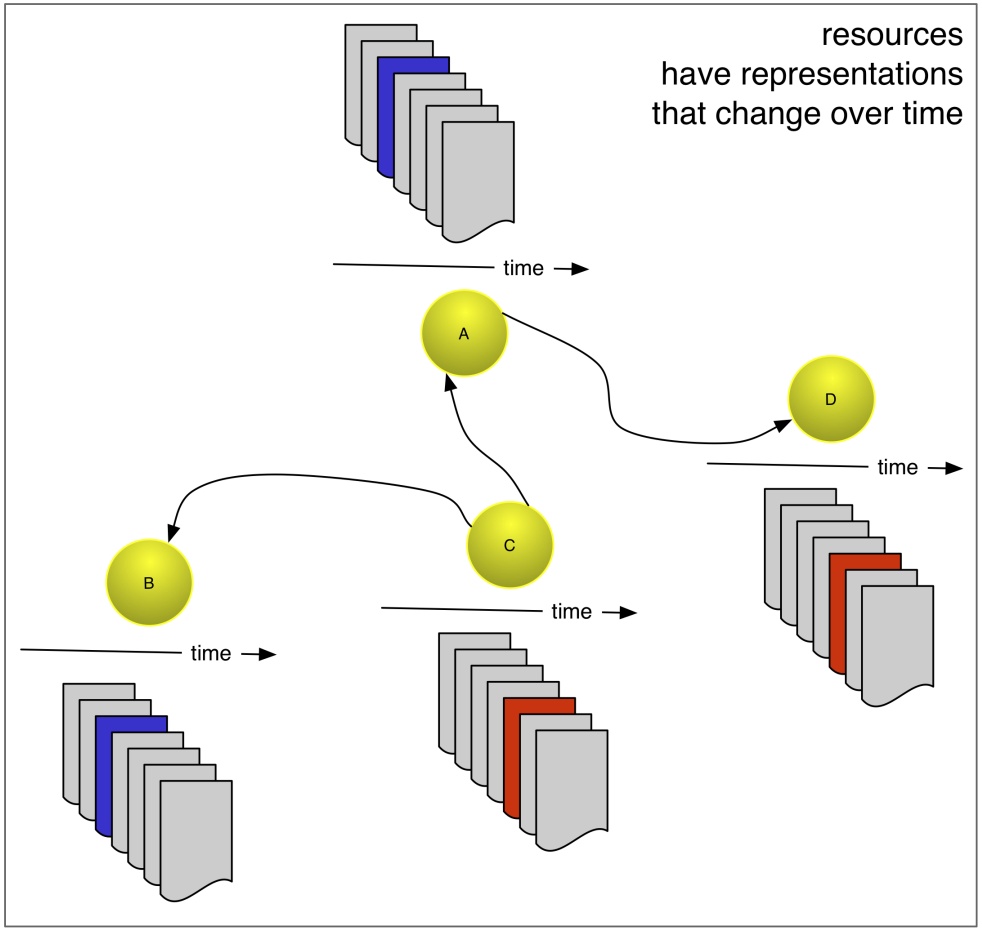


memento
cc BY NC SA

Resources have Representations

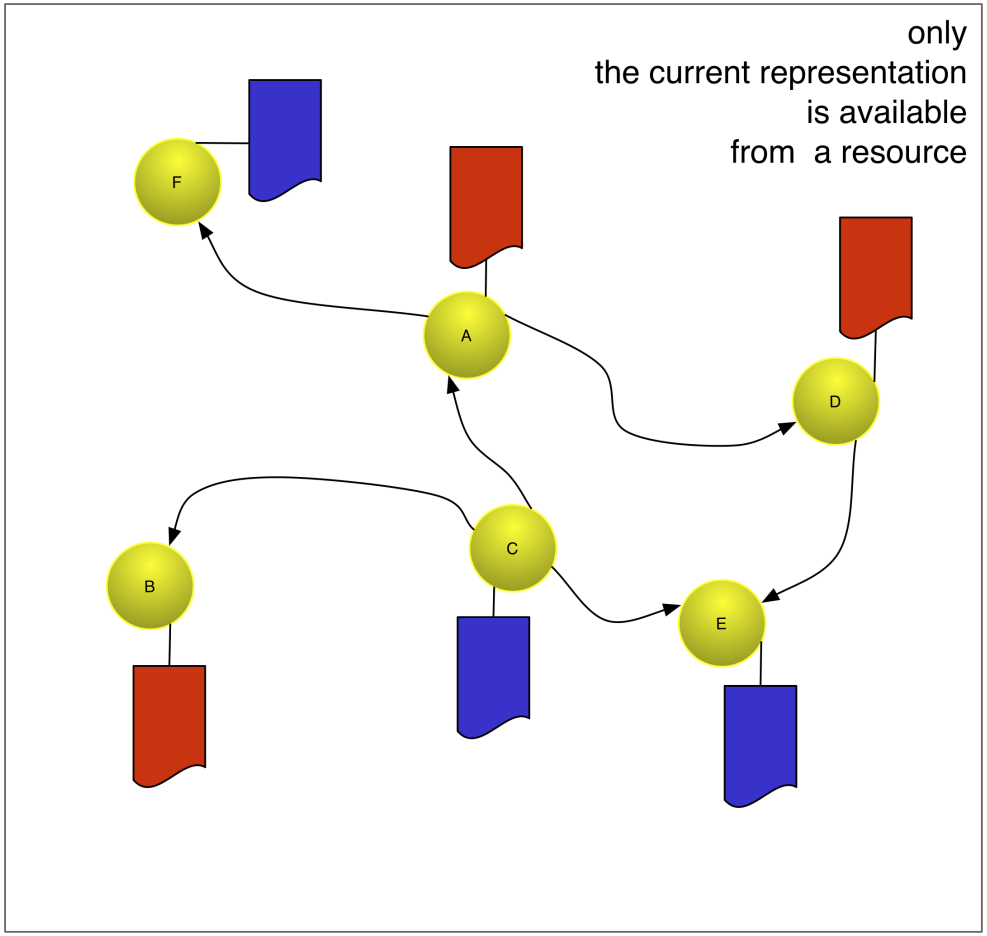


Resources have Representations that Change over Time

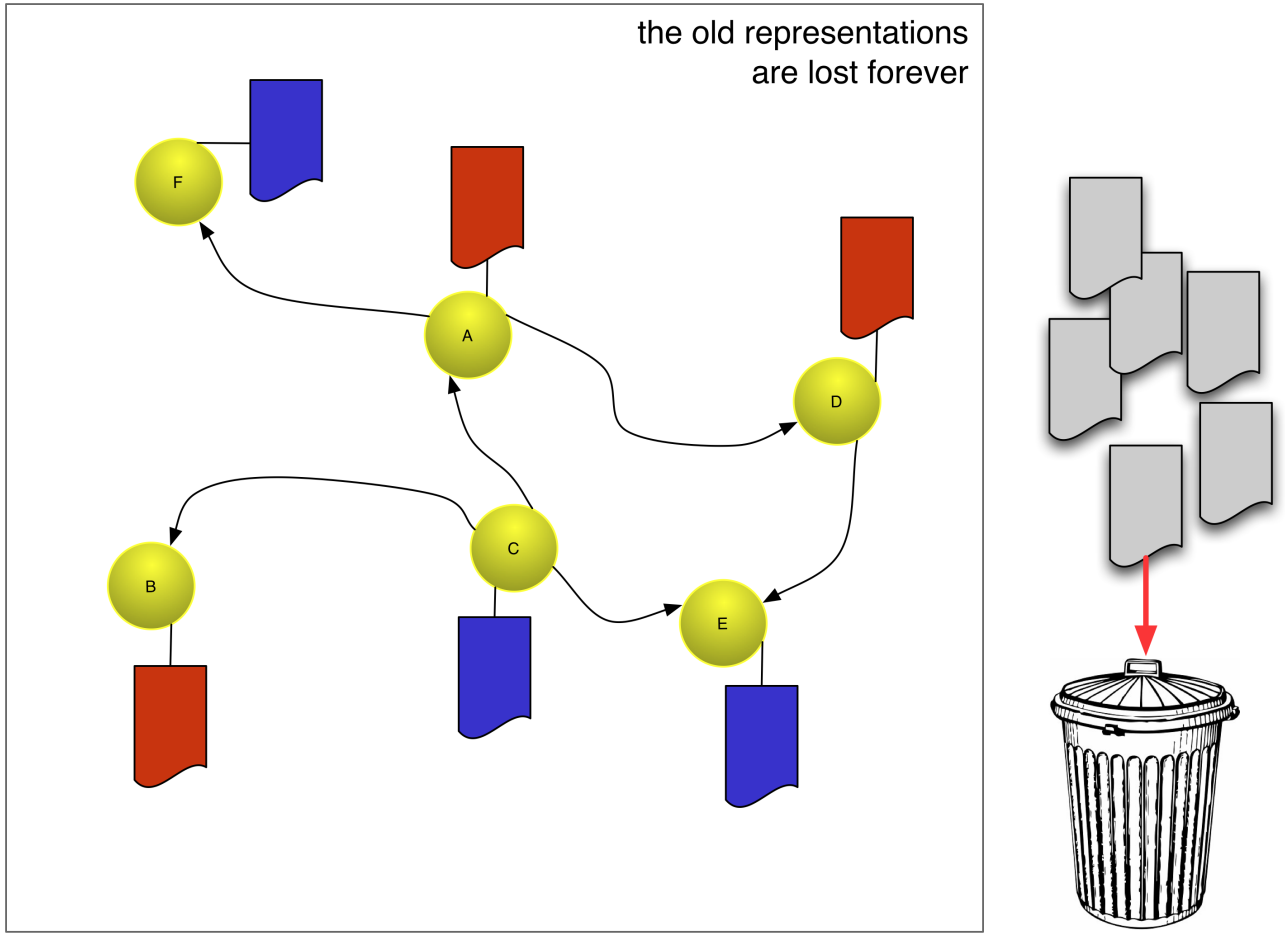


memento
CC BY NC SA

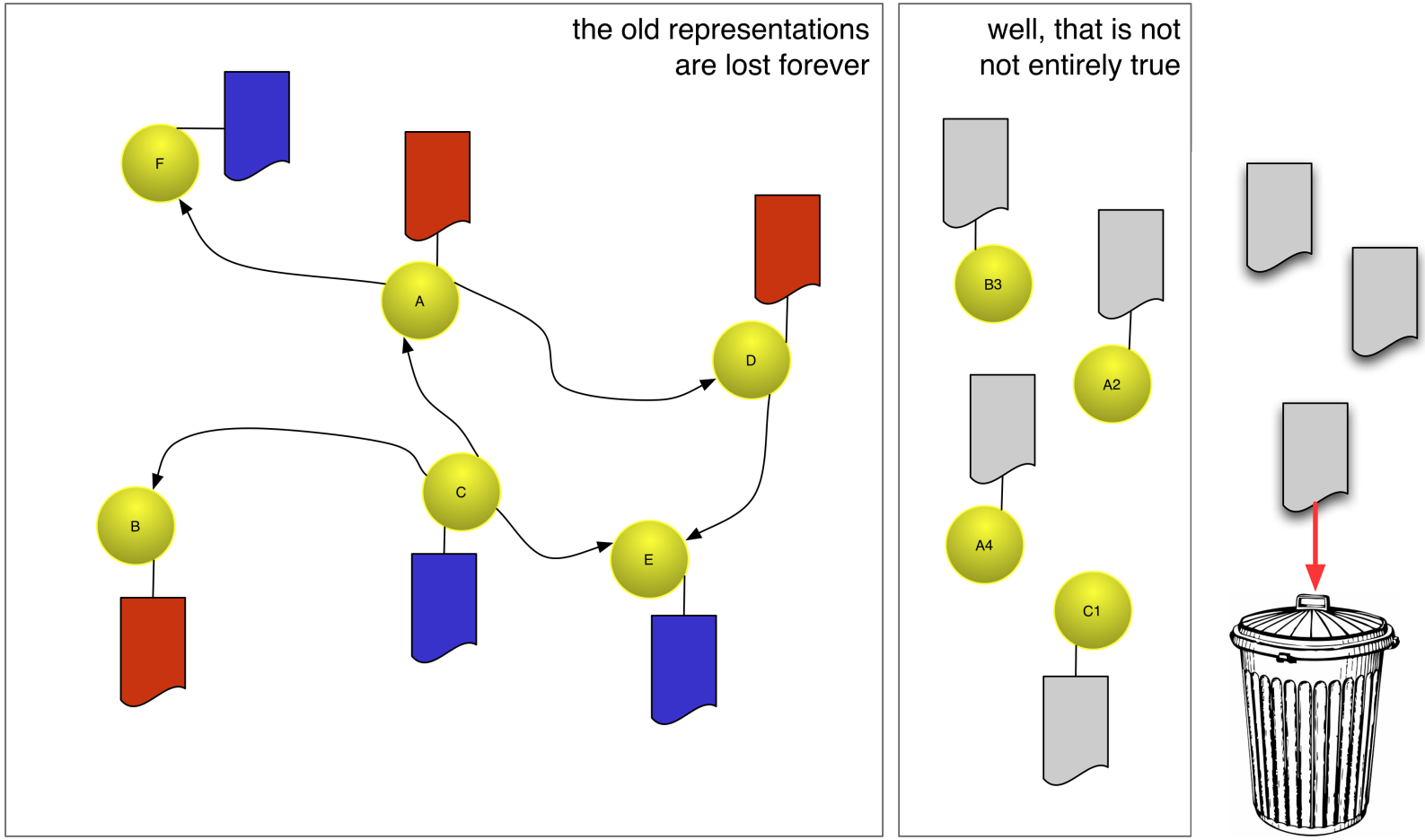
Only the Current Representation is Available from a Resource



Old Representations are Lost Forever



Archived/Version Resources Exist



Resource Versions on the Web

- Content Management Systems
- Web Archives
- Transactional archives
- Search engine caches
- …

memento
CC BY NC SA
Los Alamos NATIONAL LABORATORY
Old Dominion University

Sep 11 2001, 20:36:10 UTC

Dec 20 2001, 4:51:00 UTC

Archived Resources

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

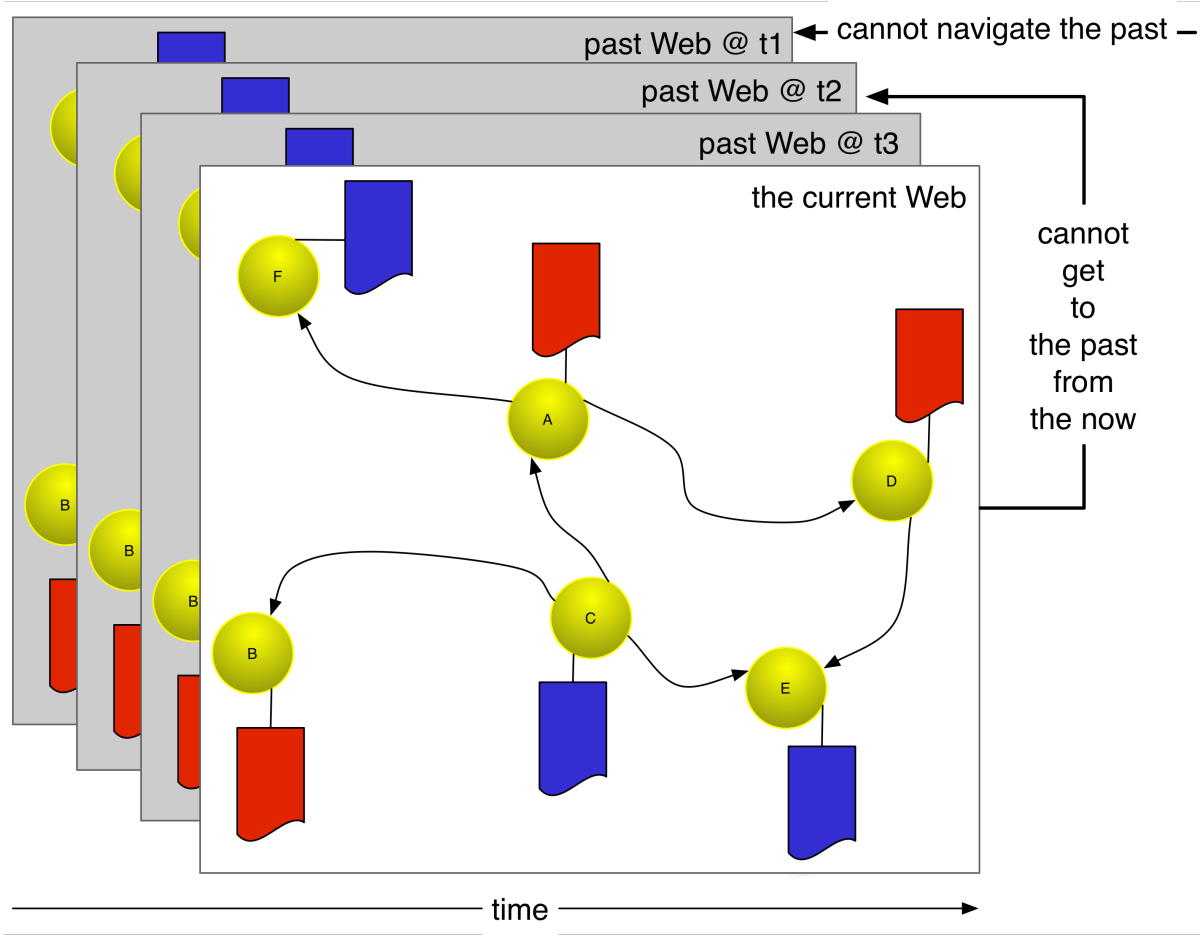
http://en.wikipedia.org/w/index.php?title=September_11_attacks&oldid=282333 archived resource for http://en.wikipedia.org/wiki/September_11_attacks

Versions are Not Integrated with the Web

- Cannot talk about a resource as it used to exist
- Cannot access a prior version knowing the current one
- Cannot access the current version knowing a prior one

BY NC SA

Memento Wants to Integrate the Past and Current Web



The Memento Framework:

Protocol to Integrate Past and Current Web

Overview

BY NC SA

Memento Framework

- Regards the Web as a big Content Management System
- Introduces a uniform capability to access versions on the Web
- Does not build new archives but leverages existing systems that host versions

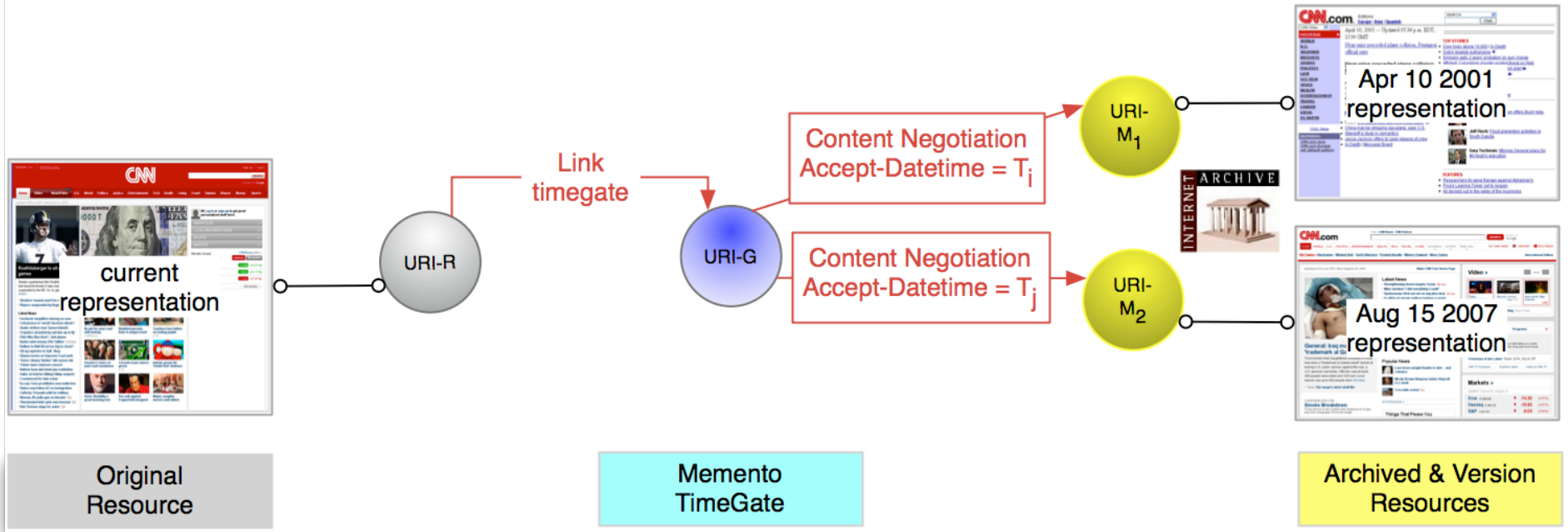
BY NC SA

Memento Framework

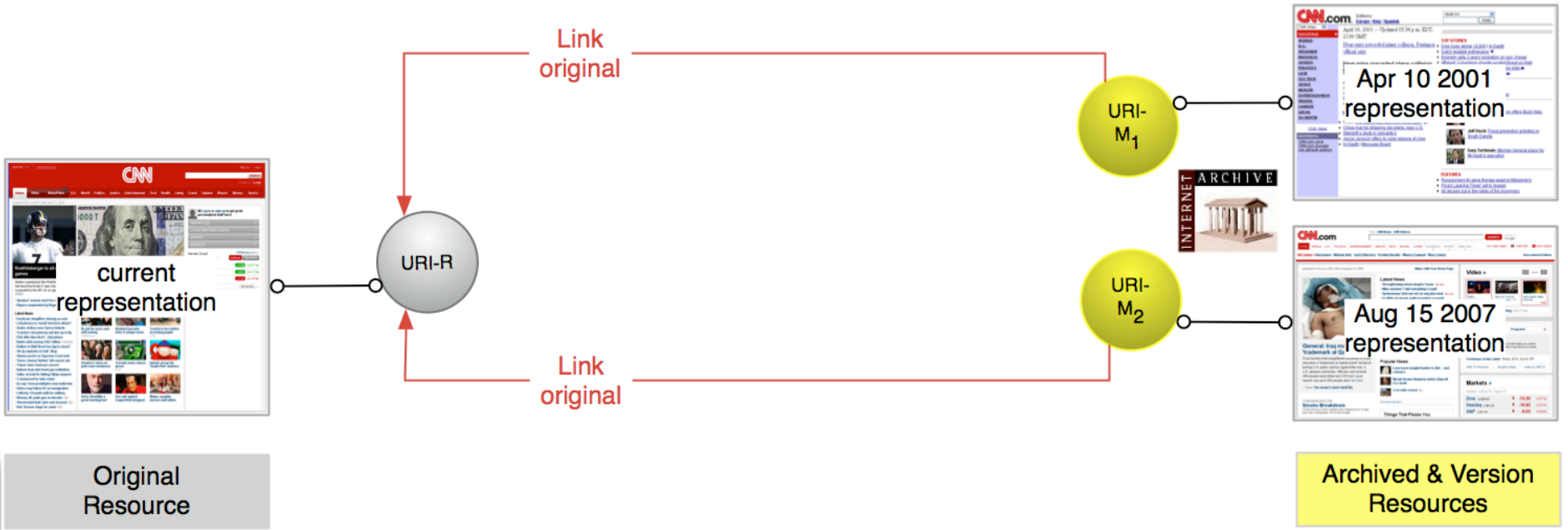
- Is distributed: versions may exist on several servers
- Uses time as a global version indicator
- Is based on the primitives of the Web: resource, resource state, representation, content negotiation, link

Original Resources and Mementos

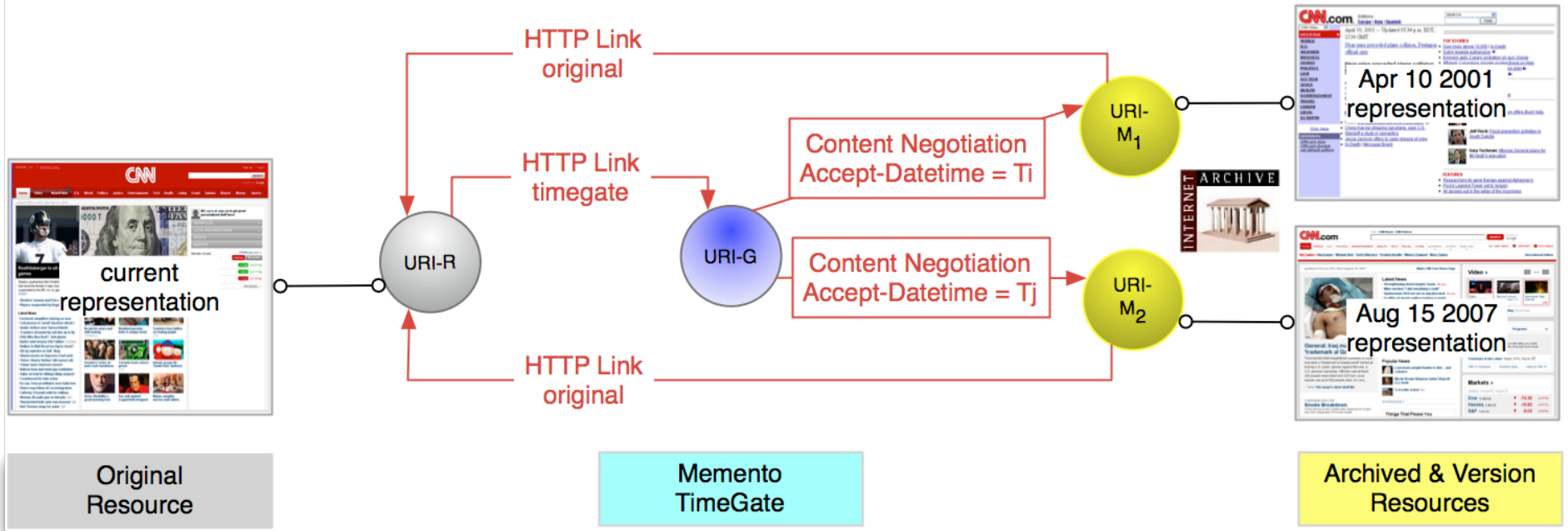
Bridge from Present to Past



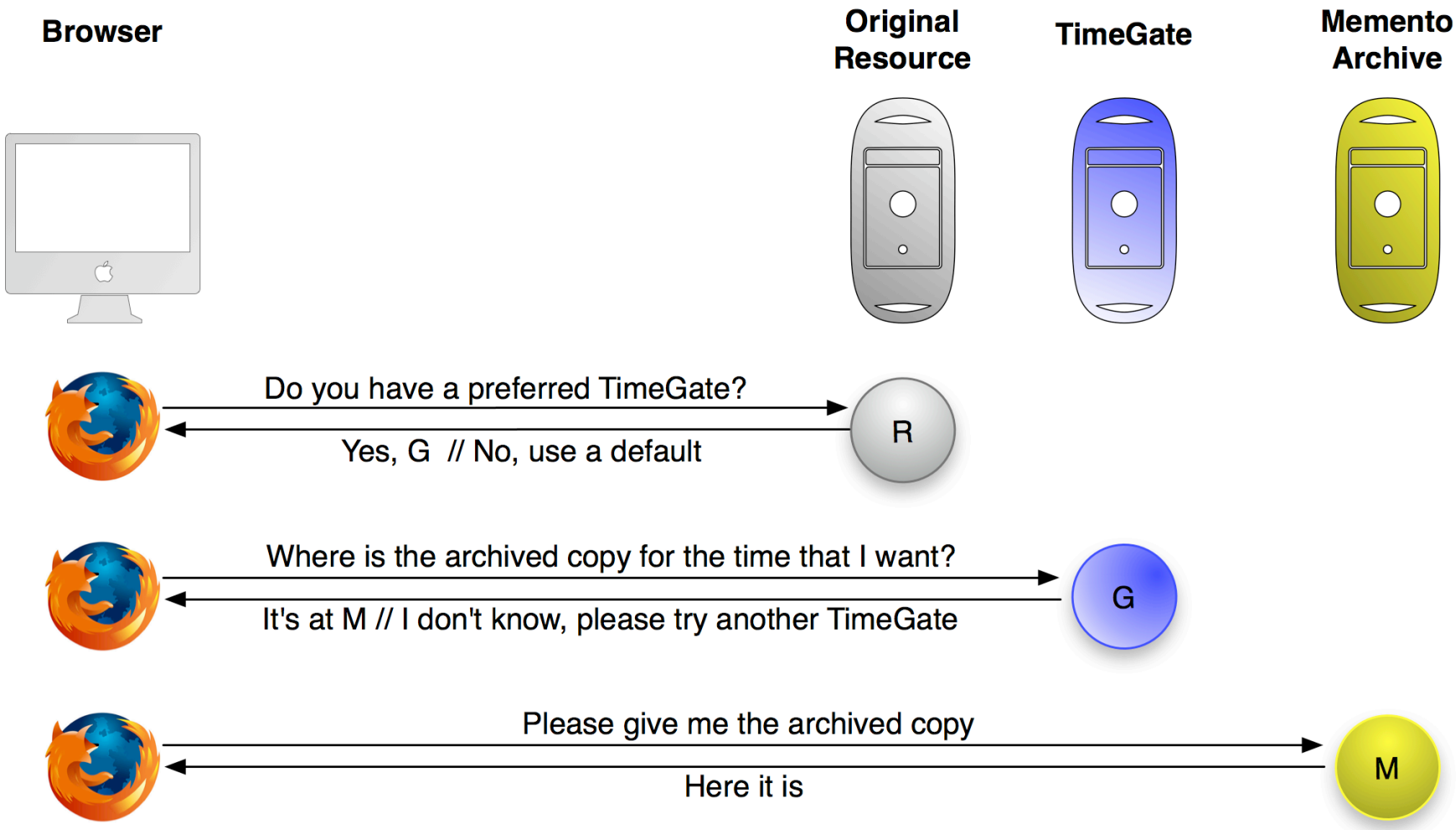
Bridge from Past to Present



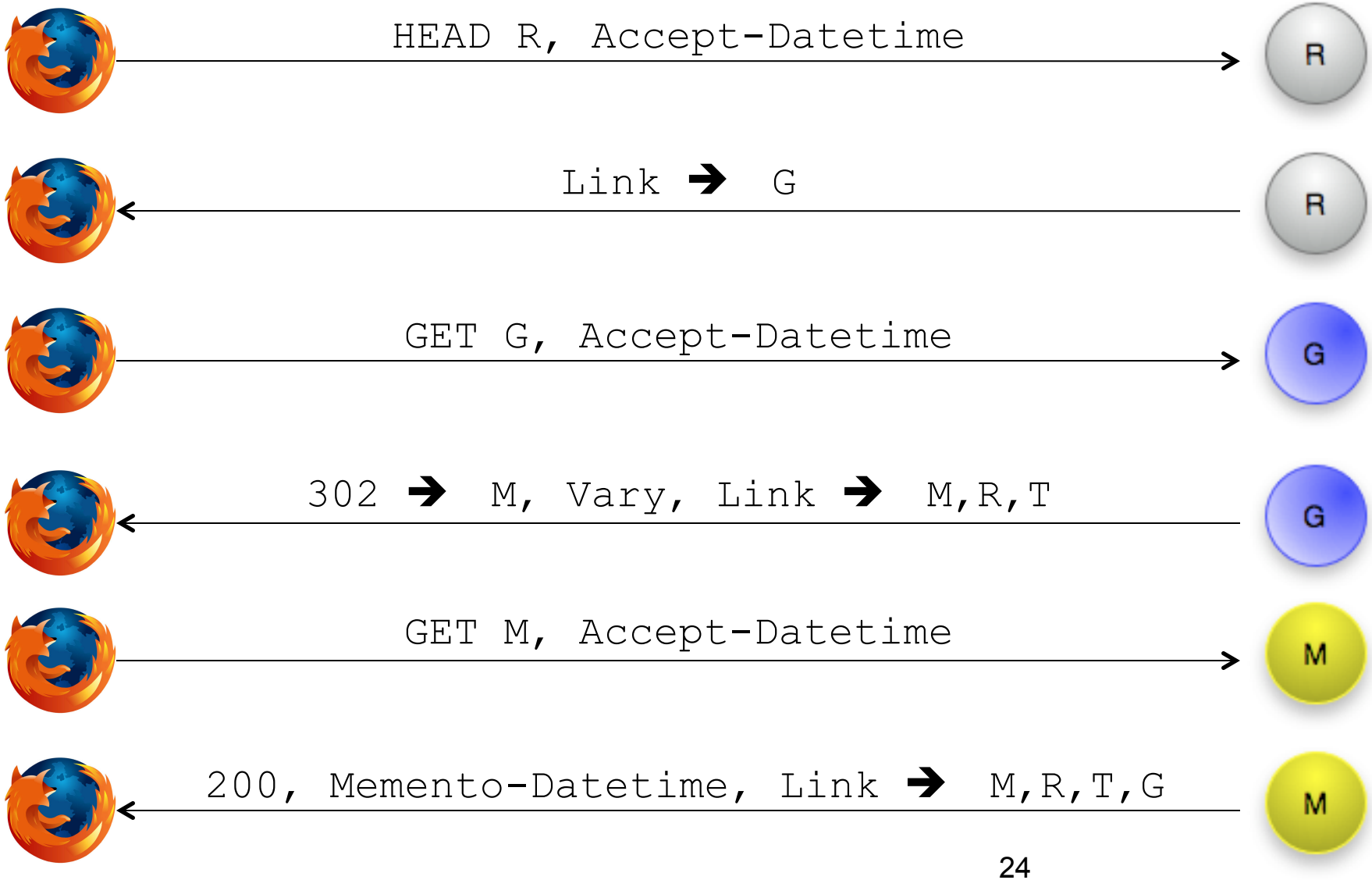
Memento Framework



Memento Client Server Interaction



Memento HTTP Flow

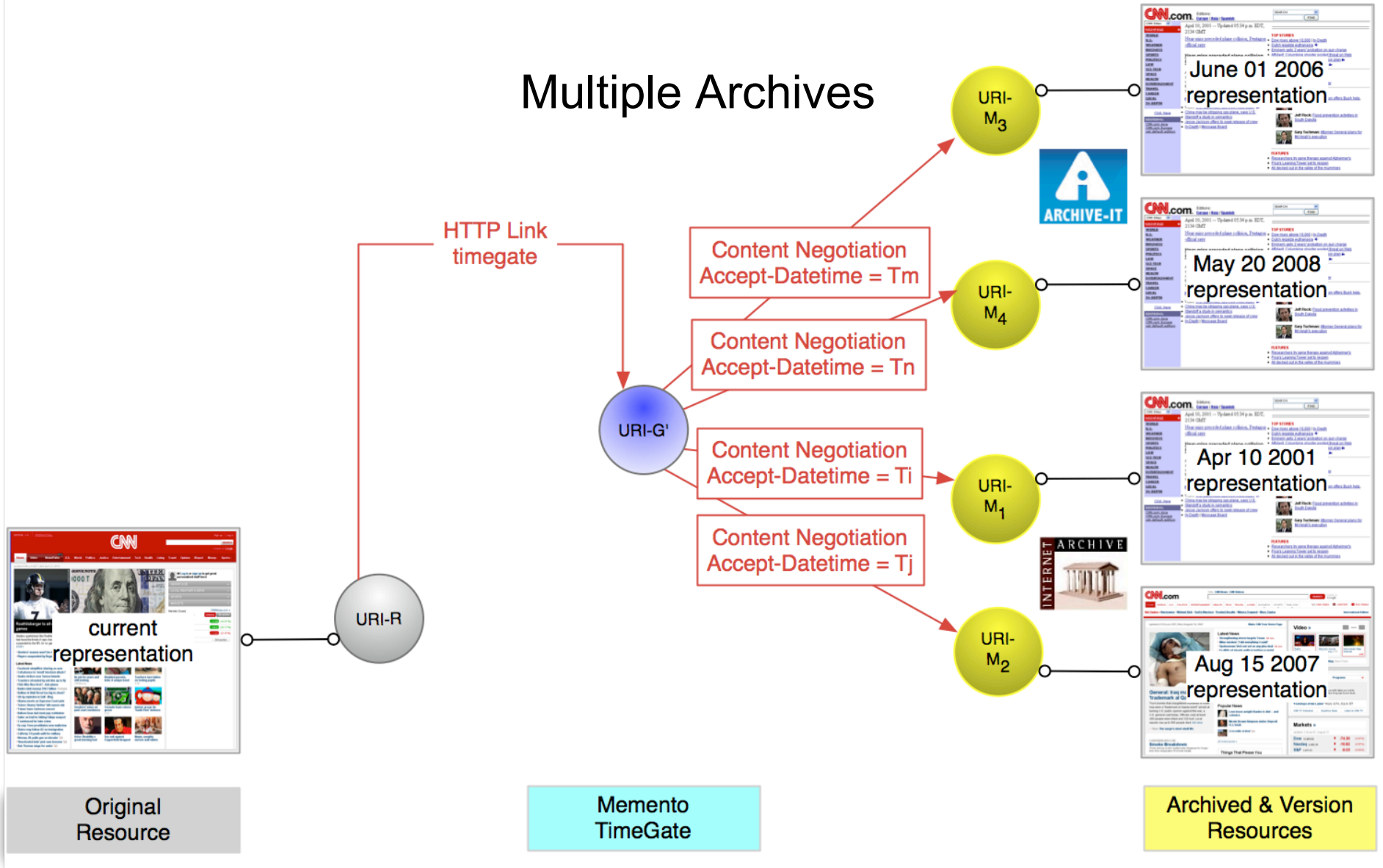


The Memento Framework:

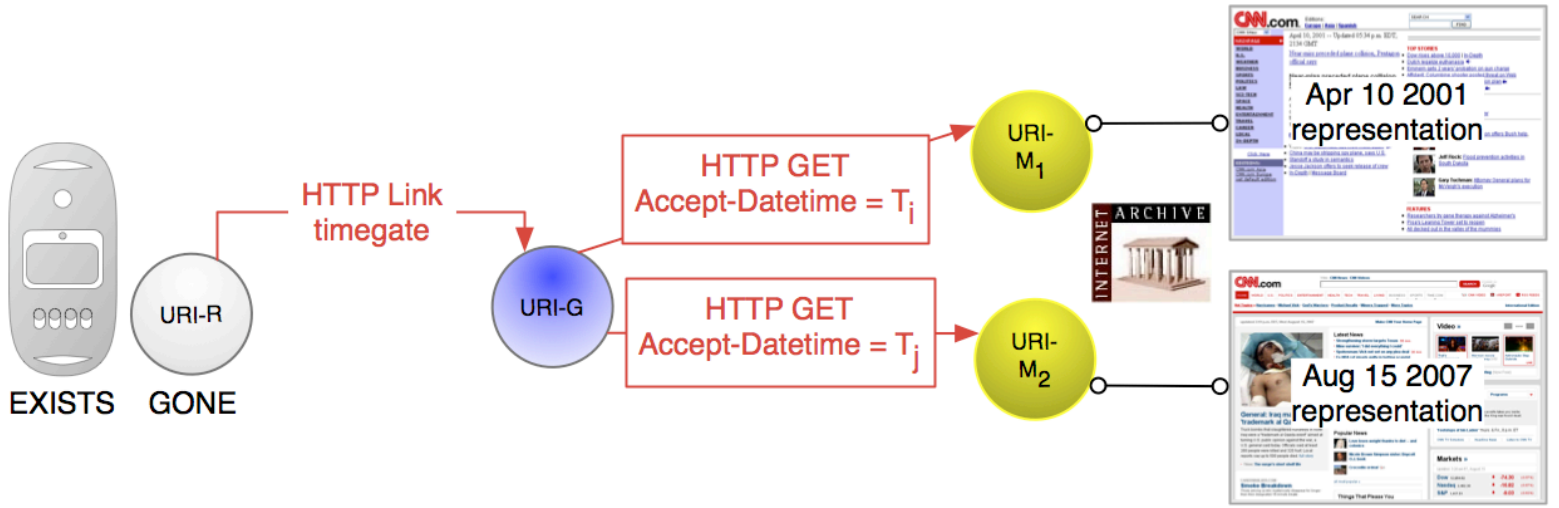
Protocol to Integrate Past and Current Web

Interesting Cases

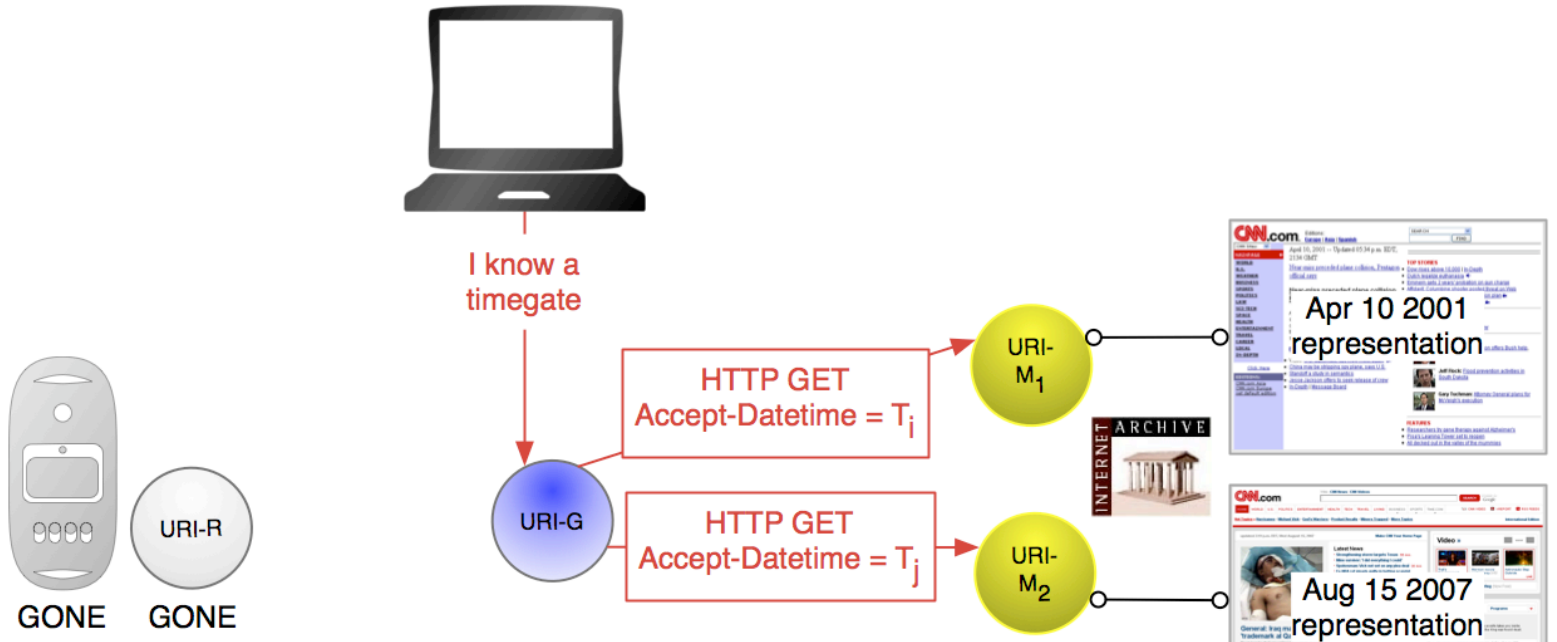
Multiple Archives



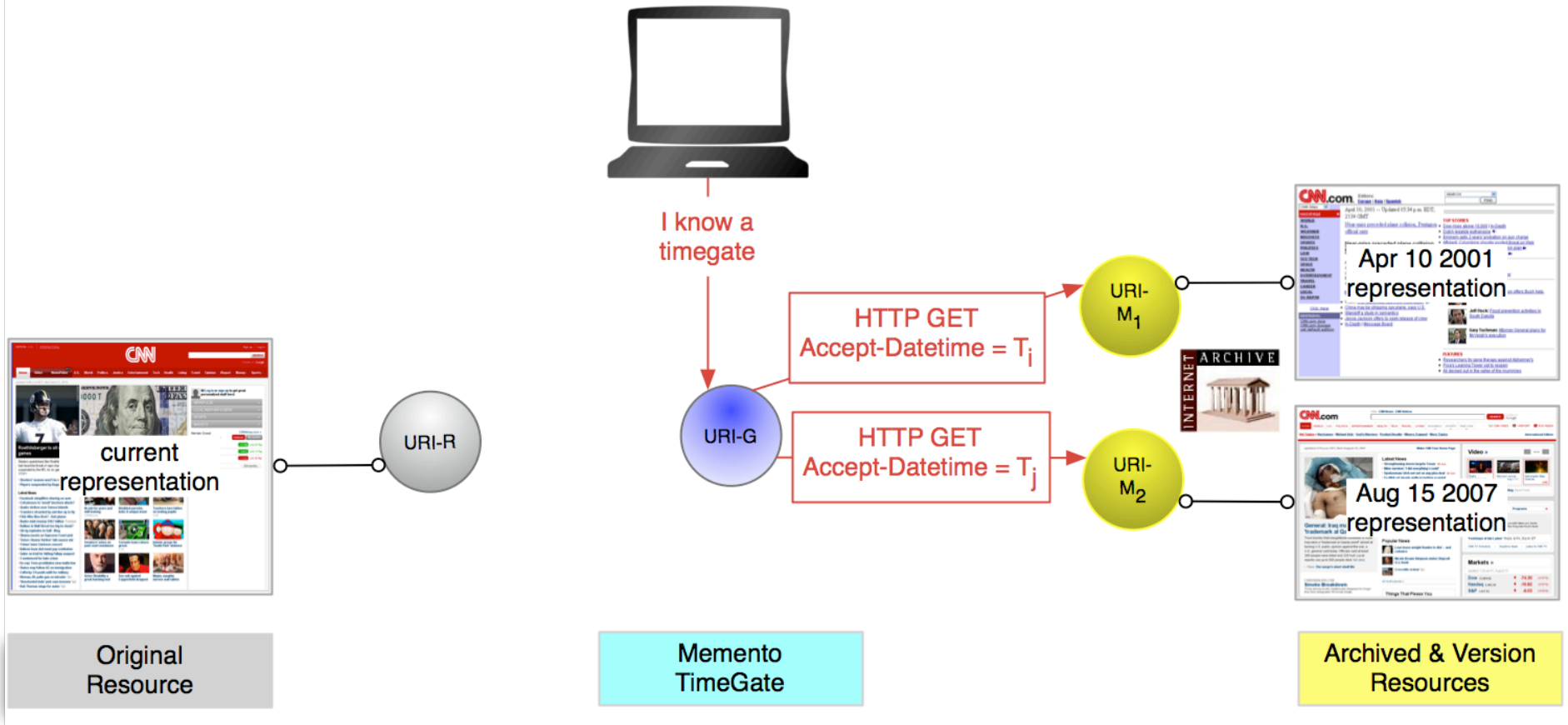
Original Resource Gone



Original Resource's Server Gone



Original Resource Provides no Link



The Memento Framework:

Protocol to Integrate Past and Current Web

HTTP Headers

Los Alamos NATIONAL LABORATORY

Old Dominion University

HTTP Headers used in Memento

- Defines two new headers:
  - request: `Accept-Datetime`
  - response: `Memento-Datetime`
- Introduce new content for two existing headers:
  - response: `Vary ; Link`
- Use one existing headers without modification:
  - response: `Location`

HTTP Request Headers: Accept-Datetime

- `Accept-Datetime`
  - Issued against TimeGate, (Original Resource), (Memento)
  - Header value:
    - Desired datetime of Memento (MANDATORY)
      Must be in RFC 1123 format and in GMT
    - Interval indicator to express the client is only interested in Mementos within the interval (OPTIONAL)
      – Expressed as two ISO8601 durations:
      `"-P3DT5H;+P2DT6H"`

```
Accept-Datetime: Mon, 12 Oct 2009 14:20:33 GMT
```

HTTP Response Headers: Memento-Datetime

- `Memento-Datetime`
  - Returned by Mementos
    - Always. Even when not via a TimeGate
  - Header value: Archival datetime of the Memento
    - Resource has not and will not change beyond that date
  - This header is sticky:
    - Once returned, must always return it with same value
    - Must also be preserved when Mementos are mirrored at different URIs
- This header is crucial to allow a client to understand it has arrived at a Memento

  See: *http://www.mementoweb.org/guide/resourcetype/*

```
Memento-Datetime: Mon, 12 Oct 2009 14:20:33 GMT
```

HTTP Response Headers: Vary

- `Vary`
  - Returned by TimeGate
  - Similar to regular content negotiation
  - Header value:
    - `negotiate, accept-datetime`

- TimeGate must first meet the datetime preference, and then – if possible – other content negotiation preferences

- Note: `accept-datetime` value in `Vary` header is crucial to allow a client to understand it has arrived at a TimeGate.
  See: *http://www.mementoweb.org/guide/resourcetype/*

```
Vary: negotiate, accept-datetime
```

HTTP Response Headers: Location

- `Location`
  - Returned by TimeGate
  - Similar to regular content negotiation
  - Header value: URI of selected Memento

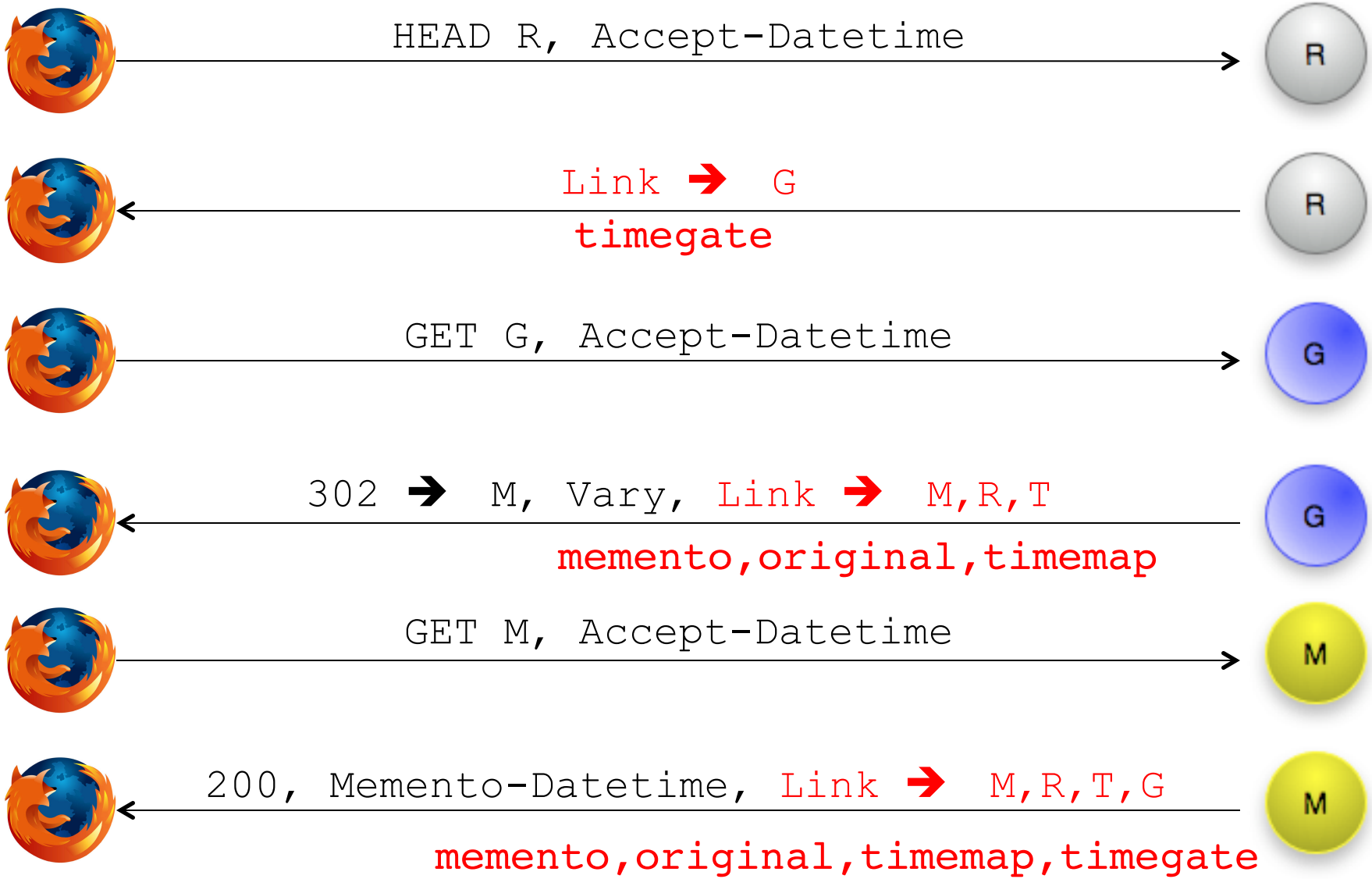
```
Location: http://web.archive.org/web/20010911223004/
                        http://cnn.com
```

HTTP Response Headers: Link

- `Link`
  - Returned by Original Resource, TimeGate and Mementos
  - Various new Relation Types are introduced:
    - `"original"`
    - `"timegate"`
    - `"memento"`
    - `"timemap"`
  - HTTP Link Header: RFC 5988
    See: *https://datatracker.ietf.org/doc/rfc5988/*

```
Link: <http://web.archive.org/web/20010911223004/http://
 cnn.com>;rel="memento";datetime="Mon, 11 Sep 2001 22:30:04 GMT"
```

Memento HTTP Flow



HTTP Response Headers: Link

- `Link`
  - Returned by Original Resource, TimeGate and Mementos
  - Various new Relation Types are introduced

Relation Type	Original Resource	TimeGate	Memento
original	NA, except see Section 3.1.2.1	REQUIRED, 1	REQUIRED, 1
timegate	RECOMMENDED, 0 or more	NA	RECOMMENDED, 0 or more
timemap	NA	RECOMMENDED, 0 or more	RECOMMENDED, 0 or more
memento	NA, except see Section 3.1.2.1	REQUIRED, 1 or more	REQUIRED, 1 or more

The Memento Framework:

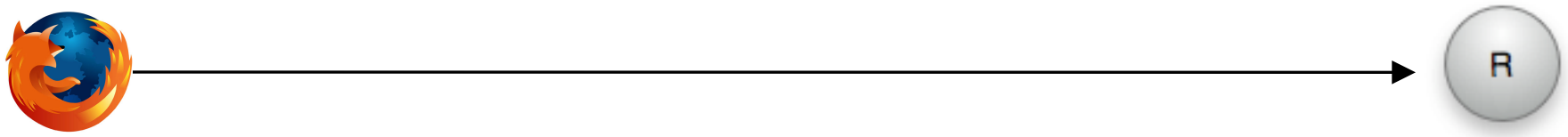
Protocol to Integrate Past and Current Web

HTTP Interactions

Los Alamos NATIONAL LABORATORY

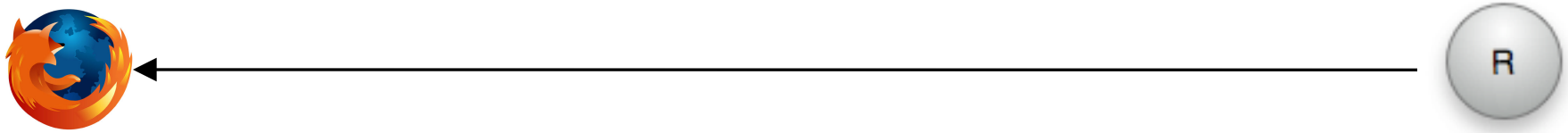
Old Dominion University

Memento HTTP Flow: Step 1



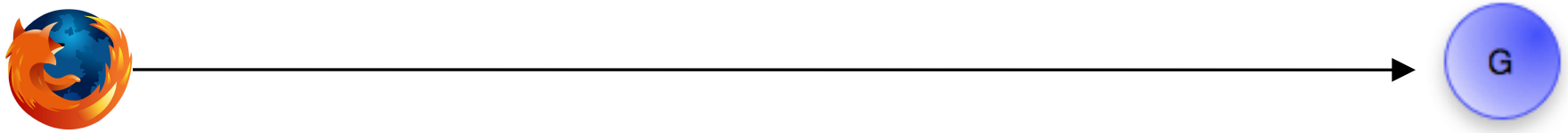

```
HEAD / HTTP/1.1
Host: a.example.org
Accept-Datetime: Tue, 11 Sep 2001 20:35:00 GMT
Connection: close
```

Memento HTTP Flow: Step 2




```
HTTP/1.1 200 OK
Date: Thu, 21 Jan 2010 00:02:12 GMT
Server: Apache
Link: <http://arxiv.example.net/timegate/http://a.example.org>
   ; rel="timegate"
Content-Length: 255
Connection: close
Content-Type: text/html; charset=iso-8859-1
```

Memento HTTP Flow: Step 3

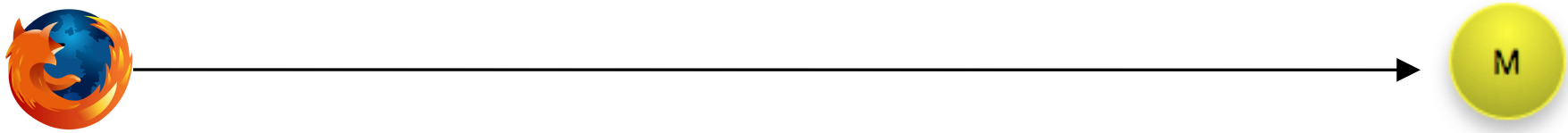



```
GET /timegate/http://a.example.org
 HTTP/1.1
Host: arxiv.example.net
Accept-Datetime: Tue, 11 Sep 2001 20:35:00 GMT
Connection: close
```

Memento HTTP Flow: Step 4

```
HTTP/1.1 302 Found
Date: Thu, 21 Jan 2010 00:06:50 GMT
Server: Apache
Vary: negotiate, accept-datetime
Location:
 http://arxiv.example.net/web/20010911203610/http://a.example.org
Link: <http://a.example.org>; rel="original",
 <http://arxiv.example.net/web/20000915112826/http://a.example.org>
   ; rel="first memento"; datetime="Tue, 15 Sep 2000 11:28:26 GMT",
 <http://arxiv.example.net/web/20080708093433/http://a.example.org>
   ; rel="last memento"; datetime="Tue, 08 Jul 2008 09:34:33 GMT",
 <http://arxiv.example.net/timemap/http://a.example.org>
   ; rel="timemap"; type="application/link-format",
 <http://arxiv.example.net/web/20010911203610/http://a.example.org>
   ; rel="memento"; datetime="Tue, 11 Sep 2001 20:36:10 GMT",
 <http://arxiv.example.net/web/20010911203610/http://a.example.org>
   ; rel="prev memento"; datetime="Tue, 11 Sep 2001 20:30:51 GMT",
 <http://arxiv.example.net/web/20010911203610/http://a.example.org>
   ; rel="next memento"; datetime="Tue, 11 Sep 2001 20:47:33 GMT"
Content-Length: 0
Content-Type: text/plain; charset=UTF-8
Connection: close
```

Memento HTTP Flow: Step 5




```
GET /web/20010911203610/http://a.example.org
 HTTP/1.1
Host: arxiv.example.net
Accept-Datetime: Tue, 11 Sep 2001 20:35:00 GMT
Connection: close
```

Memento HTTP Flow: Step 6

```
HTTP/1.1 200 OK
Date: Thu, 21 Jan 2010 00:09:40 GMT
Server: Apache-Coyote/1.1
Memento-Datetime: Tue, 11 Sep 2001 20:36:10 GMT
Link: <http://a.example.org>; rel="original",
 <http://arxiv.example.net/web/20000915112826/http://a.example.org>
   ; rel="first memento"; datetime="Tue, 15 Sep 2000 11:28:26 GMT",
 <http://arxiv.example.net/web/20080708093433/http://a.example.org>
   ; rel="last memento"; datetime="Tue, 08 Jul 2008 09:34:33 GMT",
 <http://arxiv.example.net/timemap/http://a.example.org>
   ; rel="timemap"; type="application/link-format",
 <http://arxiv.example.net/timegate/http://a.example.org>
   ; rel="timegate",
 <http://arxiv.example.net/web/20010911203610/http://a.example.org>
   ; rel="memento"; datetime="Tue, 11 Sep 2001 20:36:10 GMT",
 <http://arxiv.example.net/web/20010911203610/http://a.example.org>
   ; rel="prev memento"; datetime="Tue, 11 Sep 2001 20:30:51 GMT",
 <http://arxiv.example.net/web/20010911203610/http://a.example.org>
   ; rel="next memento"; datetime="Tue, 11 Sep 2001 20:47:33 GMT"
Content-Length: 23364
Content-Type: text/html;charset=utf-8
Connection: close
```

The Memento Framework:

Protocol to Integrate Past and Current Web

HTTP Link Header Details

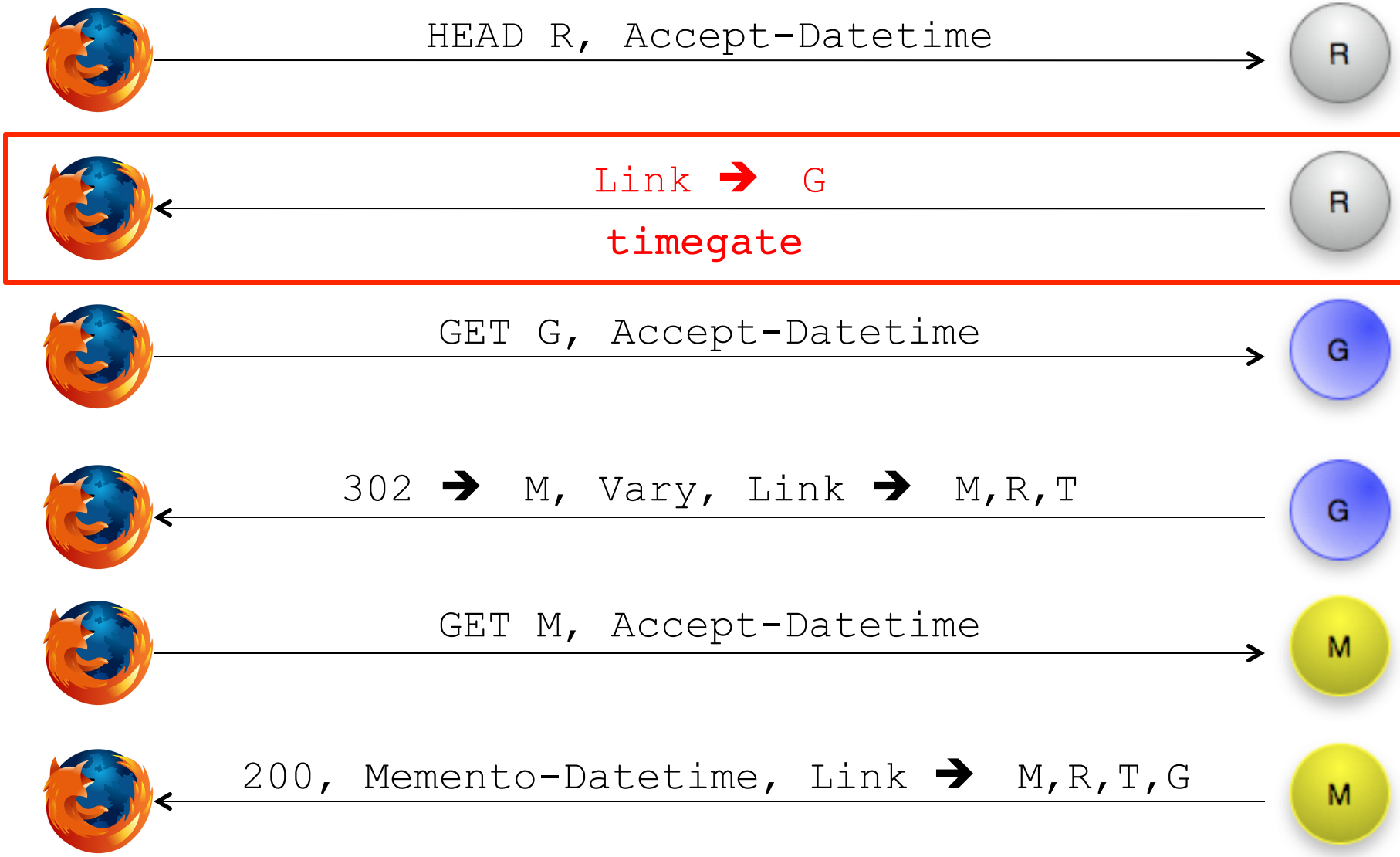
Los Alamos NATIONAL LABORATORY

Old Dominion University

HTTP Response Headers: Link

Relation Type	Original Resource	TimeGate	Memento
original	NA, except see Section 3.1.2.1	REQUIRED, 1	REQUIRED, 1
timegate	RECOMMENDED, 0 or more	NA	RECOMMENDED, 0 or more
timemap	NA	RECOMMENDED, 0 or more	RECOMMENDED, 0 or more
memento	NA, except see Section 3.1.2.1	REQUIRED, 1 or more	REQUIRED, 1 or more

Memento HTTP Flow



HTTP Response Headers: Link

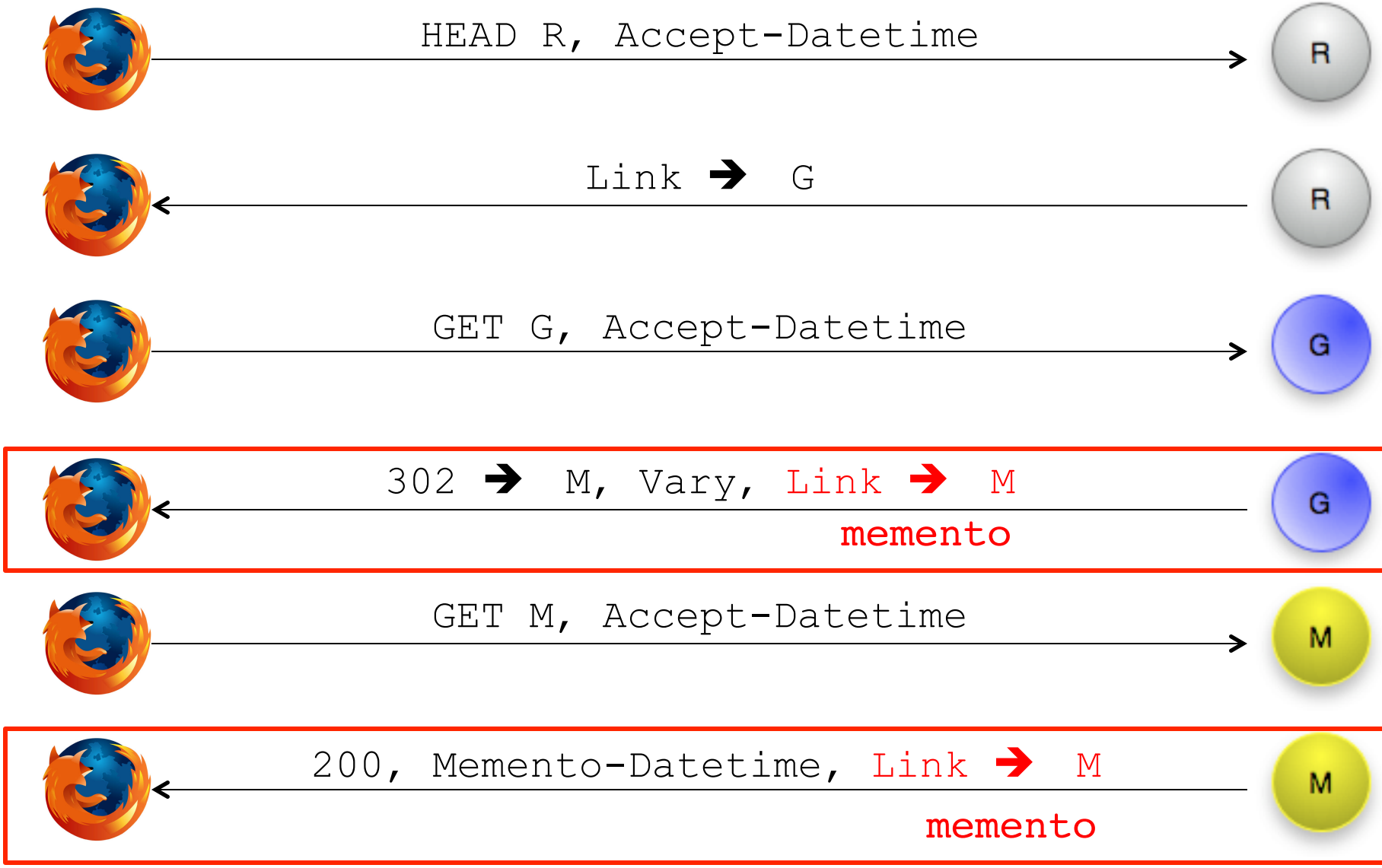
- RECOMMENDED `"timegate" Link` from Original Resource to TimeGate

- If this Link is not available, the client must try and find a TimeGate itself, via:
  - Memento Discovery approaches (see later)
  - User interaction (e.g. Preferences in an application)

HTTP Response Headers: Link

Relation Type	Original Resource	TimeGate	Memento
original	NA, except see Section 3.1.2.1	REQUIRED, 1	REQUIRED, 1
timegate	RECOMMENDED, 0 or more	NA	RECOMMENDED, 0 or more
timemap	NA	RECOMMENDED, 0 or more	RECOMMENDED, 0 or more
memento	NA, except see Section 3.1.2.1	REQUIRED, 1 or more	REQUIRED, 1 or more

Memento HTTP Flow



HTTP Response Headers: Link

- MANDATORY `"memento" Links` from TimeGate and Memento to Mementos

- `"memento" Links point` at the following Mementos know to the responding server:
  - Selected Memento (MANDATORY if a Memento is selected)
  - First Memento, Last Memento (MANDATORY)
  - Memento prev to selected one, Memento next to selected one (RECOMMENDED)
  - Other Mementos (OPTIONAL, and only if prev and next are provided)
  - Temporal order of Mementos is expressed using existing Relation Types (RFC 5829, RFC 5988): `first, last, next, prev, successor-version, predecessor-version`

HTTP Response Headers: Link

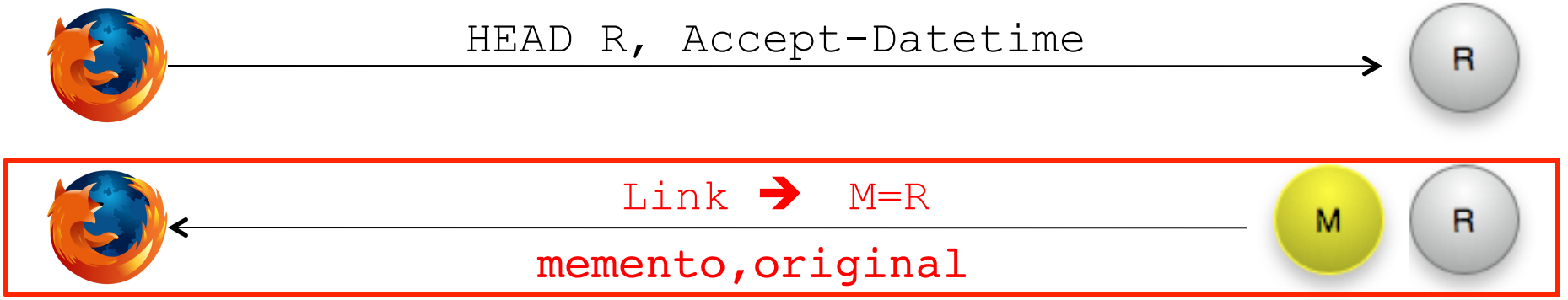
- MANDATORY `"memento" Links` from TimeGate and Memento to Mementos

- Attributes for a `"memento" Link`:
  - `datetime` (MANDATORY)
    datetime of the Memento pointed at by the link
  - `license` (OPTIONAL)
    license associated with the Memento
  - `embargo` (OPTIONAL)
    datetime until which the Memento will remain inaccessible
  - `type` (RECOMMENDED)
    mime type of the Memento.

HTTP Response Headers: Link

Relation Type	Original Resource	TimeGate	Memento
original	NA, except see Section 3.1.2.1	REQUIRED, 1	REQUIRED, 1
timegate	RECOMMENDED, 0 or more	NA	RECOMMENDED, 0 or more
timemap	NA	RECOMMENDED, 0 or more	RECOMMENDED, 0 or more
memento	NA, except see Section 3.1.2.1	REQUIRED, 1 or more	REQUIRED, 1 or more

Memento HTTP Flow



HTTP Response Headers: Link

- Both `"memento" and "original" Link` on a resource.

- The resource is its own Memento, i.e. it is a stable resource.
  - Resource that was born stable or became stable; it will not change anymore.
  - For example resources with PermaLink on news sites
  - Note the difference with `Last-Modified` header

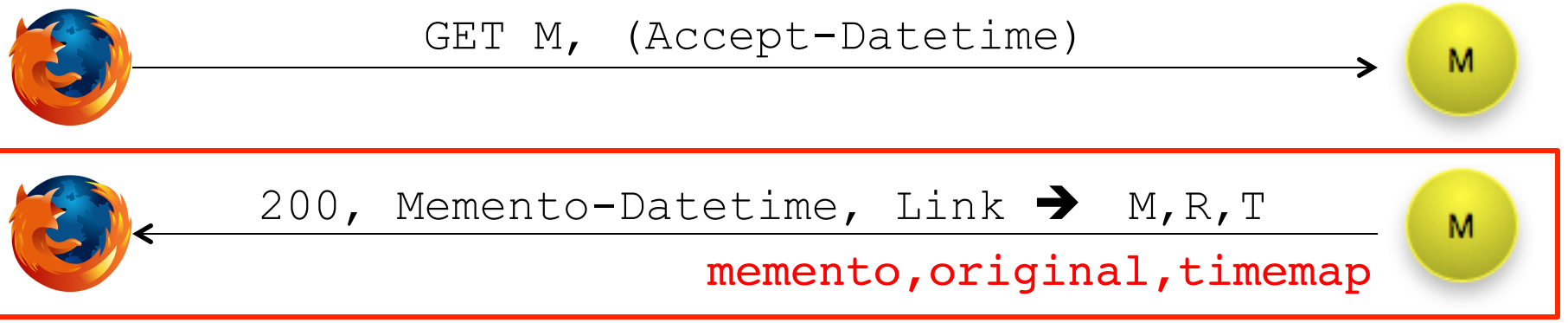
- Can still also provide a `"timegate" Link`
  - For example pointing at TimeGates for Mementos of the resource before it became stable

HTTP Response Headers: Link

Relation Type	Original Resource	TimeGate	Memento
original	NA, except see Section 3.1.2.1	REQUIRED, 1	REQUIRED, 1
timegate	RECOMMENDED, 0 or more	NA	RECOMMENDED, 0 or more
timemap	NA	RECOMMENDED, 0 or more	RECOMMENDED, 0 or more
memento	NA, except see Section 3.1.2.1	REQUIRED, 1 or more	REQUIRED, 1 or more

Los Alamos NATIONAL LABORATORY
Old Dominion University

Memento HTTP Flow



HTTP Response Headers: Link

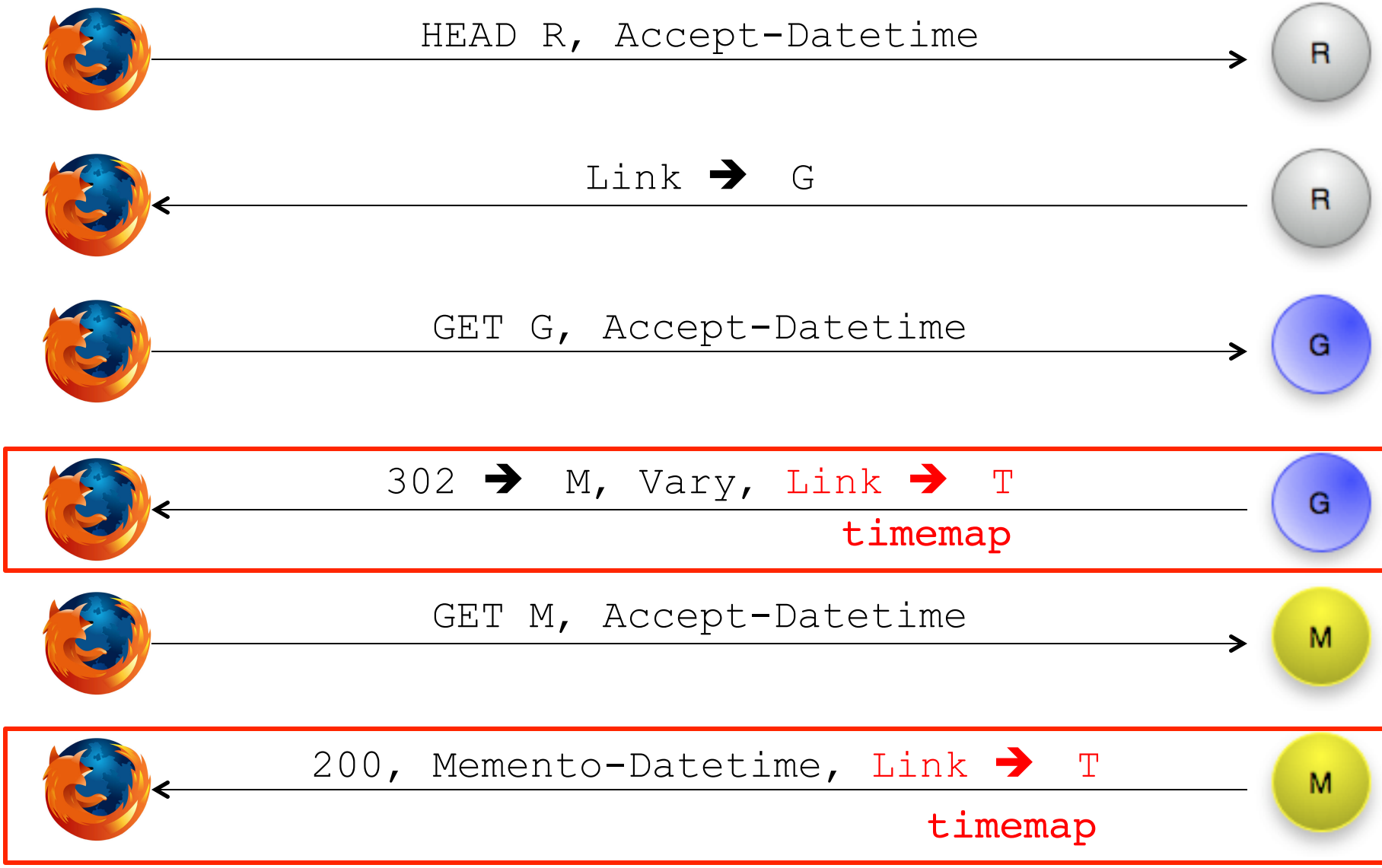
- Mementos without a TimeGate, for example:
  - Resources in snapshot archives
  - Version resources in systems that have not yet implemented TimeGates

- Should still use Memento-Datetime header
- Should still use `"original"` Link
- Can have `"timemap"` Link

HTTP Response Headers: Link

Relation Type	Original Resource	TimeGate	Memento
original	NA, except see Section 3.1.2.1	REQUIRED, 1	REQUIRED, 1
timegate	RECOMMENDED, 0 or more	NA	RECOMMENDED, 0 or more
timemap	NA	RECOMMENDED, 0 or more	RECOMMENDED, 0 or more
memento	NA, except see Section 3.1.2.1	REQUIRED, 1 or more	REQUIRED, 1 or more

Memento HTTP Flow



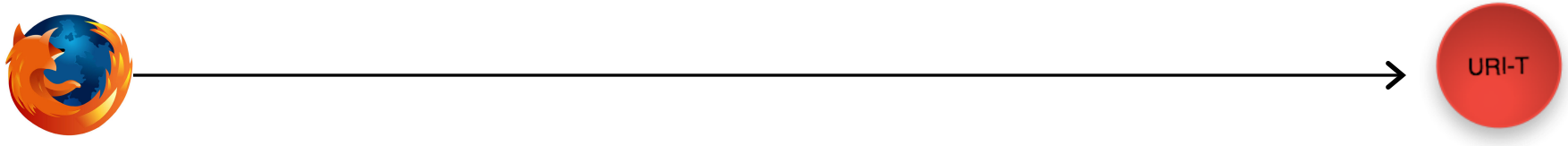
HTTP Response Headers: Link

- RECOMMENDED `"timemap" Links` from TimeGate and Mementos.

- A TimeMap is introduced to allow retrieving an inventory of Mementos for an Original Resource that the responding server is aware of. It lists:
  - URI of Original Resource (MANDATORY)
  - URI and datetime of all known Mementos (MANDATORY)
  - URI of TimeGate for Original Resource (RECOMMENDED)
  - URI of TimeMap itself (RECOMMENDED)

- Multiple TimeMap serializations possible; link-value format MANDATORY
  - application/link-format:
    see *https://datatracker.ietf.org/doc/draft-ietf-core-link-format/*

Memento HTTP Flow: GET TimeMap




```
GET /timemap/http://a.example.org HTTP/1.1
Host: arxiv.example.net
Accept: application/link-format;q=1.0
Connection: close
```

Memento HTTP Flow: TimeMap Response

```
HTTP/1.1 200 OK
Date: Thu, 21 Jan 2010 00:06:50 GMT
Server: Apache
Link: <http://arxiv.example.net/timemap/http://a.example.org>
   ; anchor="http://a.example.org"; rel="timemap"
   ; type="application/link-format"
Content-Length: 4883
Content-Type: application/link-format
Connection: close

 <http://a.example.org>;rel="original",
 <http://arxiv.example.net/timemap/http://a.example.org>
   ; rel="timemap";type="application/link-format",
 <http://arxiv.example.net/timegate/http://a.example.org>
   ; rel="timegate",
 <http://arxiv.example.net/web/20000620180259/http://a.example.org>
   ; rel="first memento";datetime="Tue, 20 Jun 2000 18:02:59 GMT"
   ; license="http://creativecommons.org/publicdomain/zero/1.0/",
 <http://arxiv.example.net/web/20091027204954/http://a.example.org>
    ; rel="last memento";datetime="Tue, 27 Oct 2009 20:49:54 GMT"
    ; license="http://creativecommons.org/publicdomain/zero/1.0/"
    ; embargo="Tue, 19 Apr 2011 00:00:00 GMT",
 <http://arxiv.example.net/web/20000621011731/http://a.example.org>
   ; rel="memento";datetime="Wed, 21 Jun 2000 01:17:31 GMT"
   ; license="http://creativecommons.org/publicdomain/zero/1.0/",
 <http://arxiv.example.net/web/20000621044156/http://a.example.org>
   ; rel="memento";datetime="Wed, 21 Jun 2000 04:41:56 GMT"
   ; license="http://creativecommons.org/publicdomain/zero/1.0/",
 ...
```

The Memento Framework:

Discovery to Support Integration of Past and Current Web

TimeGate Discovery

Los Alamos NATIONAL LABORATORY

Old Dominion University

Batch Discovery of TimeGates: robots.txt

- `robots.txt` file is used by Web servers to convey crawling policies

- Web crawlers (such as for archives) retrieve and parse it
- De-facto standard, no official endorsement
- Extended with new directives, including by Google

```
User-agent: *          # Reject all crawlers
Disallow: /
                       # Google Sitemap extension
Sitemap: http://some.example.com/me/sitemap.xml

User-agent: NiceBot    # Select only NiceBot
Crawl-delay: 10        # 10 seconds between requests
```


BY NC SA

Batch Discovery of TimeGates: robots.txt

- Add `TimeGate` and `Archived` directives to support discovery of TimeGates known to the server
- User agent should concatenate desired URL with TimeGate link
- Archived value is truncated host/path or * to describe a general web archive

```
TimeGate: http://a.example.org/w/index.php/Special:TimeGate/
Archived: a.example.org/w/
```

```
TimeGate: http://arxiv.example.net/timegate/
Archived: a.example.org/
Archived: www.a.example.org/
```

```
TimeGate: http://arxiv.example.net/timegate/
Archived: *
```

http://mementoweb.org/guide/robotstxt/

The Memento Framework:

Discovery to Support Integration of Present and Past Web

Discovery via TimeMaps: All Mementos for a given Original Resource known by an archive

TimeMap Overview

- A TimeMap is an inventory of Mementos for an Original Resource that the responding server is aware of. It lists at least:
  - URI of Original Resource
  - URI and datetime of all known Mementos
  - URI of TimeGate for Original Resource
  - URI of TimeMap itself

- Multiple TimeMap serializations possible:
  - application/link-format mandatory

    see *https://datatracker.ietf.org/doc/draft-ietf-core-link-format/*

  - RDF TimeMaps proposed

TimeMaps: Link Format Syntax

- Document in the format of the value of the Link HTTP Header
- Format:

```
<URI>;rel="type";attr="val",
<URI2>…
```

- `rel` is the relationship between context URI and the URI in <>s
- The Context URI for TimeMaps is the URI with rel="original"
- Other `rel` types link to Mementos, TimeGates etc.

```
<http://cnn.com/>;rel="original",

<http://web.archive.org/web/20010911223004/http://
 cnn.com>;rel="memento";datetime="Mon, 11 Sep 2001 22:30:04 GMT"
```

TimeMaps: Link Attributes

- Existing Attributes for Links
  - `"rel"` The type of relationship
  - `"type"` The (mime) formatt of the linked resource
  - `"title"` Title of the linked resource
  - `"hreflang"` Language of linked resource
  - `"media"` Intended media (eg screen)
  - `"anchor"` URI to override context URI for link

- New `rel` types introduced:
  - `"original"` The Original Resource
  - `"memento"` A Memento of the Original
  - `"timegate"` A TimeGate for the Original
  - `"timemap"` A TimeMap of Mementos of the Original

TimeMaps: Link Attributes

- New Attributes for Mementos:
  - `datetime` The Memento-Datetime
  - `license` License associated with Memento
  - `embargo` Time after which Memento is available
  - Are others necessary?

```
<http://cnn.com/>;rel="original",
<http://web.archive.org/timegate/cnn.com/>;rel="timegate",
<http://web.archive.org/timemap/link/cnn.com/>;rel="timemap",
<http://web.archive.org/web/20010911223004/http://
 cnn.com>;rel="memento";datetime="Mon, 11 Sep 2001 22:30:04 GMT"
<http://web.archive.org/web/…/cnn.com/>;rel="memento";
    license="http://archive.org/license/1";
    datetime="…";embargo="Mon, 20 Jul 2011 00:00:00 GMT",
…
```

The Memento Framework:

Discovery to Support Integration of Past and Current Web

Memento Discovery: All Mementos known by an archive

Batch discovery of Mementos: robots.txt

• `robots.txt` file is used by Web servers to convey crawling policies

• Support discovery of Mementos through robots.txt via existing `User-agent` and `Allow` directives
• Use value `memento` for `User-agent` to convey that the value for the Memento-Datetime header must remain sticky when crawling/mirroring Mementos

```
User-agent: *
Disallow: /
User-agent: memento
Allow: /web/
```


Los Alamos NATIONAL LABORATORY
Old Dominion University

Batch discovery of Mementos: Memento Feeds

- Concept:
  - Archives publish feeds in which each entry provides details about a specific Memento, e.g. Memento-Datetime, Original Resource, etc.
  - As new Mementos become available, new feeds with new entries are published
  - Once published, feeds remain static

- Technology:
  - To be decided in collaboration with IIPC
  - Inspired by the approach and functionality of CDX files (see http://www.archive.org/web/researcher/cdx_legend.php) but:
    - With Memento-specific extensions;
    - Possibly using different serialization;
    - Including mechanisms to discover these feeds.

The Memento Framework:

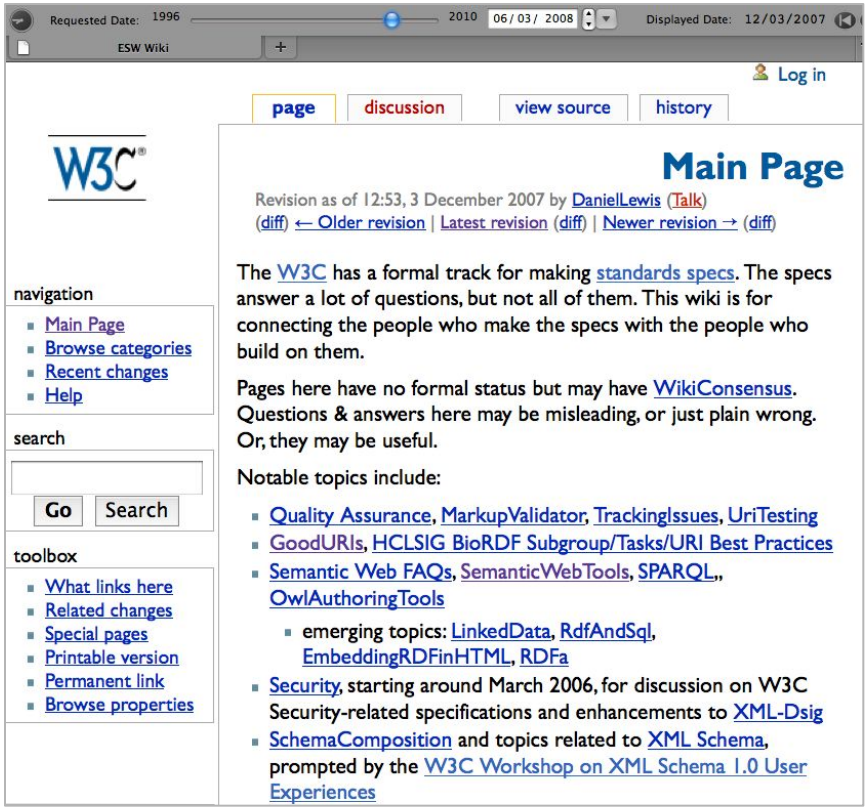
Tools

Memento Client Support

- Several client tools developed by us and others
- Add-ons for FireFox (operational) and Internet Explorer (experimental)
- Applications for Android (operational) and iPhone/iPad (in development)
- Paper in Code4Lib Journal *http://journal.code4lib.org/articles/4979*

CC BY NC SA

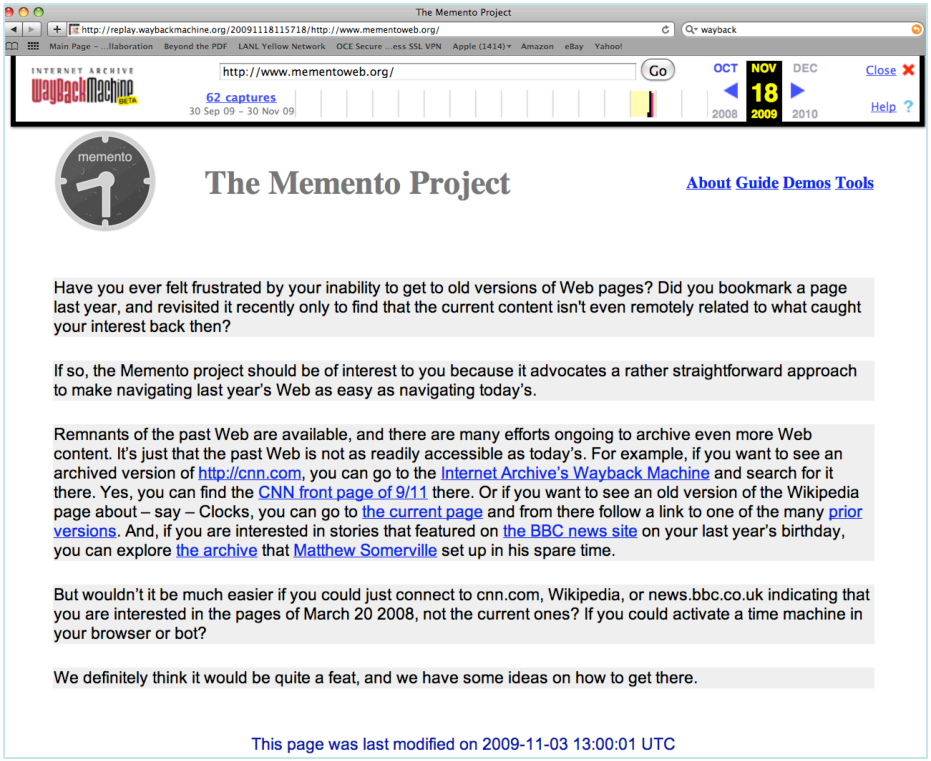
Memento Server Support



- Plug-in for MediaWiki (operational)
- Used on W3C's main wiki
- Please install it for your MediaWiki!

http://www.mediawiki.org/wiki/Extension:Memento

Memento Server Support



- Memento-compliant Wayback software:
  - In production at the Internet Archive
  - Available to Web archives, worldwide
  - Please have your favorite Web Archive experiment with the new 1.6 version!

http://mementoweb.org/tools/wayback/

Memento Server Validator

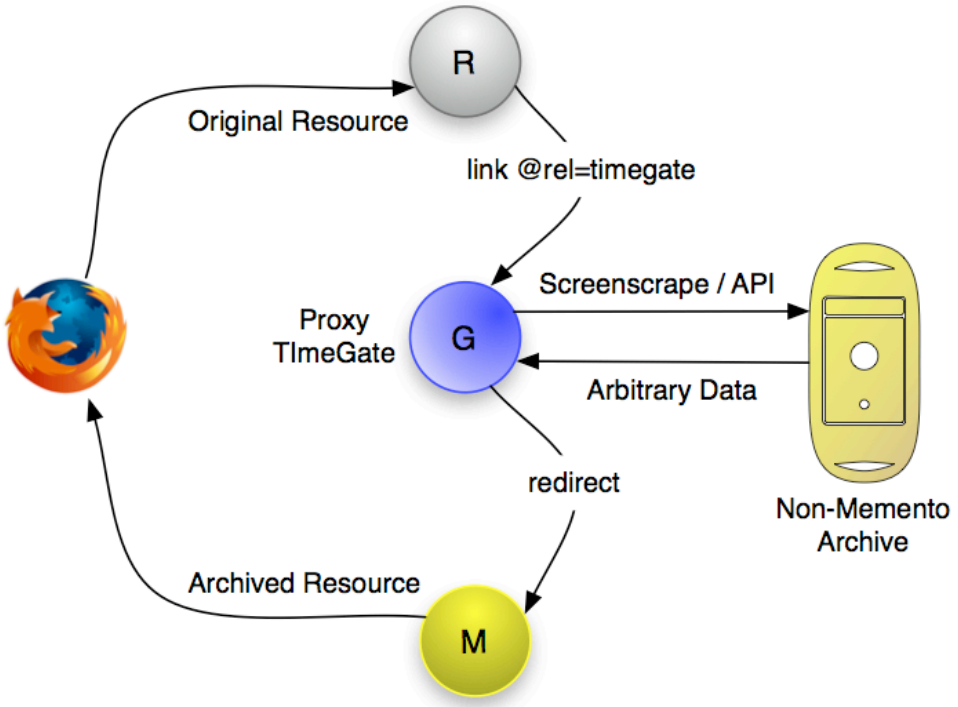
Test Details

Tests for TimeGate: *http://memento.waybackmachine.org/memento/timegate/http://www.ietf.org/*	
Test Result	URI/Date Returned
✓ TimeGate returns status 302	
✓ TimeGate has Location header for Memento	http://memento.waybackmachine.org/memento/20100402222407/http://www.ietf.org/
✓ TimeGate acknowledges Accept-Datetime in Vary header	
✓ Link Header has link with rel=original	http://www.ietf.org/
✓ Link Header has link with rel=timebundle	http://memento.waybackmachine.org/list/timebundle/http://www.ietf.org/
✓ Link Header has link with rel=timemap	http://memento.waybackmachine.org/list/timemap/link/http://www.ietf.org/
✓ TimeMap Link has application/link-format type	
✓ Link Header has one or more links with rel=memento	http://memento.waybackmachine.org/memento/20100219084310/http://www.ietf.org/
✓ -- Selected Memento in Link Header with rel=memento	
✓ Memento has parsable datetime attribute	Fri, 19 Feb 2010 08:43:10 GMT
✓ Memento has parsable datetime attribute	Wed, 06 Nov 1996 11:49:54 GMT

- Server side client:
  - Attempts to perform all Memento actions against a given URI
  - Reports success/failure of the interactions and warnings for optional aspects
  - Kept up to date with IETF Internet Draft

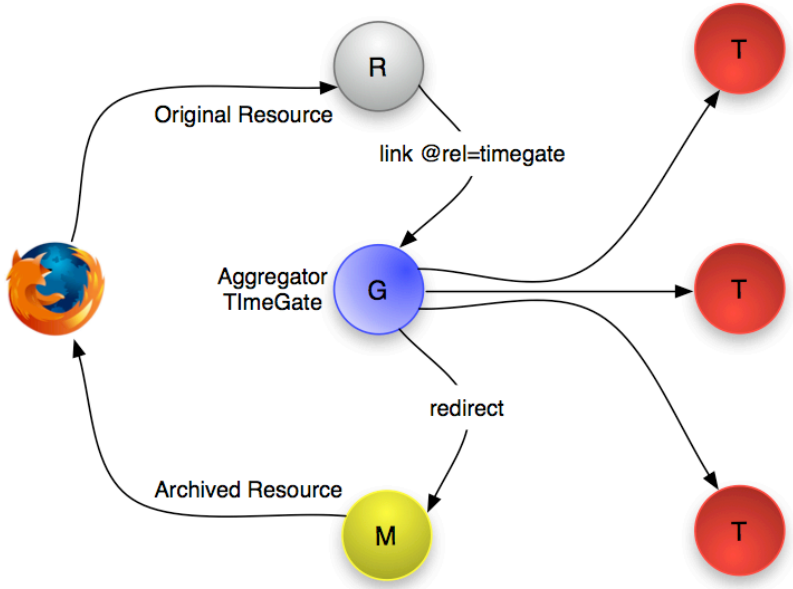
http://mementoweb.org/tools/validator/

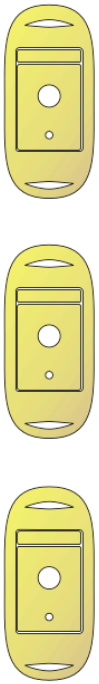
Memento Proxy Support



- Several systems that host Mementos made Memento-compliant "by proxy"
  - Many major Web Archives that do not yet run Memento-compliant software
  - 3,000+ MediaWiki systems, including Wikipedia, Wikia
- We would love all of these to become natively Memento compliant!

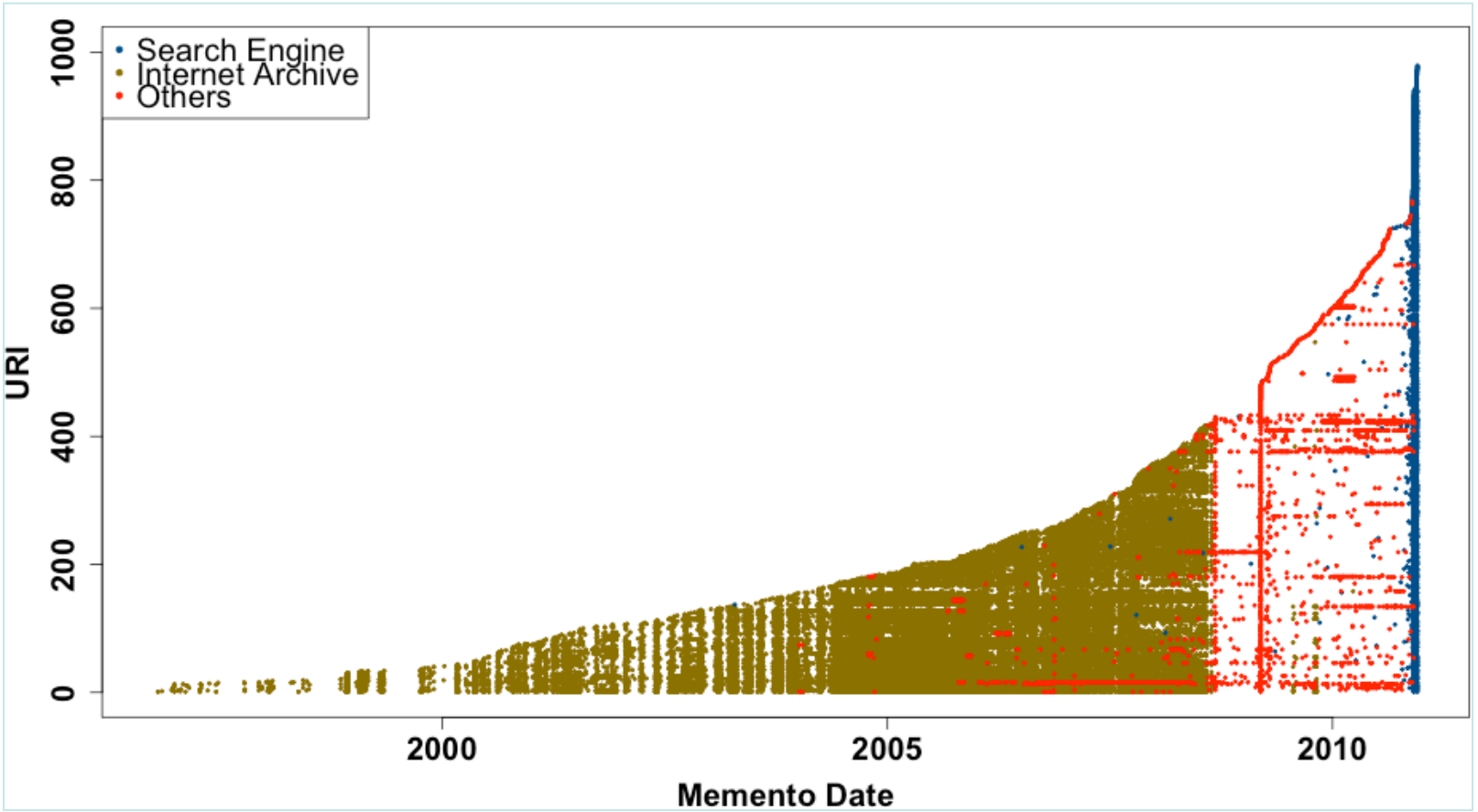
Memento Aggregator TimeGate





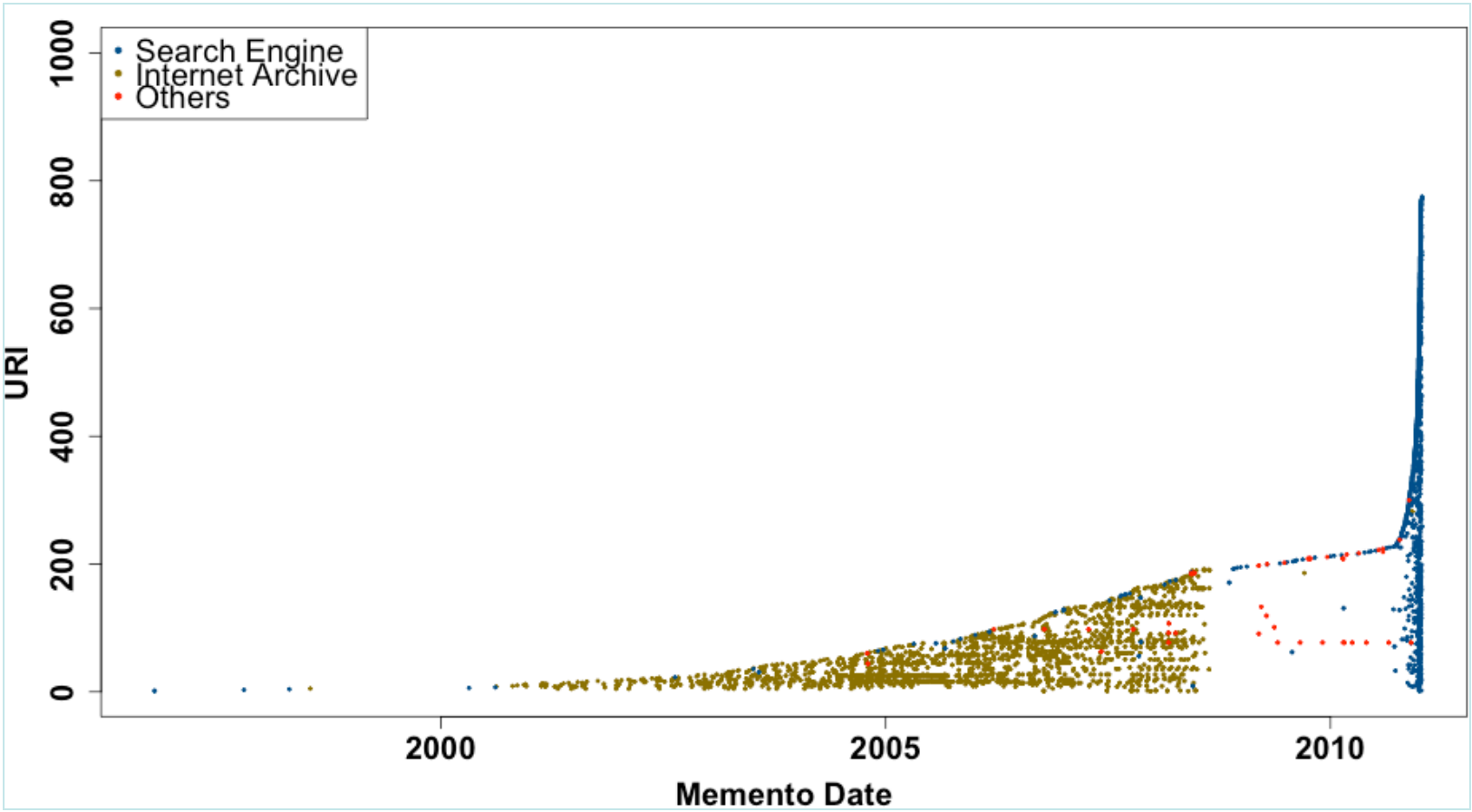
- Aggregates all known TimeGates
  - Proxies
  - Native Implementations
- Redirects to authoritative TimeGates (Wikipedia, Transactional Archives)
- Currently implemented with BerkeleyDB
- Future version to use FaceBook's Cassandra platform

Aggregators Find More Mementos!



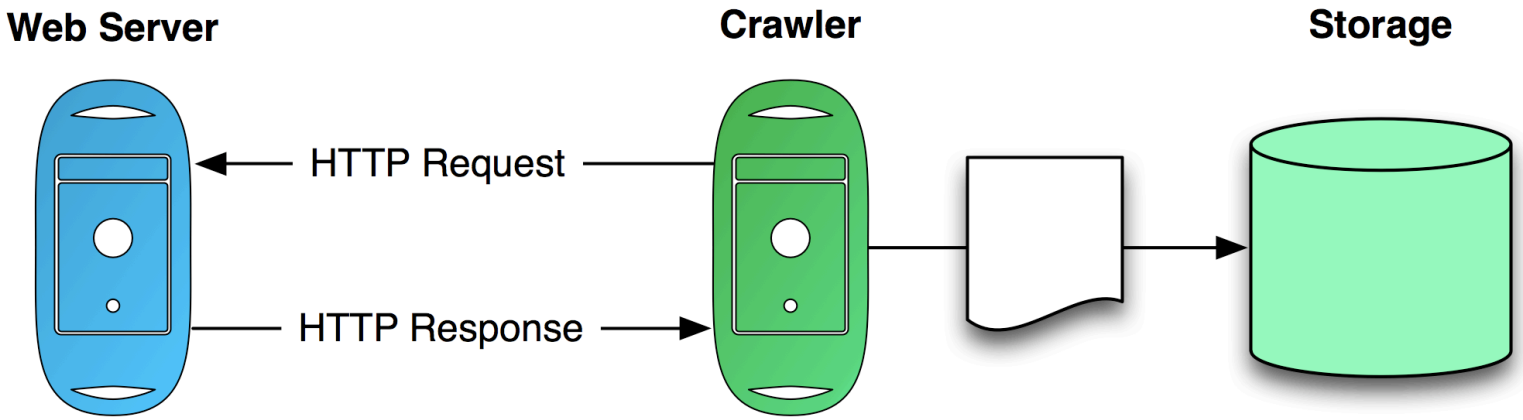
- 1000 URIs sampled from delicious.com
- 1 dot = 1 Memento (x=Memento-Datetime, y=URI of Original Resource)
- Sorted by URI longevity

But Still Too Few Mementos To Be Found…



- 1000 URIs sampled from search engine result pages;
- See: “How Much of the Web is Archived?” JCDL 2011

Crawl-Based Web Archives



Observations
For example: Heritrix crawler for Internet Archive

Crawl-Based Web Archives

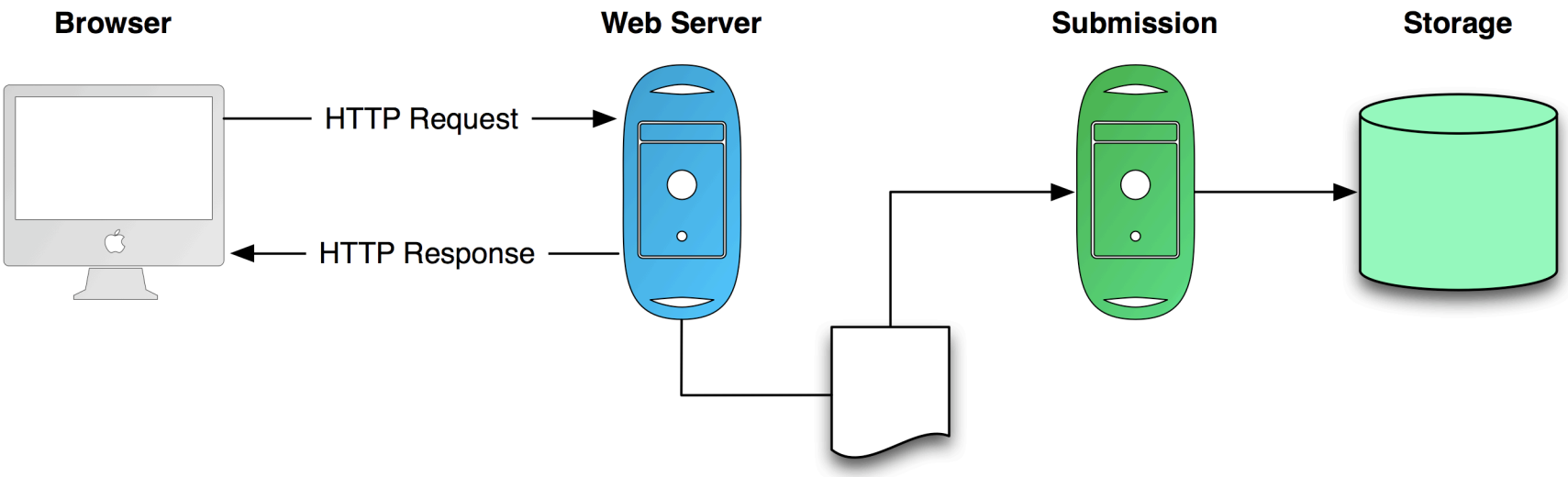
• Collect discreet observations of resources, not their entire evolution.

• Can be rejected (`robots.txt`, by user-agent, by host IP)

• Can be deceived (cloaking, by geo-location, by user-agent).

• Coverage of particular Web server dependent on crawl-strategy.

Server-Side Transactional Web Archives



Change History
For example: TTApache, PageVault, Vignette Web Capture

Server-Side Transactional Web Archives

- Collect all representations served by to-be-archived server.

- To-be-archived server needs to cooperate.
    - Incentives e.g. institutional memory, official record of Web presence.

- Archival coverage restricted by to-be-archived server, does not include external servers (e.g. embedded resources).

- To be archived server can submit falsified information.

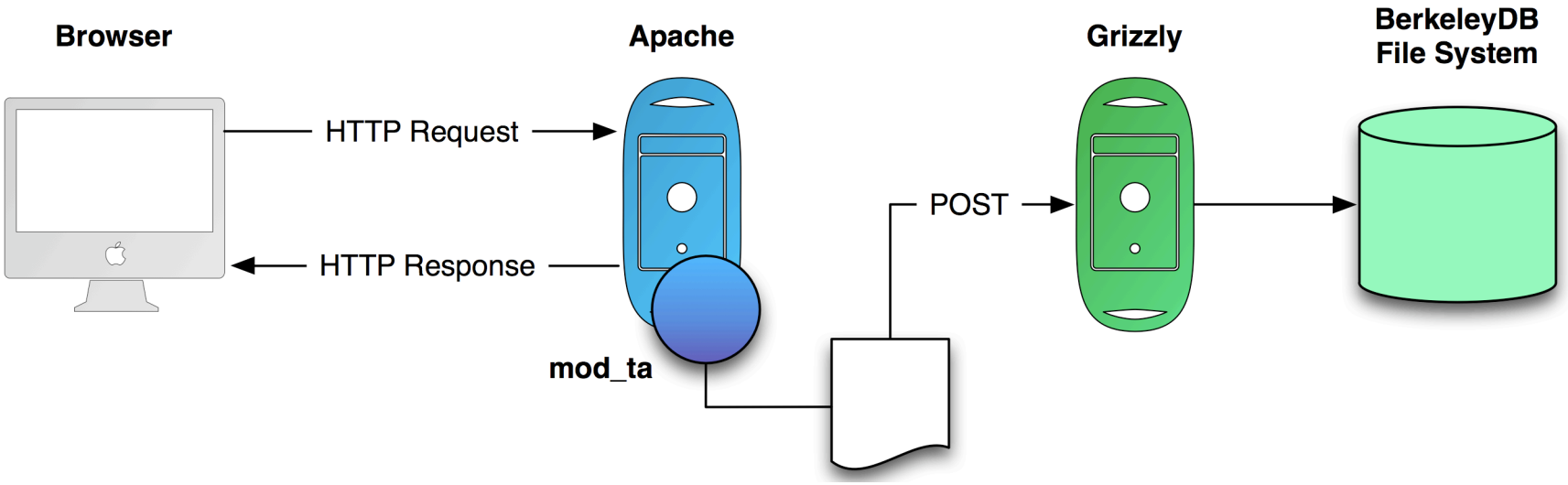
- Archival collection management: what to keep, what not (e.g. significant changes, deduplication, …).

BY NC SA

Development of Transactional Web Archive Software

Capture:

- Apache connection filter module captures URI, headers, body
- POSTs in real-time to transactional archive



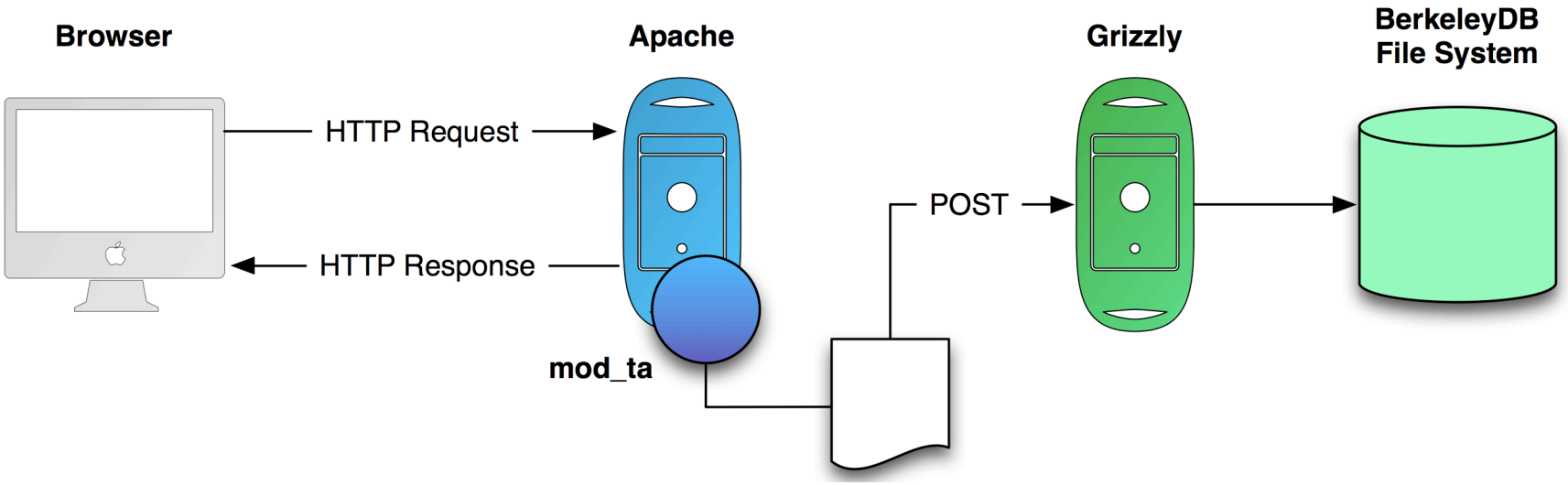
Access:

- Online, real time access via Memento TimeGates
- Batch Export via WARC files for long term preservation

Development of Transactional Web Archive Software

Capture:

- Apache connection filter module captures URI, headers, body
- POSTs in real-time to transactional archive



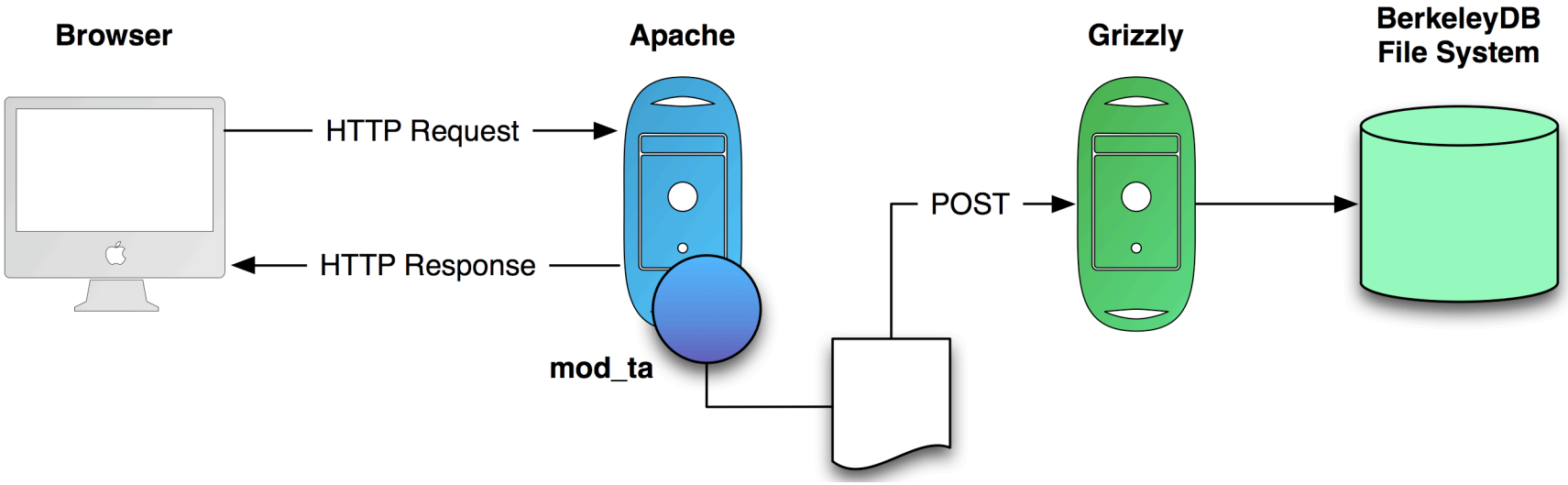
Submit:

- Java-Grizzly-Jersey submission interface application
- Berkeley DB metadata store
- FS store for body and headers

Development of Transactional Web Archive Software

Access:

- Transactional archive natively supports Memento
- Immediate availability of archived content
- Export of WARC, e.g. for long-term archiving in other environment

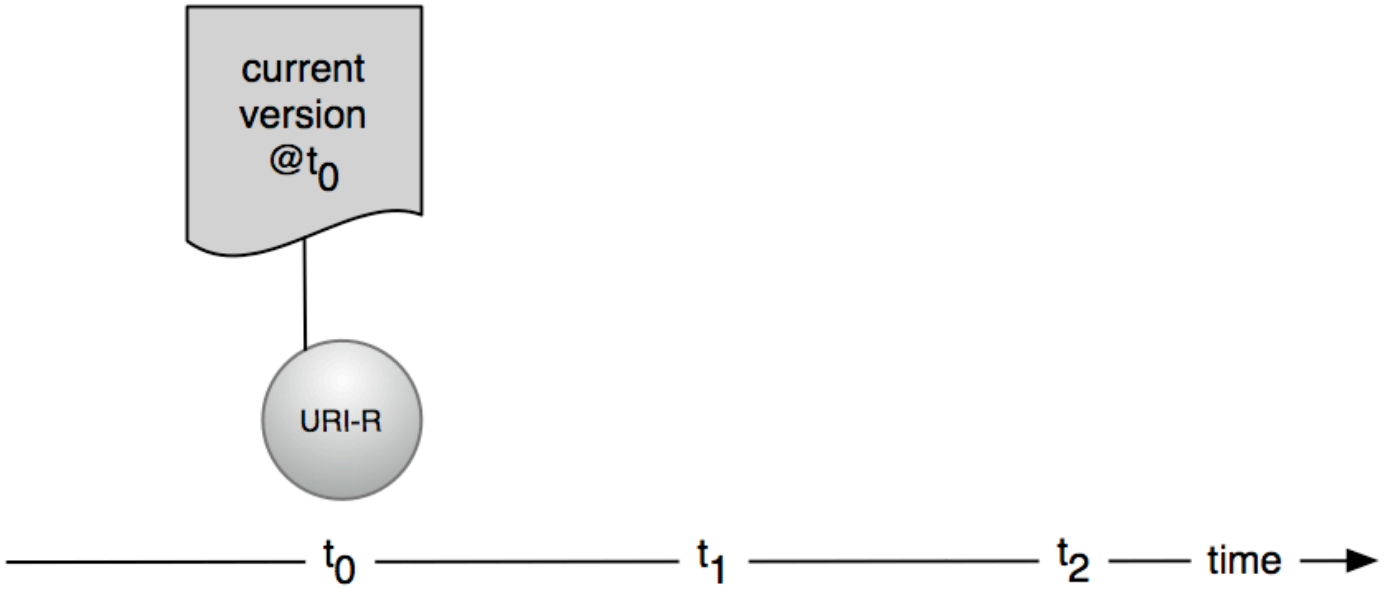


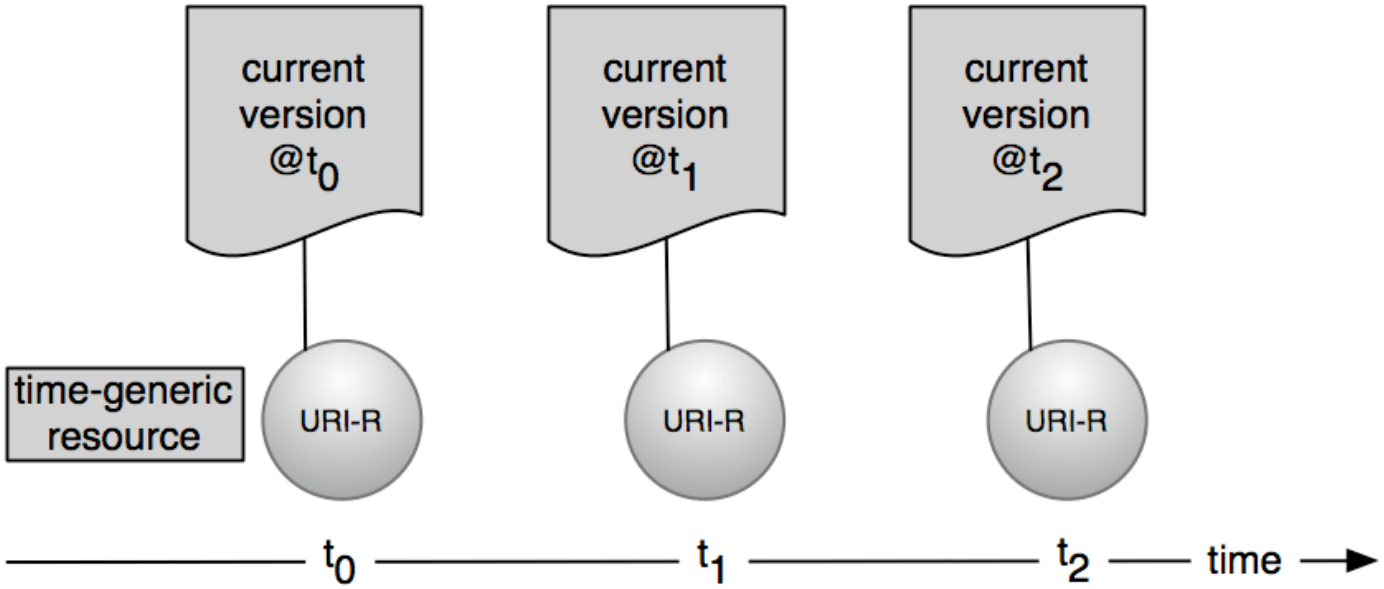
Development Timeline:

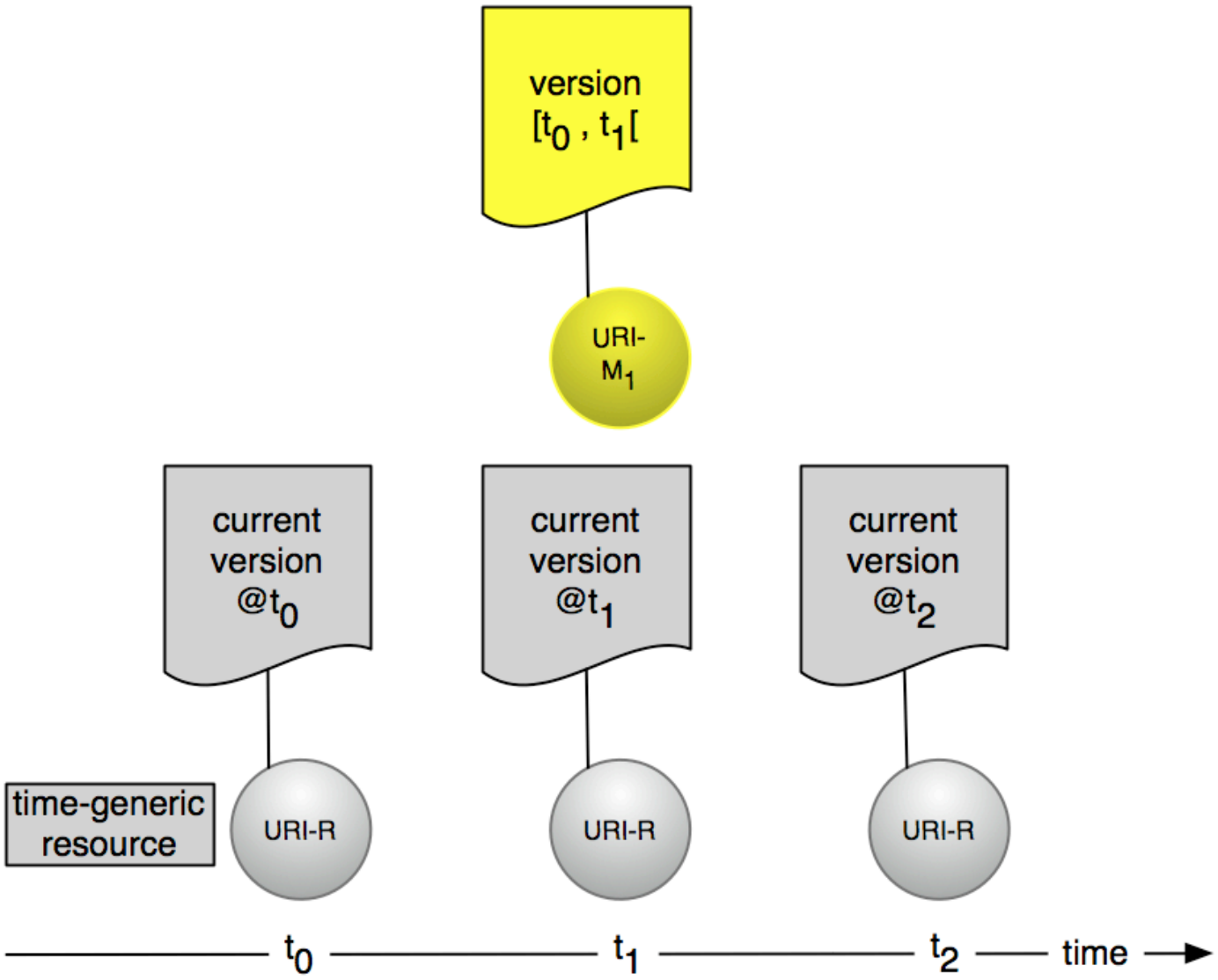
- Ongoing development (LANL) and testing (ODU)
- Submit/Access finalized; coding focus on collection management
- Expected release as open source, 3rd Quarter 2011

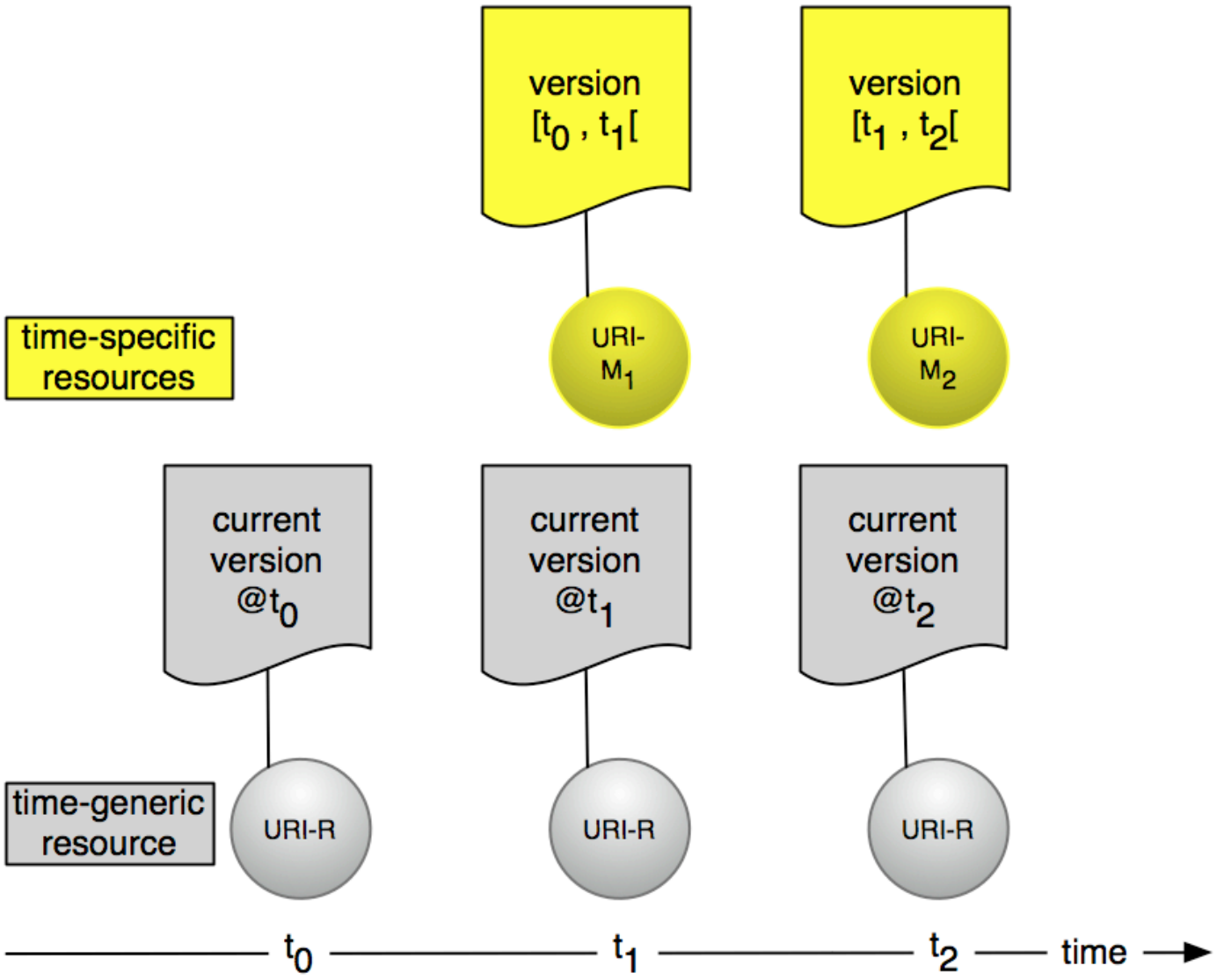
The Memento Framework:

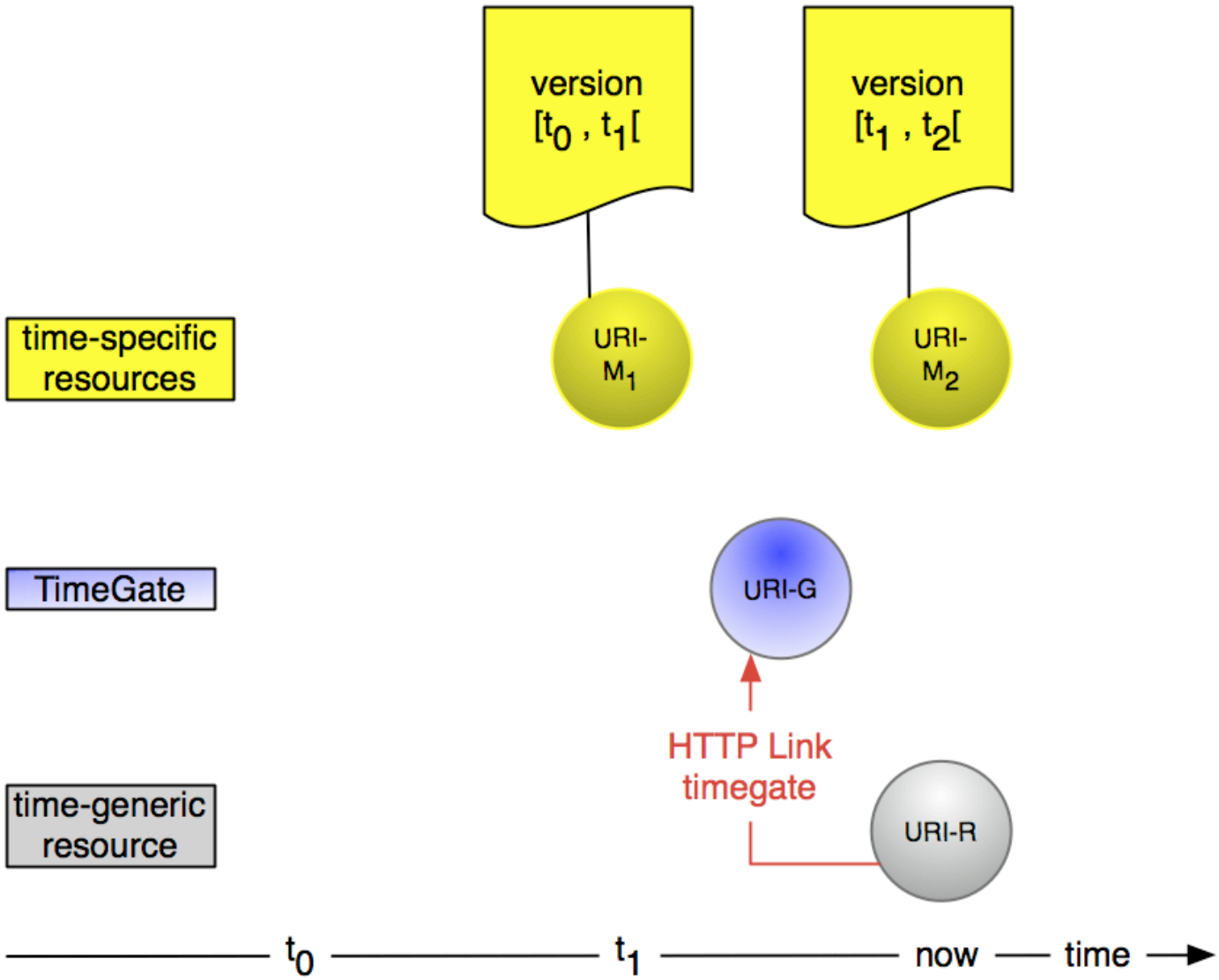
Resource Versioning

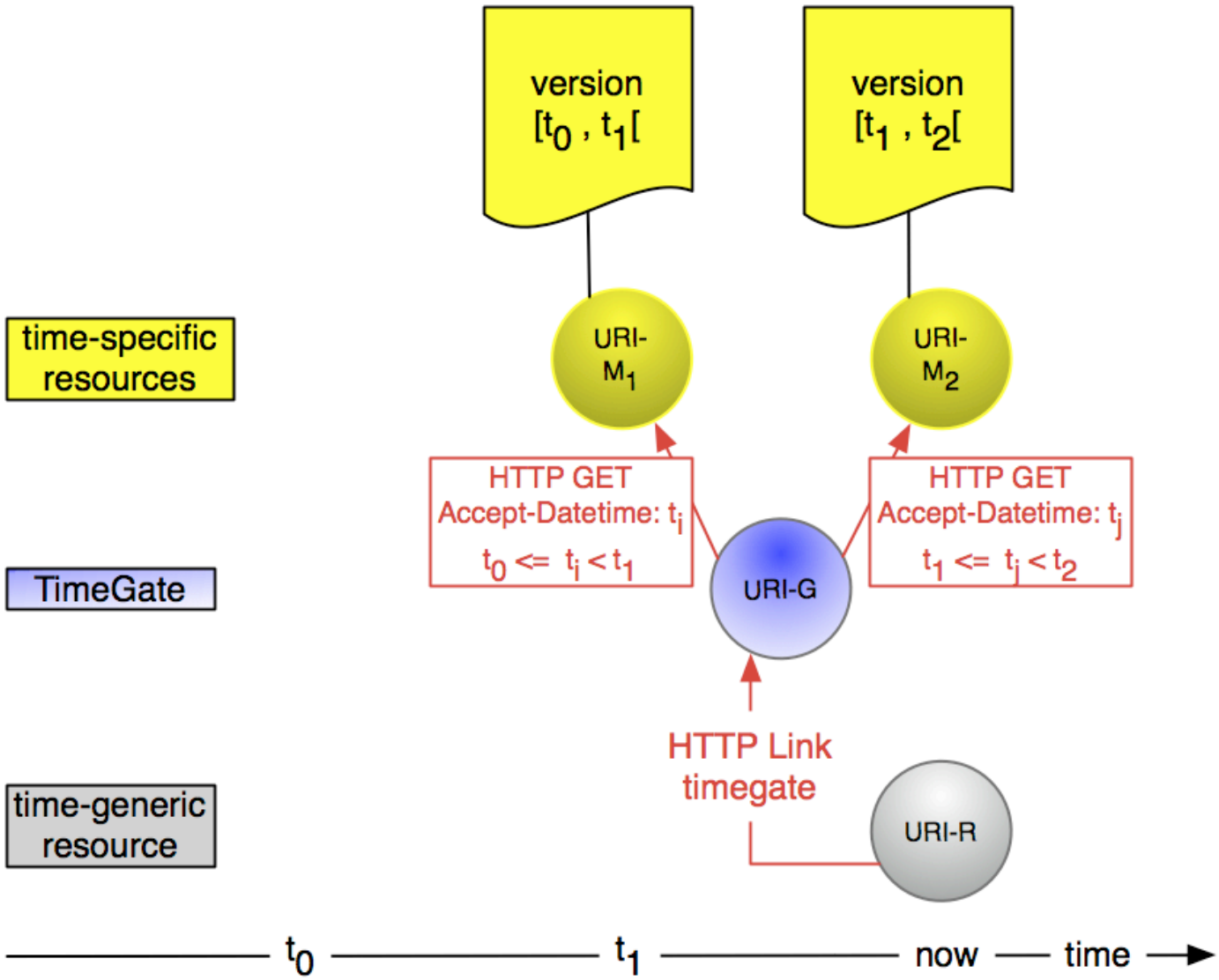

current
version
@t0
URI-R
t0
t1
t2
time

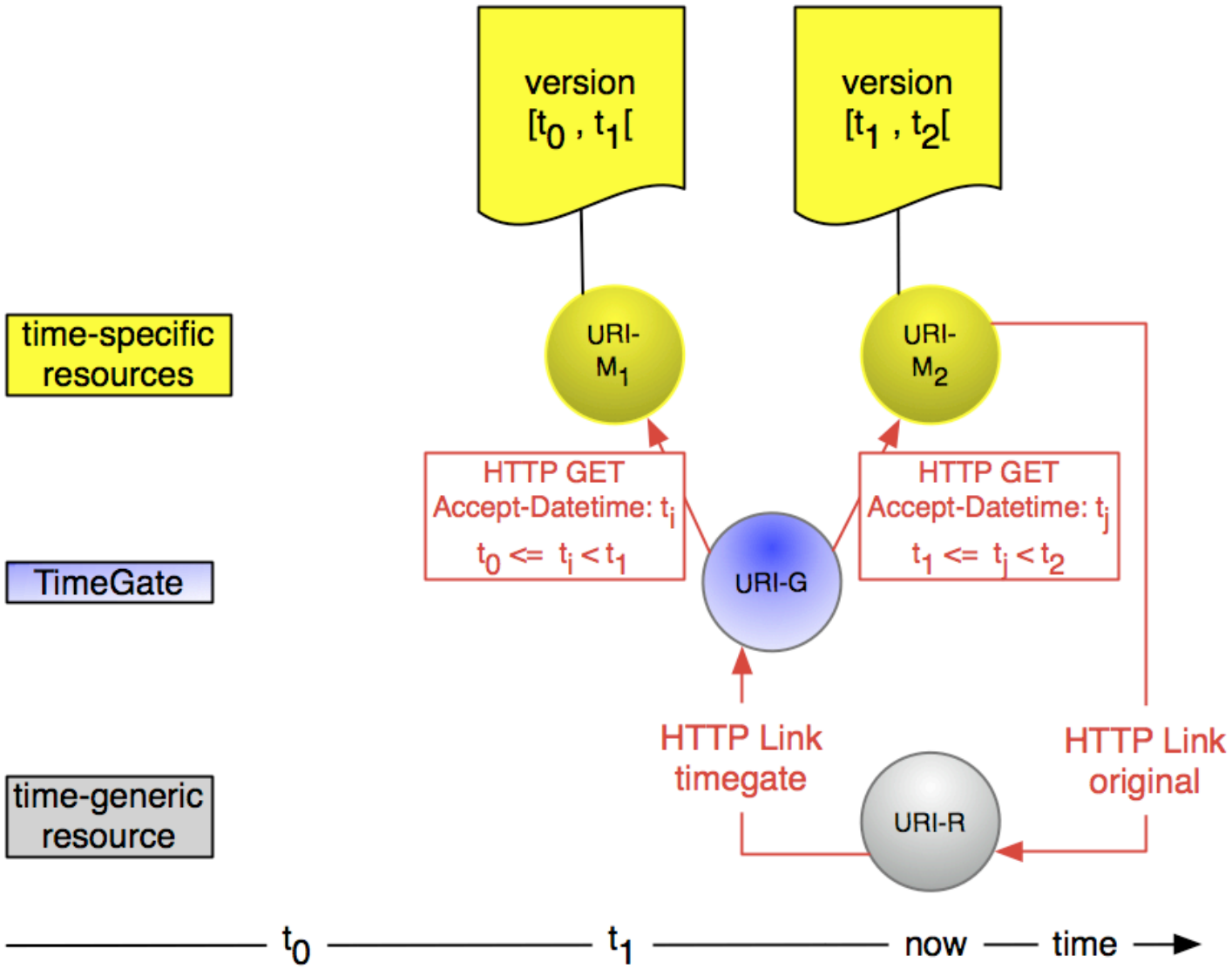

current version @t0
current version @t1
current version @t2
time-generic resource
URI-R
URI-R
URI-R
t0
t1
t2
time


version
[t0 , t1[
URI-
M1
current
version
@t0
current
version
@t1
current
version
@t2
time-generic
resource
URI-R
URI-R
URI-R
t0
t1
t2
time


version
[t0 , t1[
version
[t1 , t2[
time-specific
resources
URI-
M1
URI-
M2
current
version
@t0
current
version
@t1
current
version
@t2
time-generic
resource
URI-R
URI-R
URI-R
t0
t1
t2
time

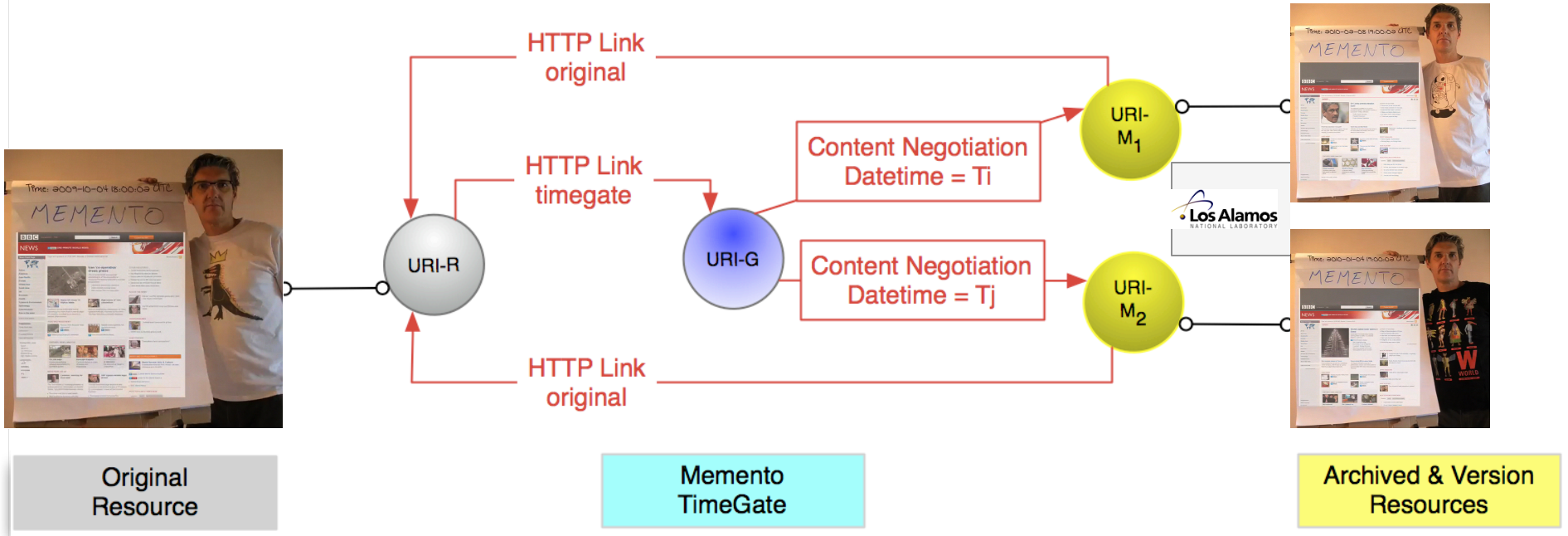

version
[t0 , t1[
version
[t1 , t2[
time-specific
resources
URI-
M1
URI-
M2
TimeGate
URI-G
HTTP Link
timegate
time-generic
resource
URI-R
t0
t1
now
time


version
[t0 , t1[
version
[t1 , t2[
time-specific resources
URI-M1
URI-M2
HTTP GET
Accept-Datetime: ti
t0 <= ti < t1
HTTP GET
Accept-Datetime: tj
t1 <= tj < t2
TimeGate
URI-G
HTTP Link
timegate
time-generic resource
URI-R
t0
t1
now
time


version
[t0 , t1[
version
[t1 , t2[
time-specific resources
URI-M1
URI-M2
HTTP GET
Accept-Datetime: ti
t0 <= ti < t1
HTTP GET
Accept-Datetime: tj
t1 <= tj < t2
TimeGate
URI-G
HTTP Link
timegate
HTTP Link
original
time-generic resource
URI-R
t0
t1
now
time

Memento Framework

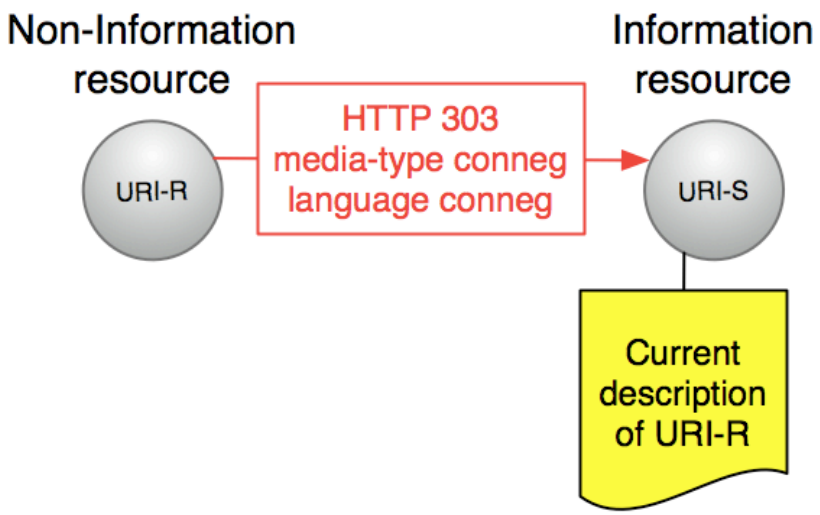
Original Resource: *http://lanlsource.lanl.gov/pics/picoftheday.png*

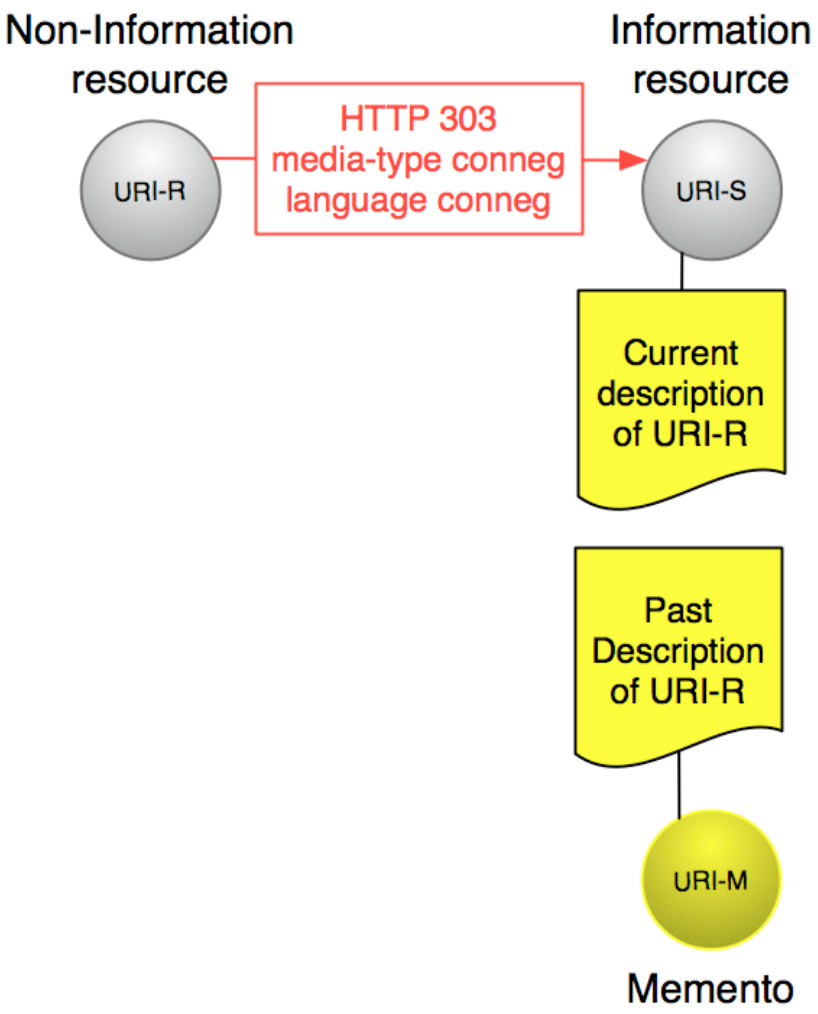


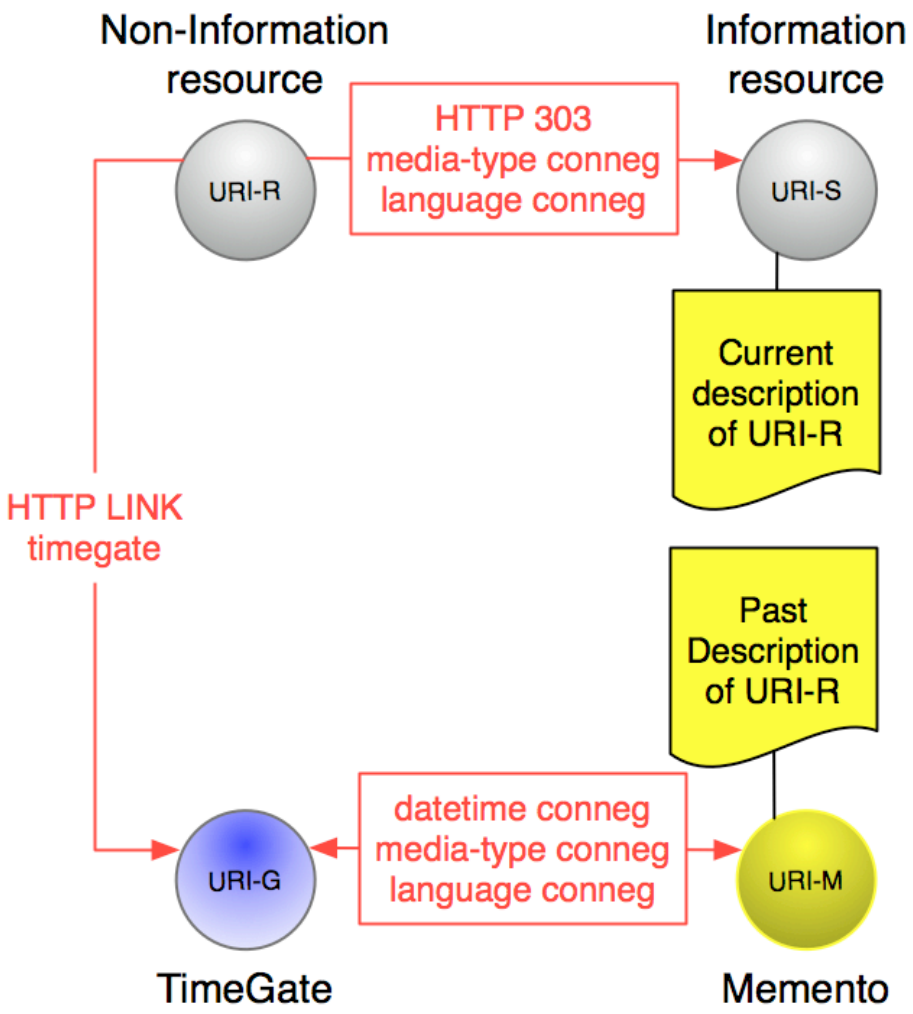
Time Travel across Versions of a Picture of the Day

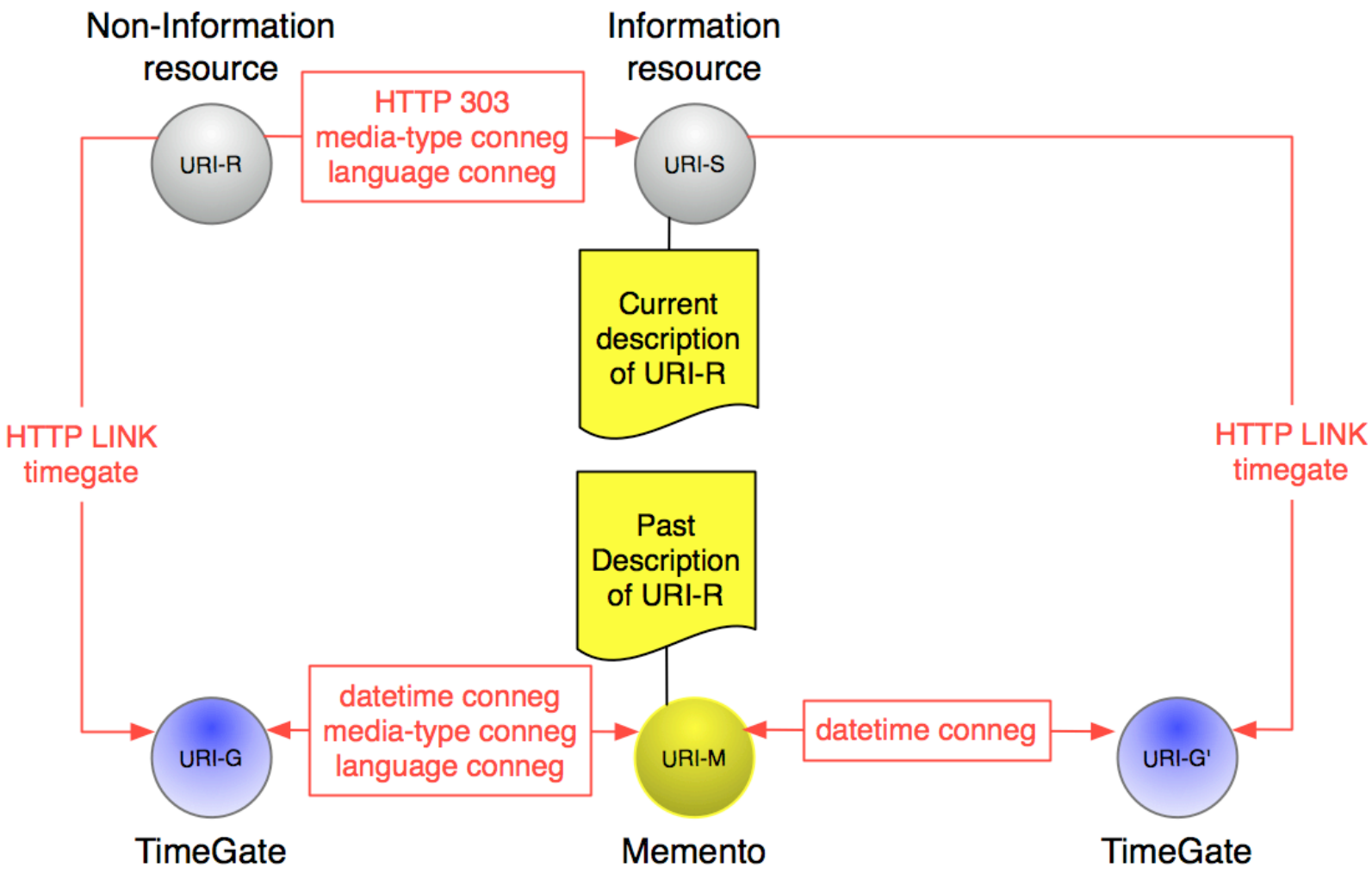
Movie at: *http://www.mementoweb.org/demo/picoftheday.mov*

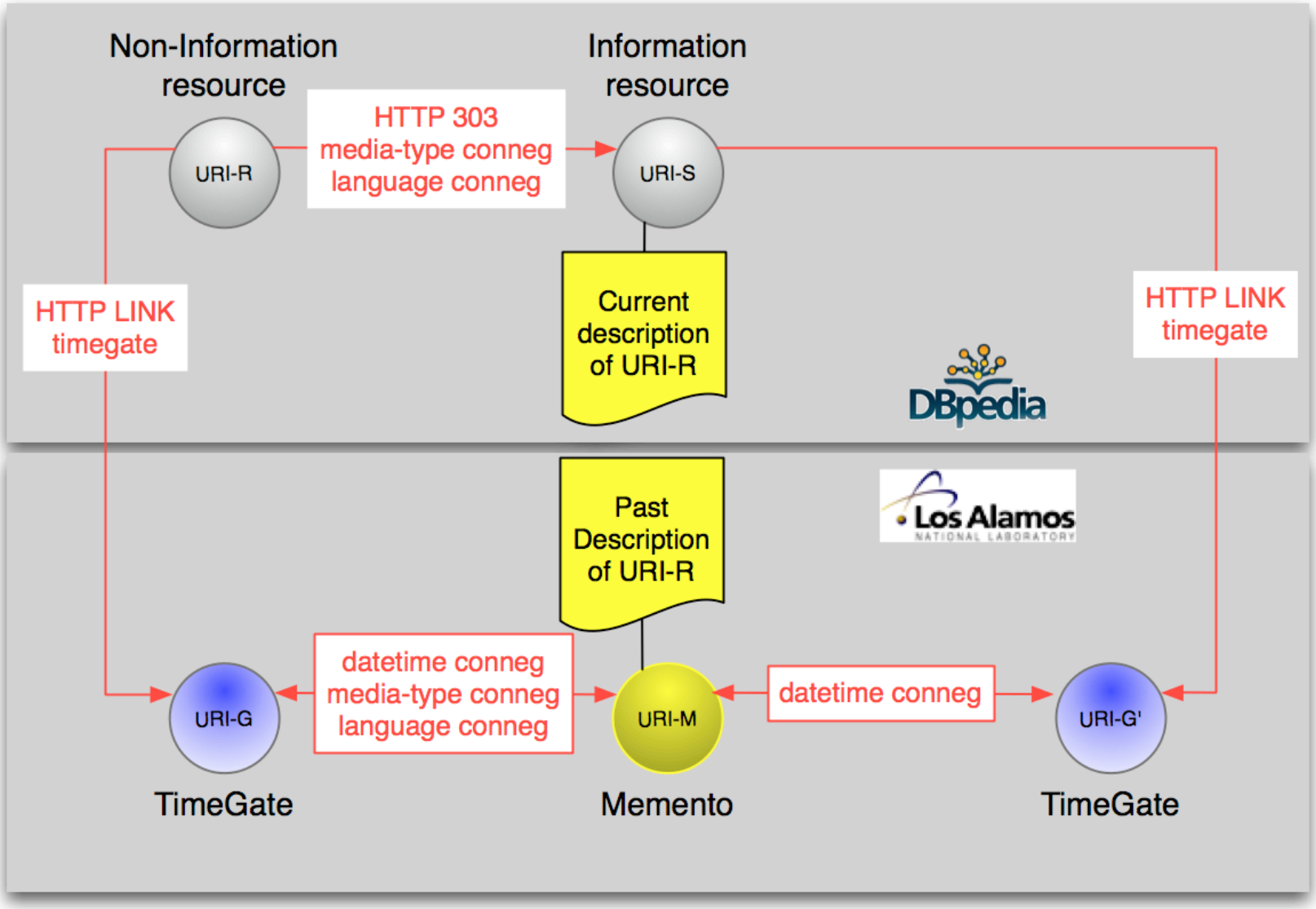
BY NC SA


Non-Information resource
Information resource
URI-R
HTTP 303
media-type conneg
language conneg
URI-S
Current description of URI-R


Non-Information resource
Information resource
HTTP 303
media-type conneg
language conneg
URI-R
URI-S
Current description of URI-R
Past Description of URI-R
URI-M
Memento

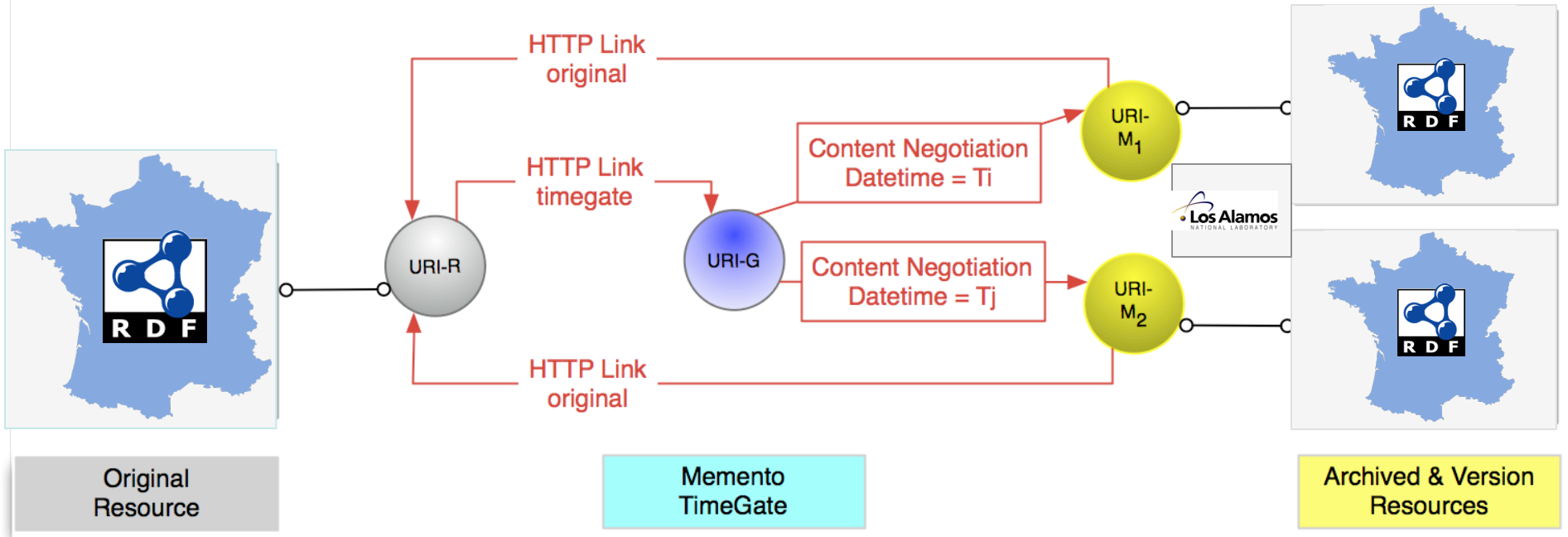

Non-Information resource
Information resource
HTTP 303
media-type conneg
language conneg
URI-R
URI-S
Current description of URI-R
HTTP LINK
timegate
Past Description of URI-R
datetime conneg
media-type conneg
language conneg
URI-G
URI-M
TimeGate
Memento


Non-Information resource
Information resource
URI-R
HTTP 303
media-type conneg
language conneg
URI-S
Current description of URI-R
HTTP LINK
timegate
HTTP LINK
timegate
Past Description of URI-R
URI-G
datetime conneg
media-type conneg
language conneg
URI-M
datetime conneg
URI-G'
TimeGate
Memento
TimeGate

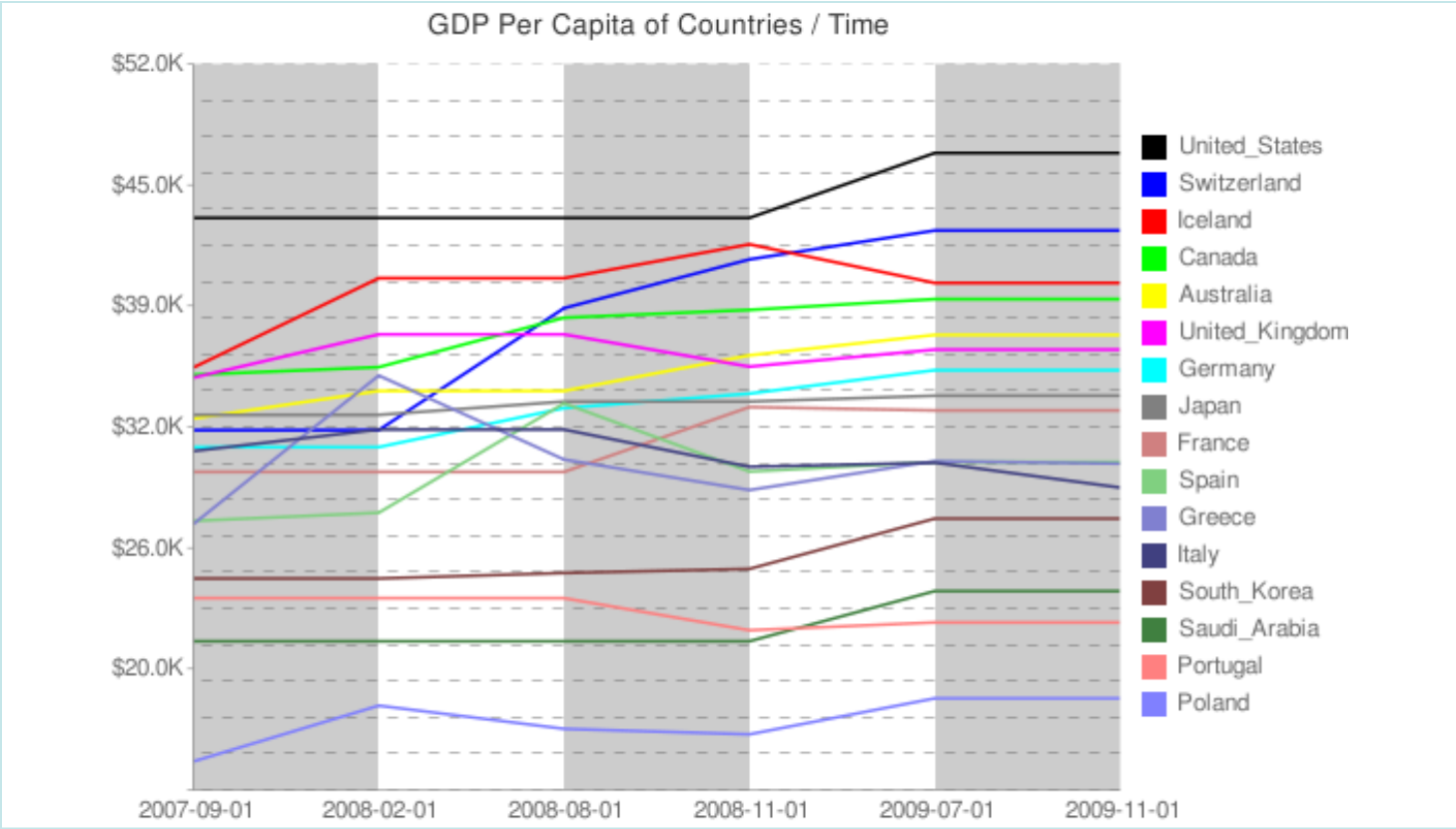

Non-Information resource
Information resource
URI-R
HTTP 303
media-type conneg
language conneg
URI-S
HTTP LINK
timegate
Current description of URI-R
HTTP LINK
timegate
DBpedia
Past Description of URI-R
Los Alamos
NATIONAL LABORATORY
datetime conneg
media-type conneg
language conneg
URI-G
URI-M
datetime conneg
URI-G'
TimeGate
Memento
TimeGate

Memento Framework

Original Resource: http://dbpedia.org/resource/France



Time-Series Analysis across DBpedia Versions



Data collected through HTTP Navigation

paper at *http://arxiv.org/abs/1003.3661*

The Memento Framework:

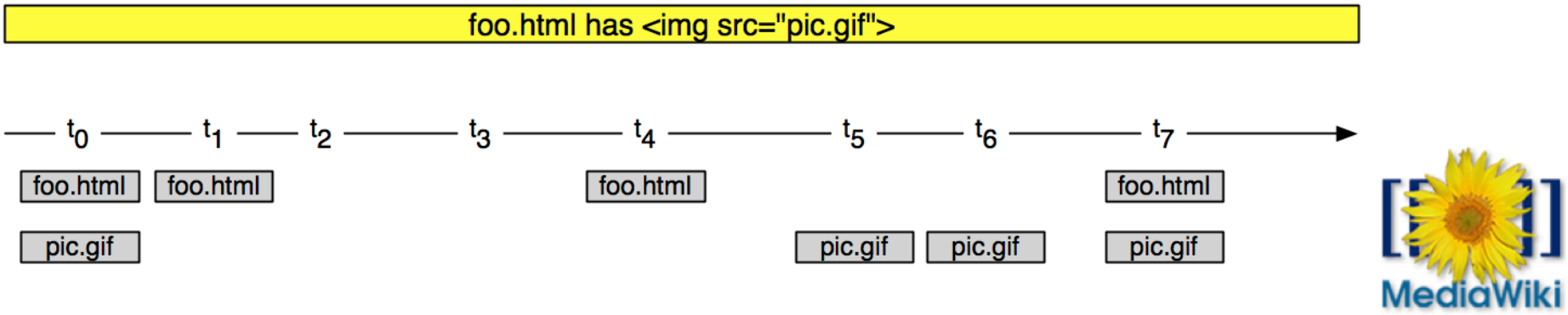
About Memento-Datetime:

Archive Navigation Coherence

Los Alamos NATIONAL LABORATORY

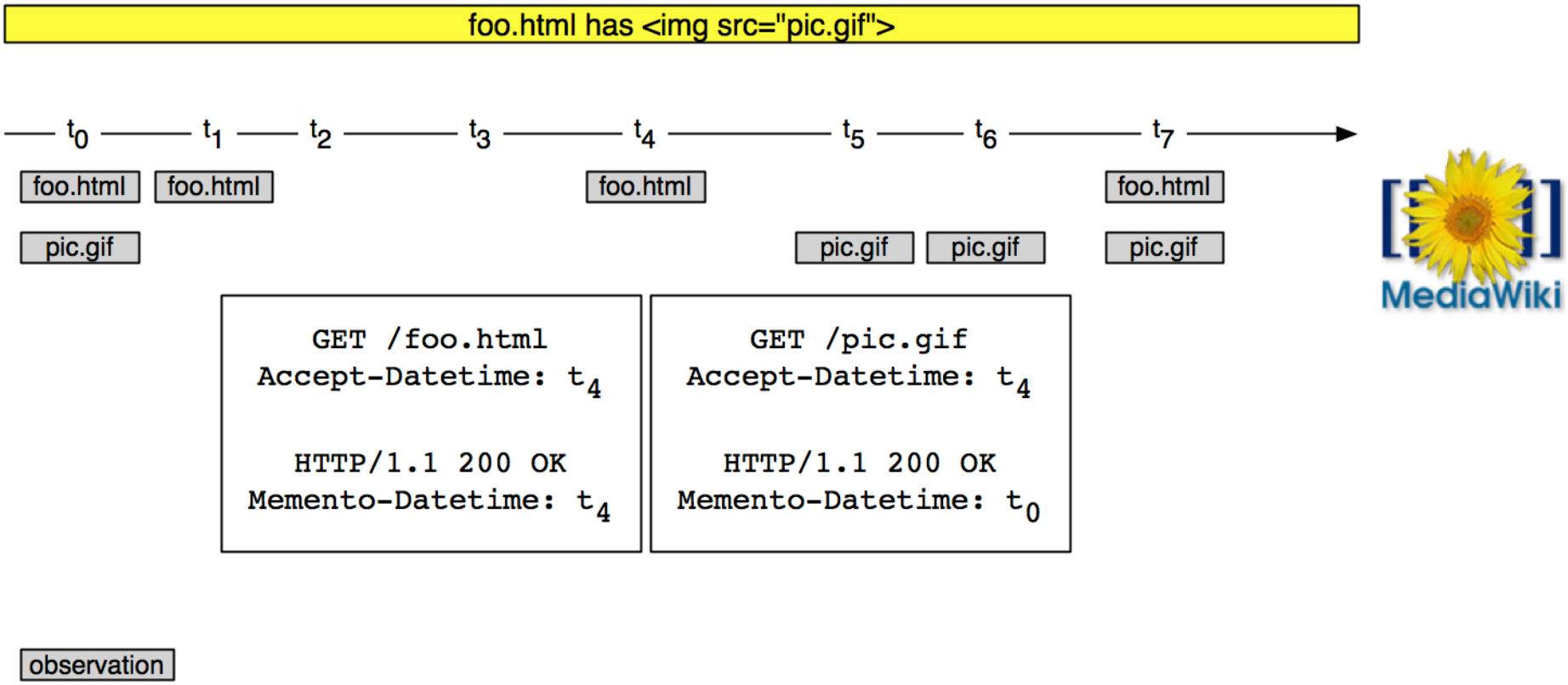
Old Dominion University

Resource History Recorded by CMS and Transactional Archives

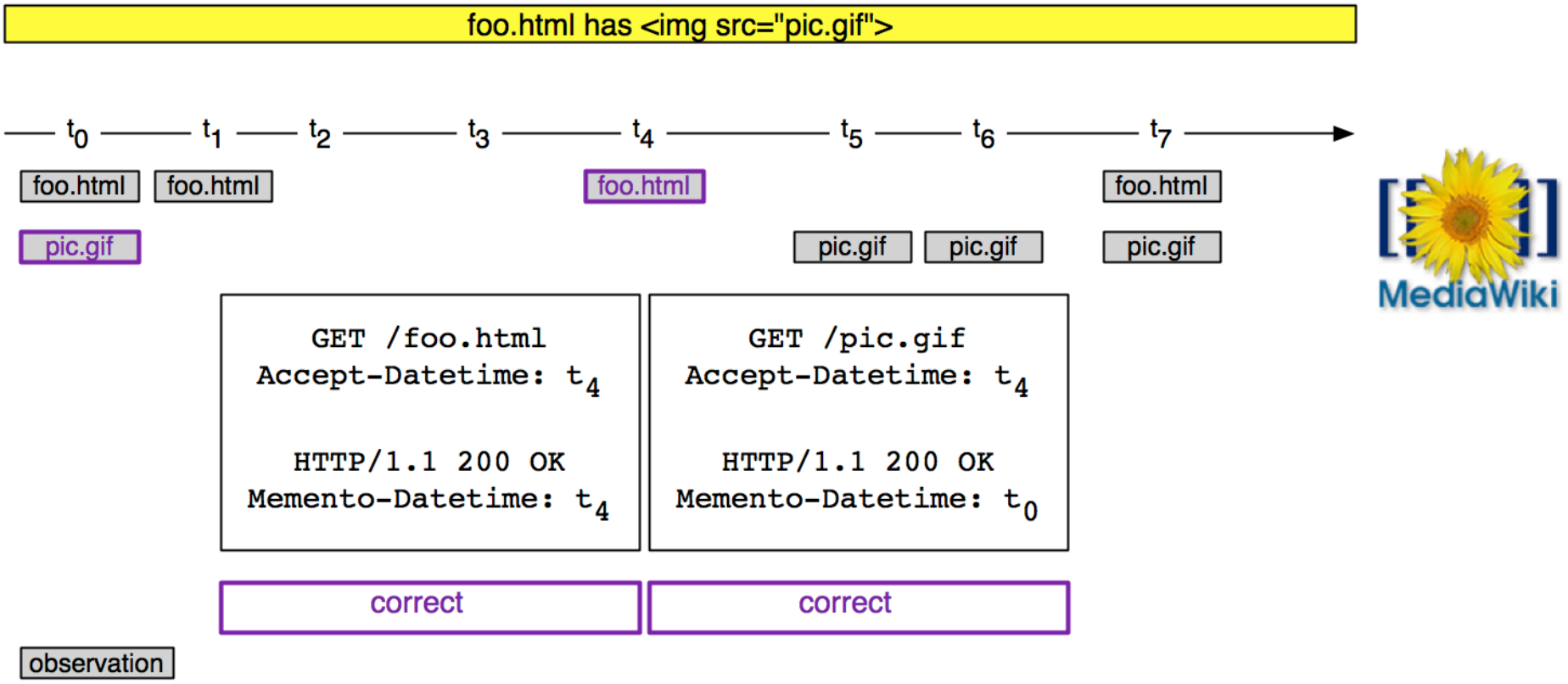


observation

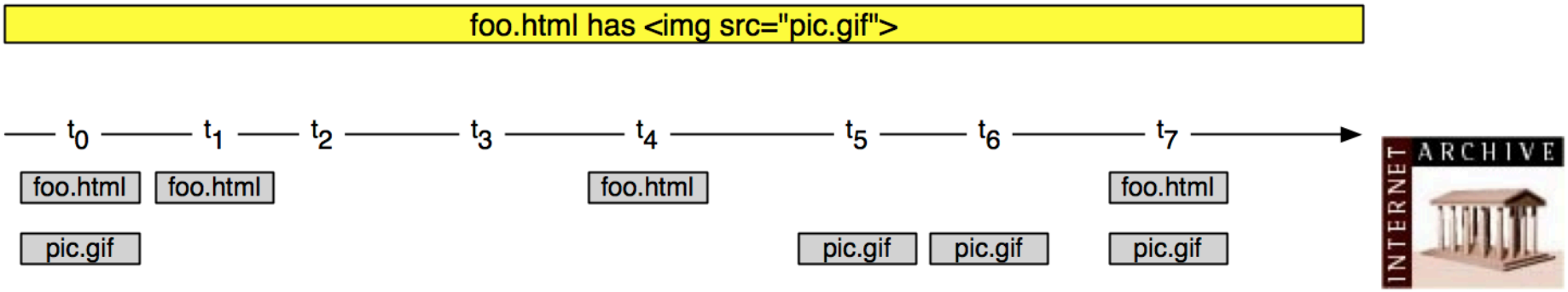
Navigate foo.html @ t4



Navigation Coherence for foo.html @ t4

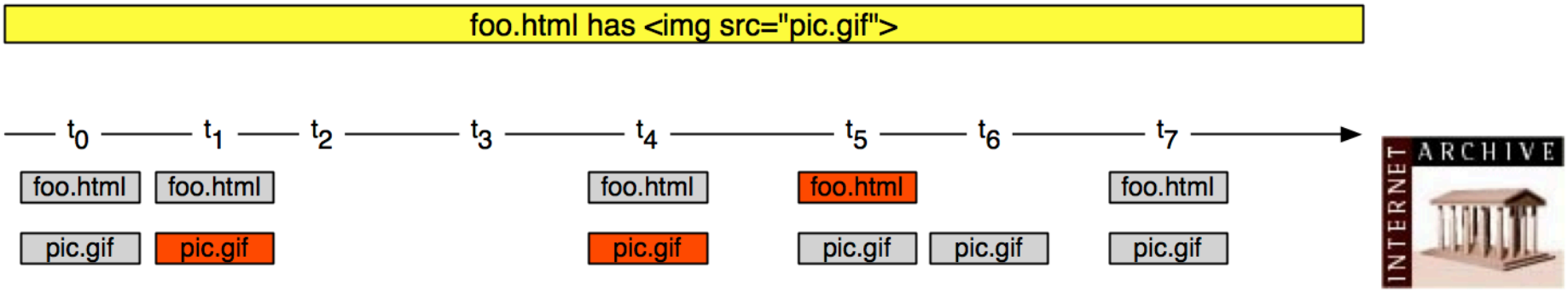


Resource Observations Recorded by Crawler-Based Archives



observation

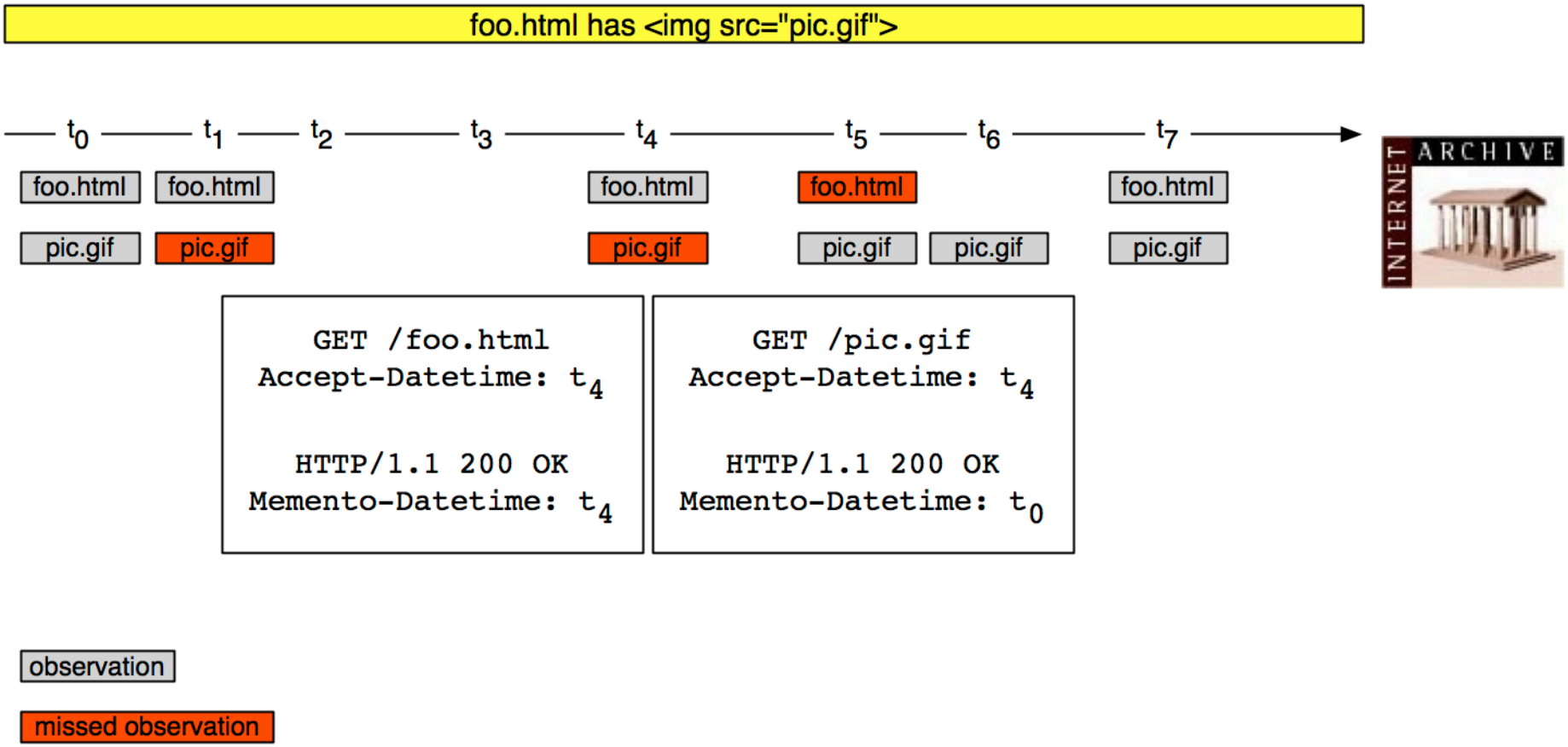
Missed Observations



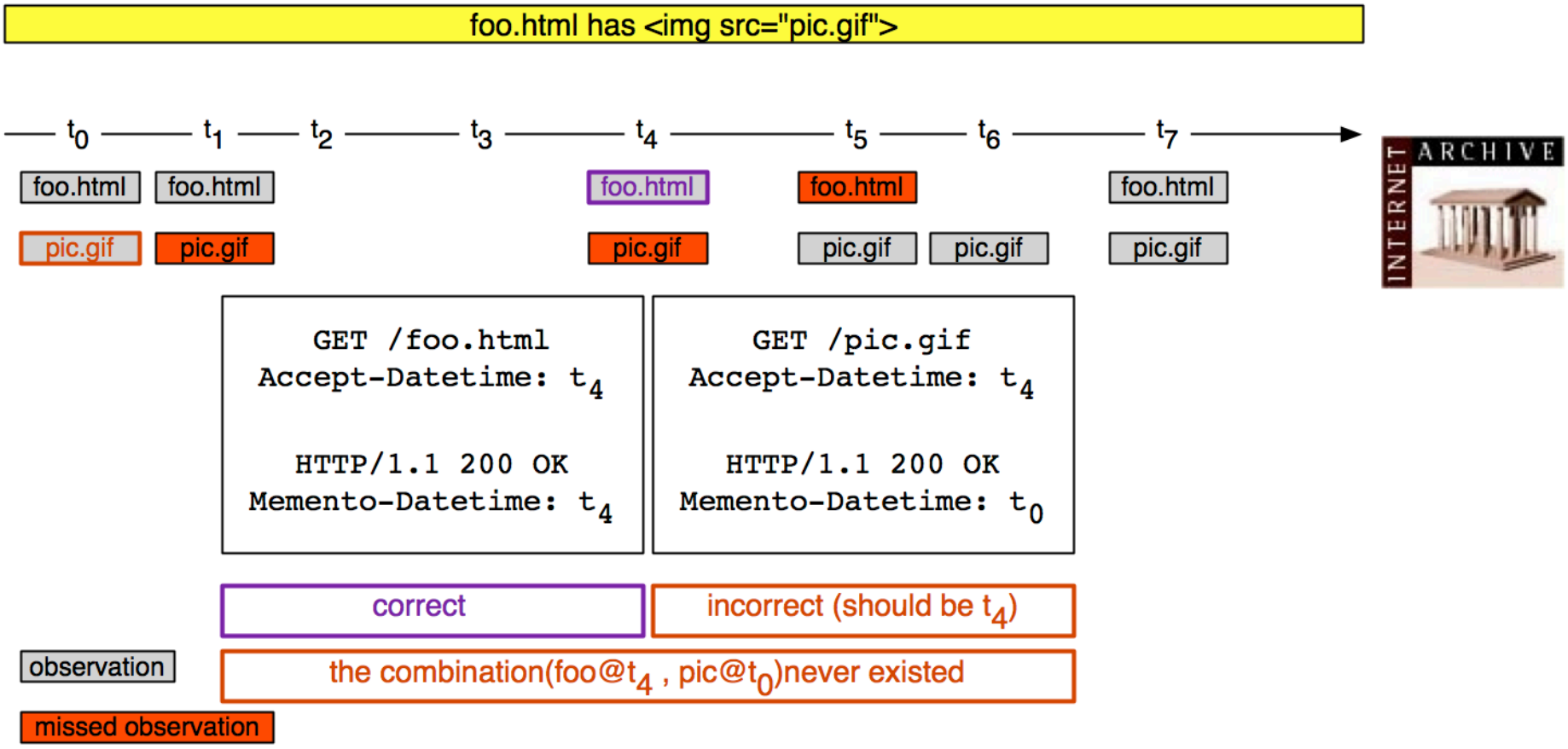
observation

missed observation

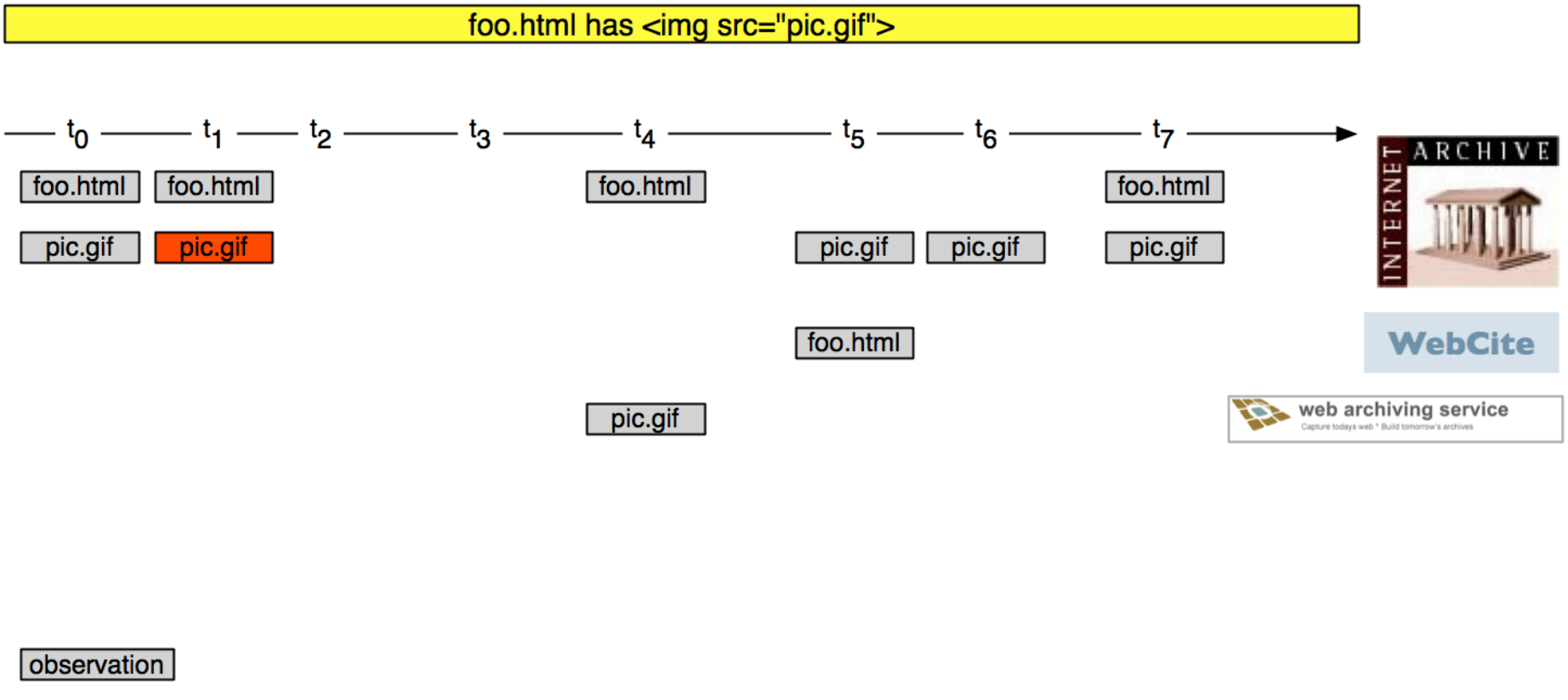
Navigate foo.html @ t4



Navigation Incoherence foo.html @ t4



Increase Coherence with Observations from Multiple Archives?



The Memento Framework:

About Memento-Datetime:

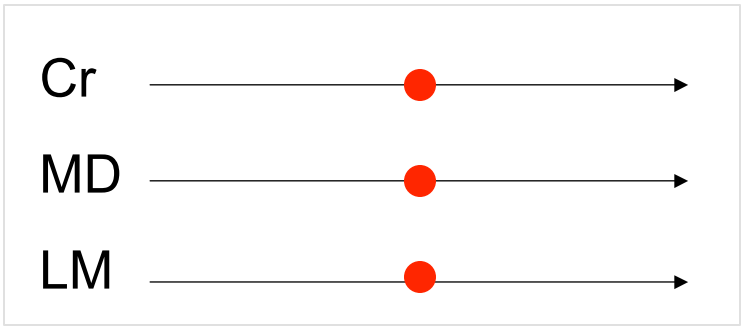
Relation to Creation Datetime and Last-Modified

BY NC SA

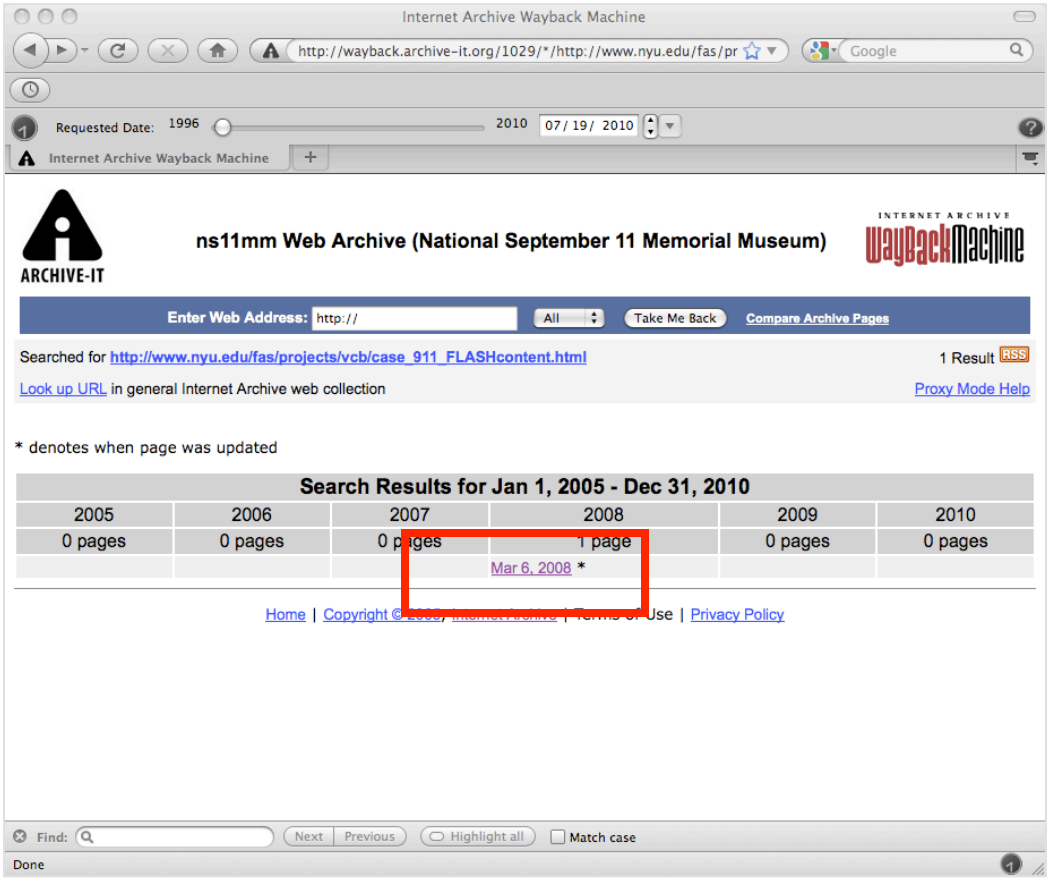
Three Notions of Time

- Creation: Datetime when the resource first came into being
- Last-Modified: Datetime when the resource was last changed
- Memento-Datetime: Datetime that the resource was “frozen”, e.g. as a result of:
  - Archiving it at a different URI (e.g. in a CMS, Web Archive, Transactional Archive, Snapshot Archive);
  - Deciding never to change it anymore and keeping it at its original URI.

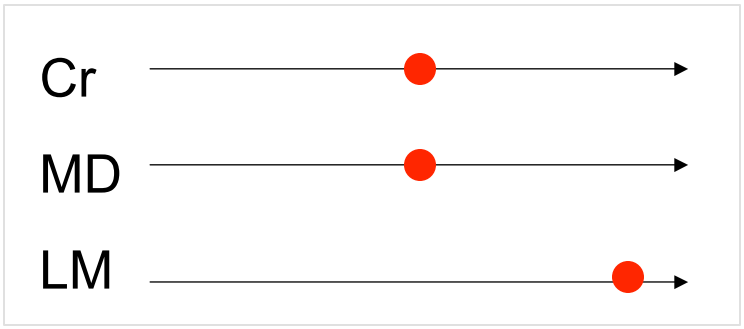
Creation = Memento-Datetime = Last-Modified



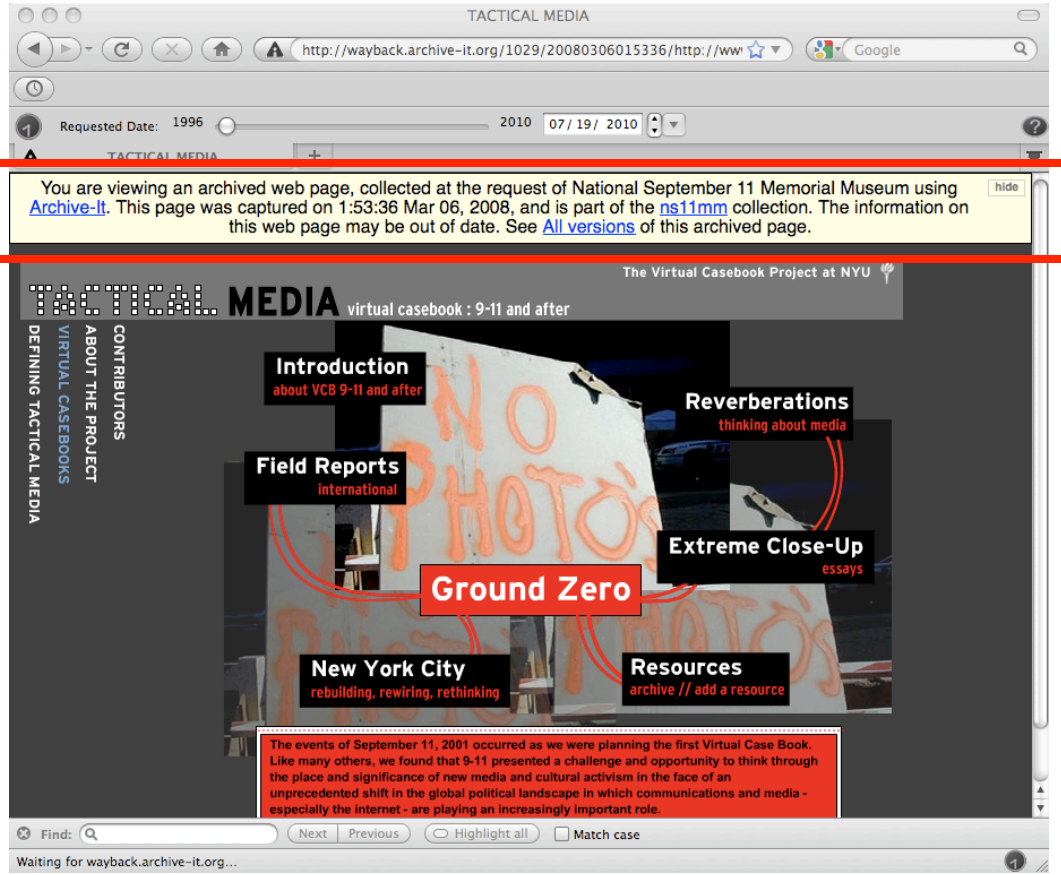
At a particular point in time, the Original Resource is observed, and the associated Memento is created. All time values for the Memento are the same.



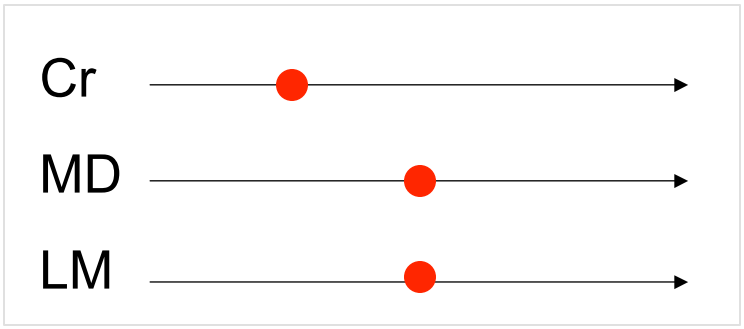
Creation = Memento-Datetime < Last-Modified



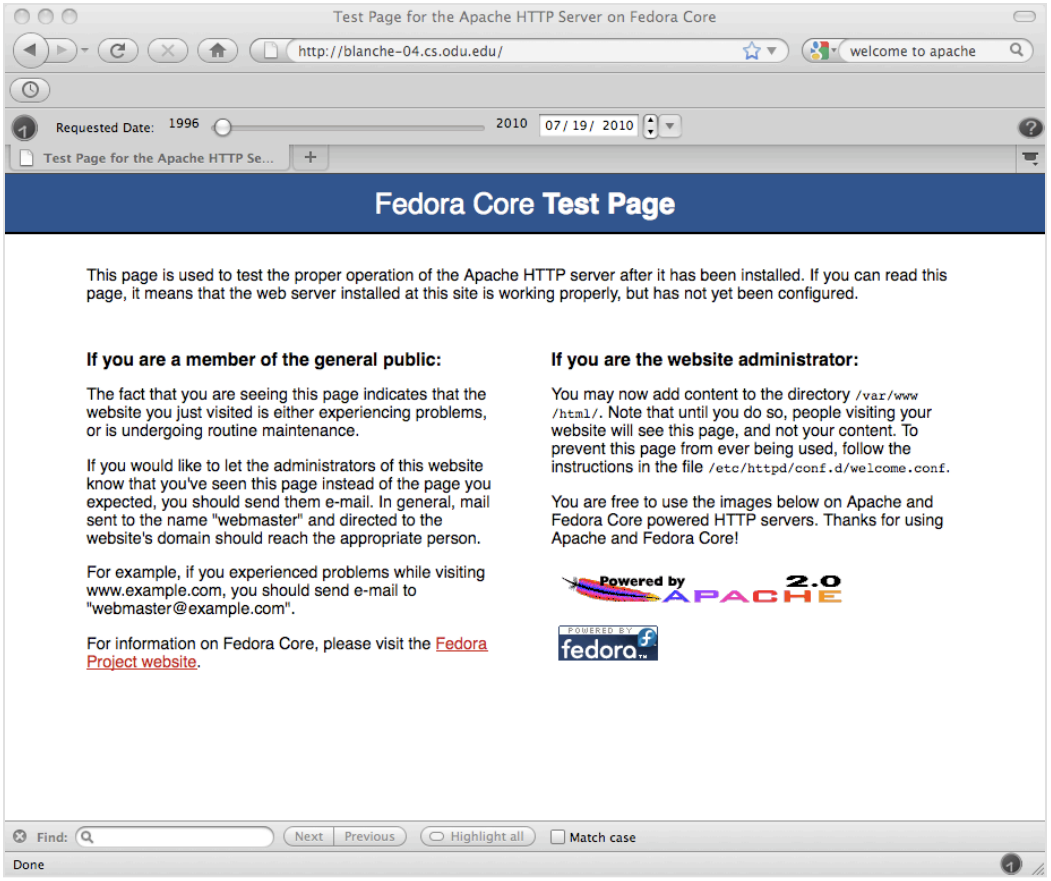
The HTML archive banner added to a Memento necessitates a change in Last-Modified of the Memento, but Creation date and Memento-Datetime remain unchanged.



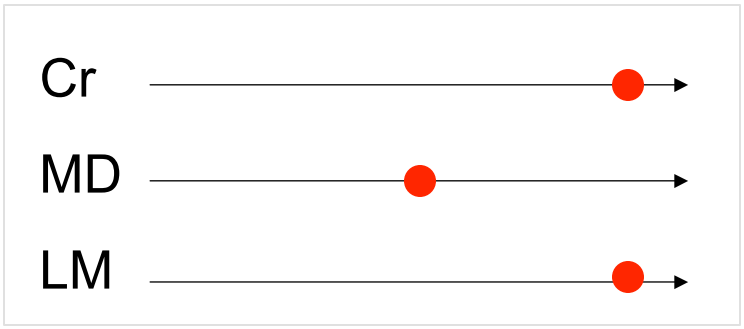
Creation < Memento-Datetime <= Last-Modified



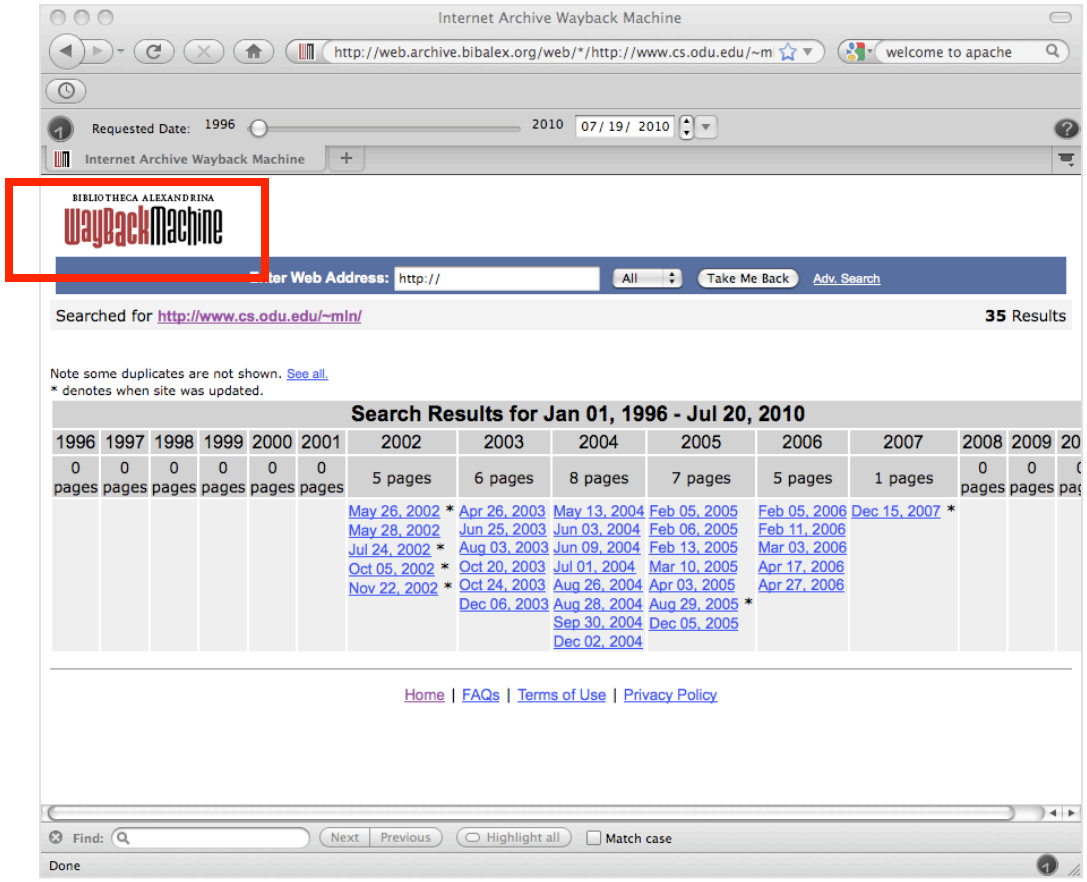
It is possible that the Original Resource was a placeholder resource and returned a 200 response before it started to identify a Memento (URI-R=URI-M).



Memento-Datetime < Creation <= Last-Modified



If a Memento is copied to a new archive, the copied Memento has a Creation and Last-Modified equal to the time of copying. The Memento-Datetime is “sticky” and is the same for the Memento and its copy.



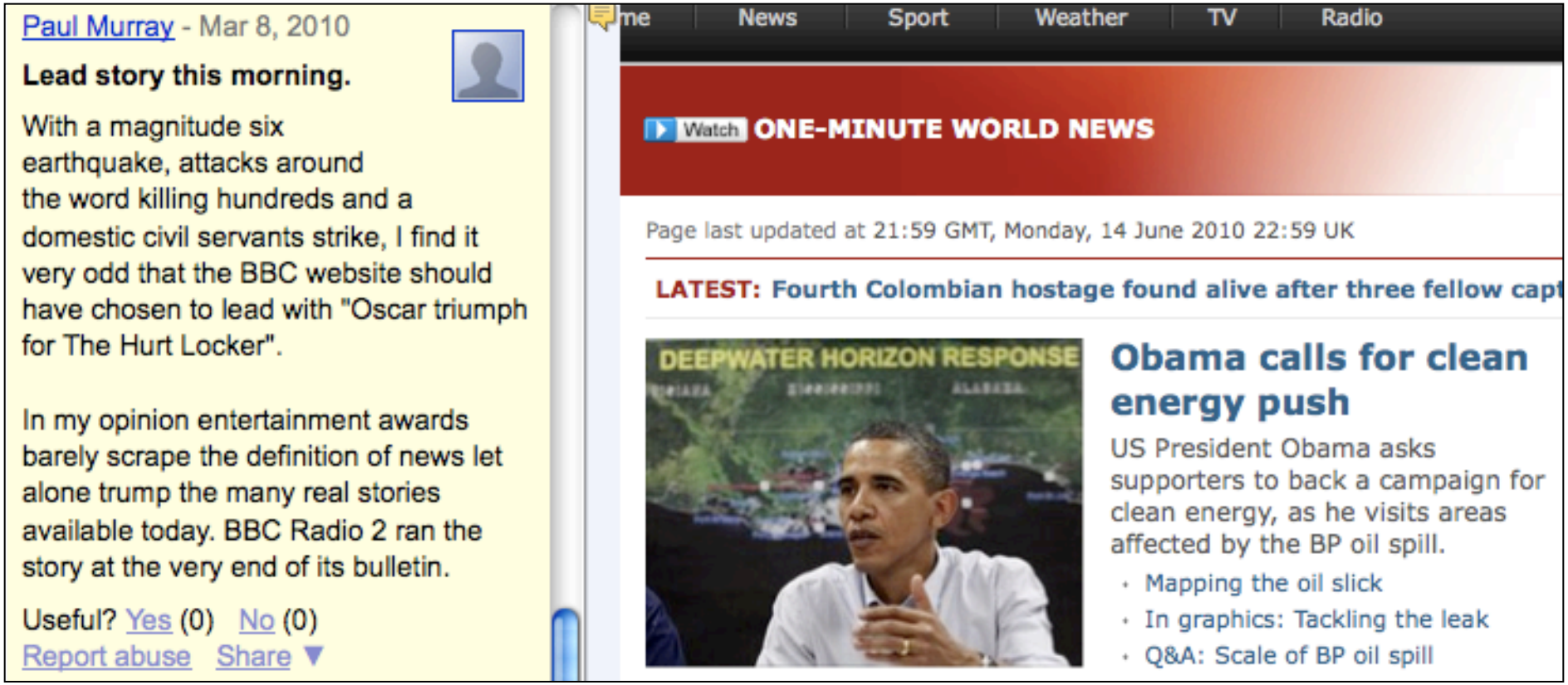
The Memento Framework:

Persistent Web Annotations

Los Alamos NATIONAL LABORATORY

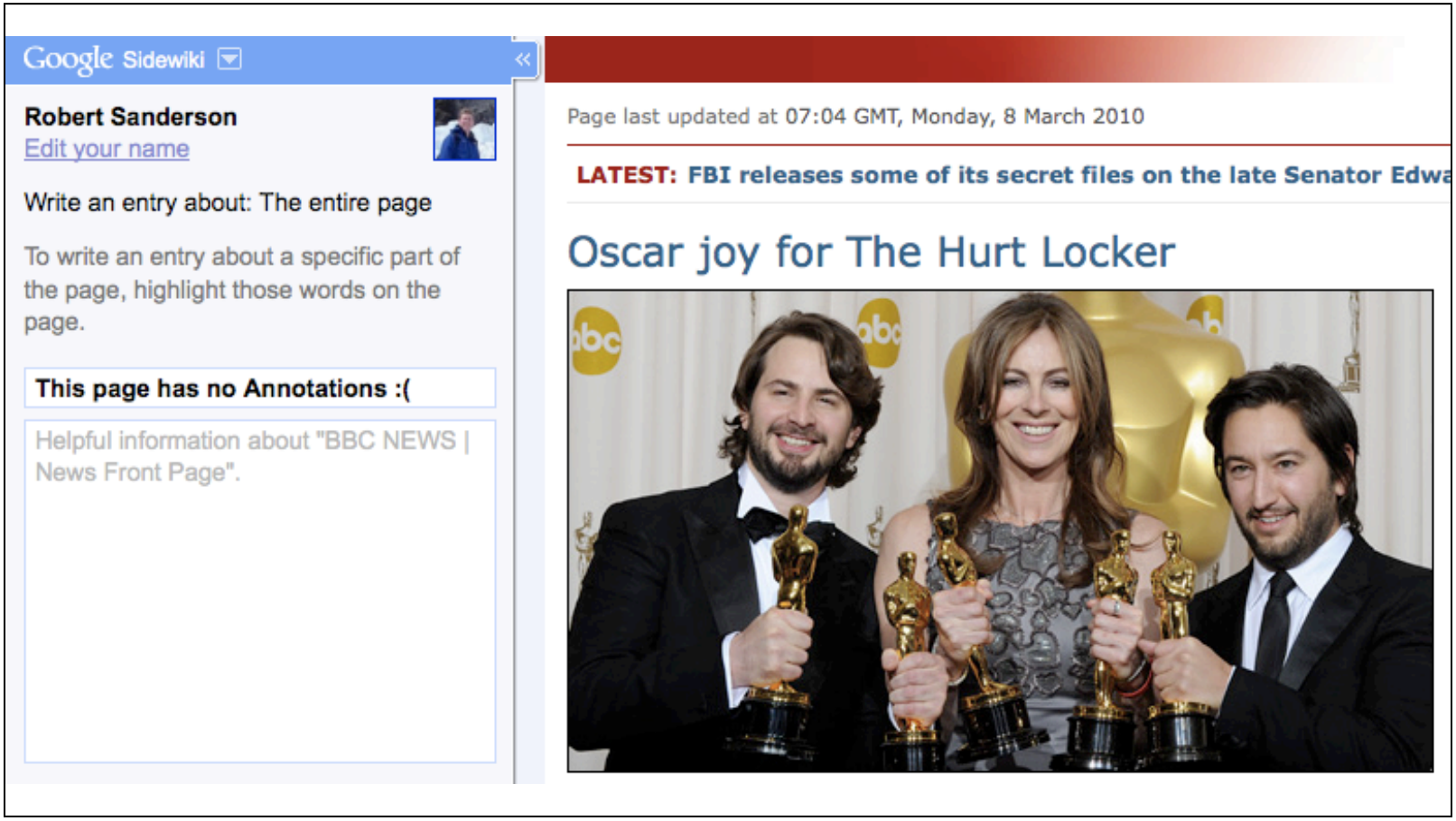
Old Dominion University

Web-Centric Annotation: No Persistence



Google Sidewiki Annotation on *http://news.bbc.co.uk/* as of 2010-06-14

Web-Centric Annotation: No Persistence



Archived page from:
http://www.dracos.co.uk/work/bbc-news-archive/2010/03/08/07.05.html

Web-Centric Annotation: Desired Persistence

Paul Murray - Mar 8, 2010

Lead story this morning.

With a magnitude six earthquake, attacks around the word killing hundreds and a domestic civil servants strike, I find it very odd that the BBC website should have chosen to lead with "Oscar triumph for The Hurt Locker".

In my opinion entertainment awards barely scrape the definition of news let alone trump the many real stories available today. BBC Radio 2 ran the story at the very end of its bulletin.

Useful? Yes (0) No (0)
Report abuse Share ▼

Page last updated at 07:04 GMT, Monday, 8 March 2010

LATEST: FBI releases some of its secret files on the late Senator

Oscar joy for The Hurt Locker

Open Annotation: Dealing with Web Time

• As regular Web resources, Body and Target of an Annotation have representations that can change over time.

• Body and Target can change independently of each other.

• If an Annotation involves resources as they existed at a particular point in time, this needs to be recorded.

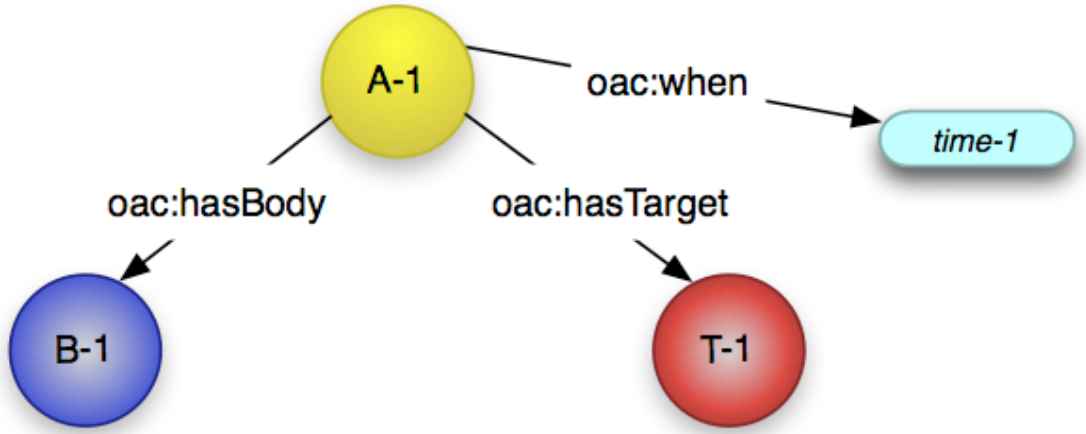
• The OAC model provides hooks for doing so:
  • Timeless Annotations;
  • **Uniform Time Annotations;**
  • Varied Time Annotations.

BY NC SA

Open Annotation: Uniform Time Annotations

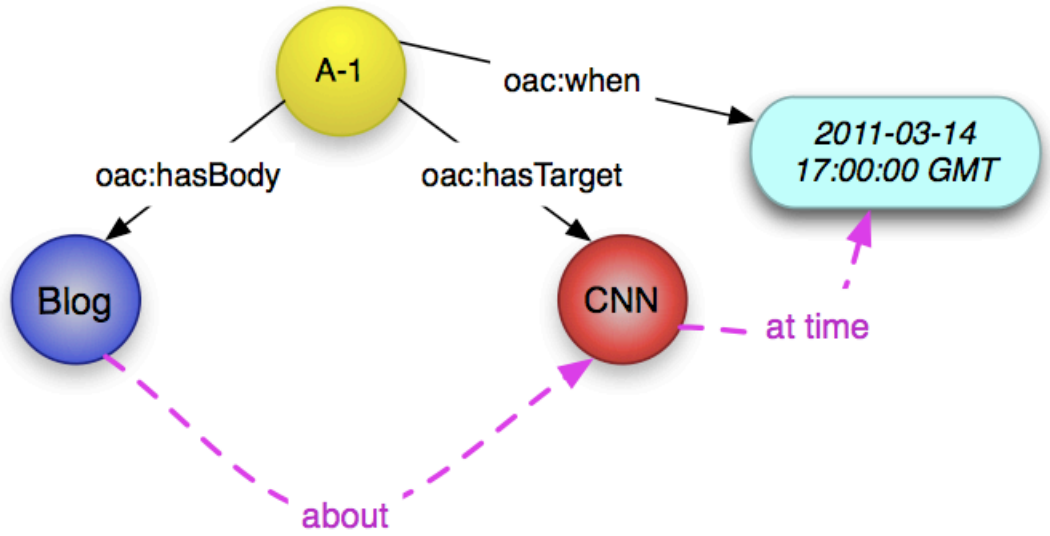
• The Annotation is not always applicable, but pertains to the state of the Body and Target at a specific moment in time.

• Add `oac:when` property to the Annotation.

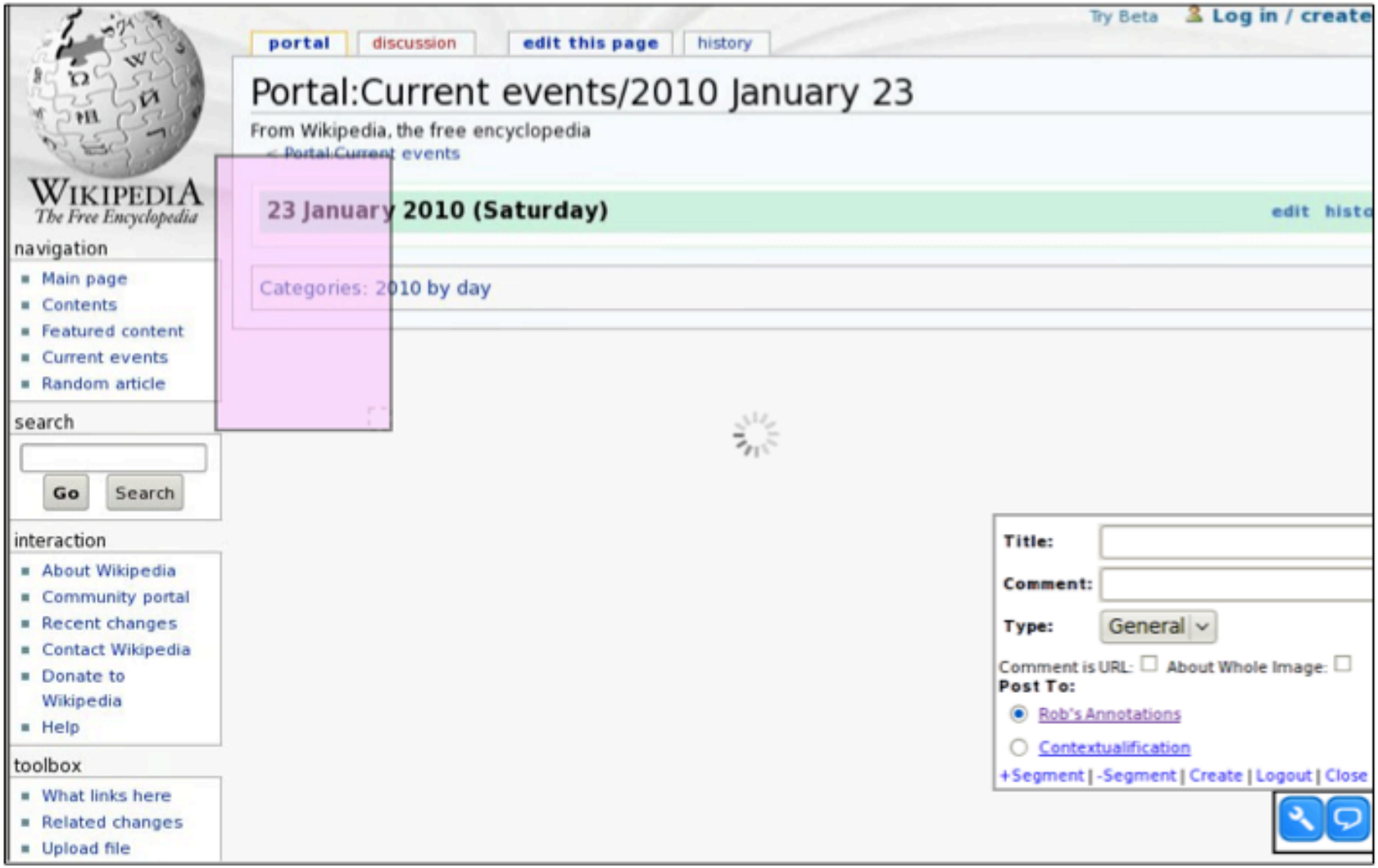


Memento + Open Annotation: Persistent Annotations

• In order to reconstruct the Annotation as intended: Use Memento to obtain an archived representation of B and T as they existed at the `oac:when` datetime.

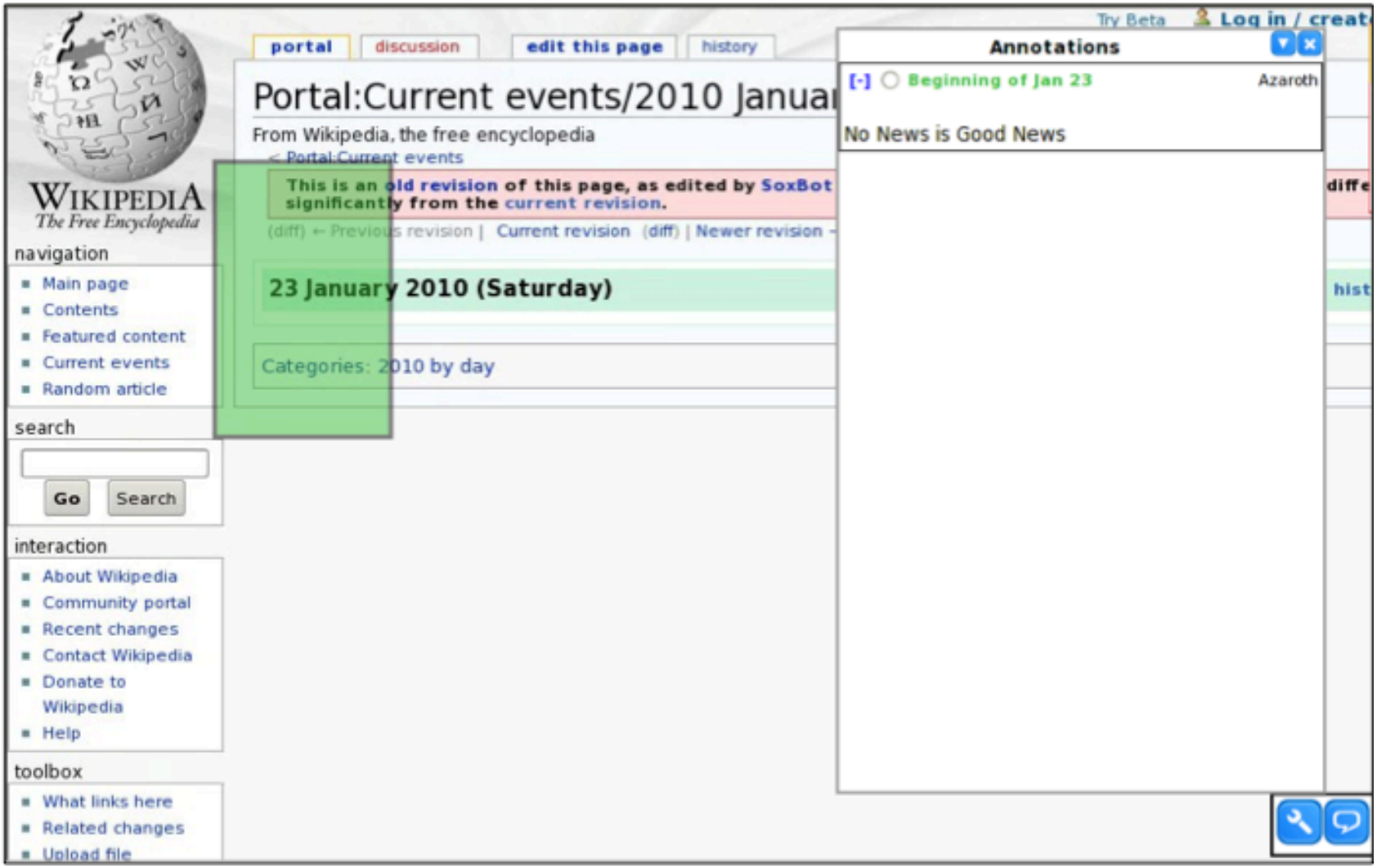


Create an Annotation



Reconstruct the Annotation without Memento

Reconstruct the Annotation with Memento



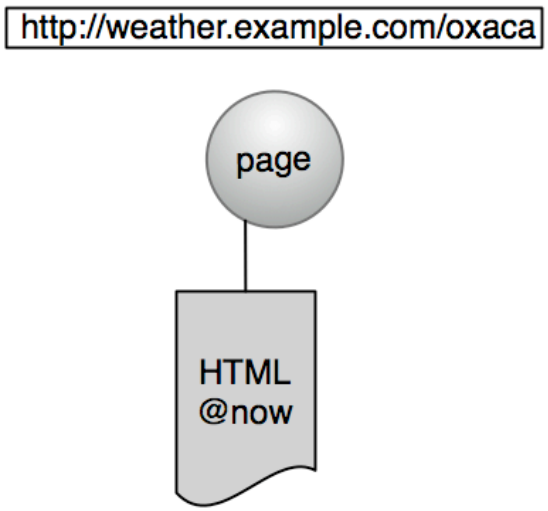
The Memento Framework:

The Increasing Value of a URI

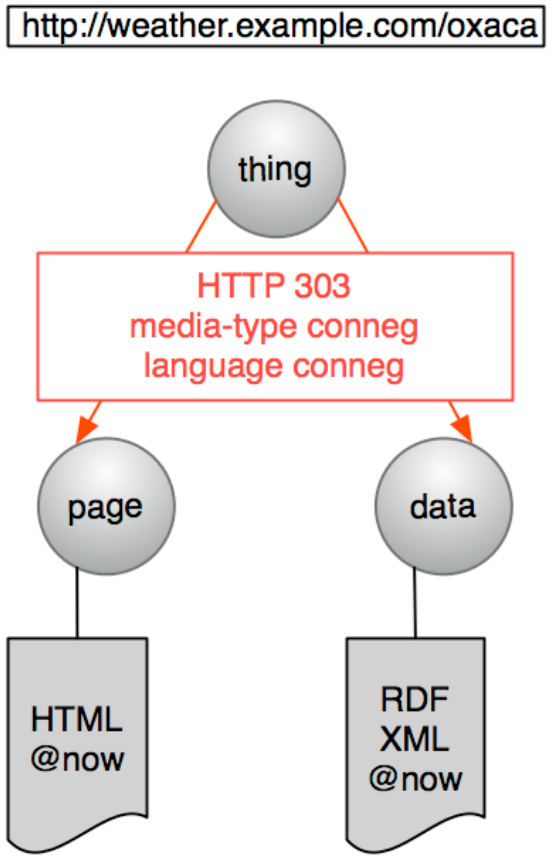
Los Alamos NATIONAL LABORATORY

Old Dominion University

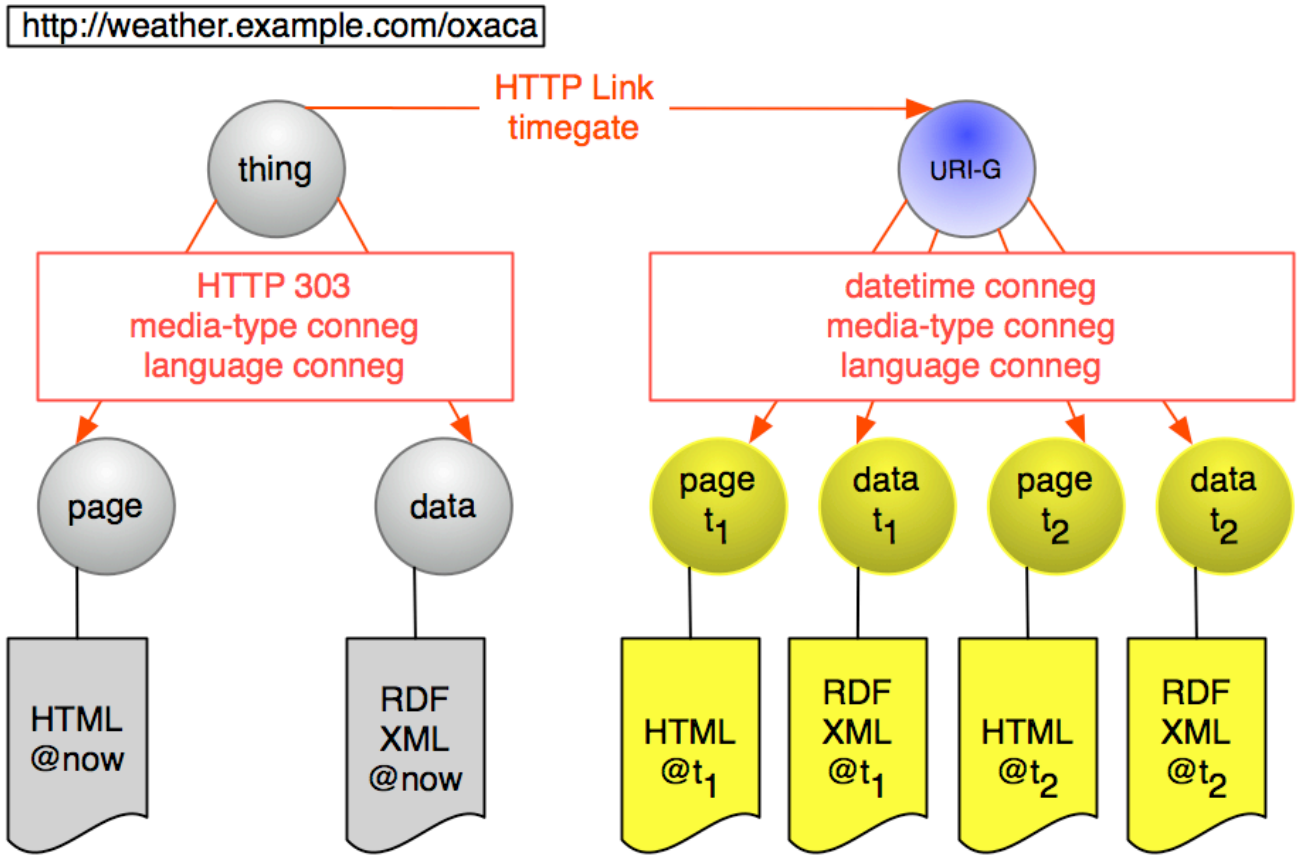
URI as Access Point to a Page



URI as Access Point to Page and Data



URI as Access Point to Current and Past Pages and Data



References

- 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
- 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
- Sanderson, R., Van de Sompel, H. (2011) Open Annotation Alpha3 Data Model Guide. http://www.openannotation.org/spec/alpha3/

Memento wants to make navigating the Web's Past Easy

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