Memento
Accessing Resource Versions Using Datetime Negotiation

Herbert Van de Sompel
Robert Sanderson
Michael L. Nelson

http://mementoweb.org/

Memento is funded by
The Library of Congress

http://www.slideshare.net/hvdsomp/memento-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/
Memento
SWIB 2011, Hamburg, Germany, November 2011

Apple
Today

Select Date
Dec 5 2009

Apple
Dec 5 2009

From
Internet Archive
Memento achieves this by introducing a uniform version access capability to integrate the past and current Web.
Problem Statement …
Resources
Resources have Representations

![Diagram showing the relationship between resources and their representations.]

resources have representations
Resources have Representations that Change over Time
Only the Current Representation is Available from a Resource
Old Representations are Lost Forever
Archived/Version Resources Exist

the old representations are lost forever

well, that is not entirely true
Resource Versions on the Web

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


Versions are Not Integrated with the Web
Architecture of the World Wide Web, Volume One

W3C Recommendation 15 December 2004

This version:
http://www.w3.org/TR/2004/REC-webarch-20041215/
Latest version:
http://www.w3.org/TR/webarch/
Previous version:
http://www.w3.org/TR/2004/PR-webarch-20041105/

Editors:
Ian Jacobs, W3C
Norman Walsh, Sun Microsystems, Inc.

Resource state may evolve over time. Requiring a URI owner to publish a new URI for each change in resource state would lead to a significant number of broken references. For robustness, Web architecture promotes independence between an identifier and the state of the identified resource.
Finding Resource Versions


Finding Resource Versions

Go to
and click History

Browse History
Navigating Resource Versions


Dec 20 2001, 4:51:00 UTC

Pentagon

http://en.wikipedia.org/wiki/The_Pentagon
Navigating Resource Versions


Memento
SWIB 2011, Hamburg, Germany, November 2011
Versions are Not Integrated with the Web

• Functionally:
  • Need to know about archives
  • Need to search each of them individually
  • Cannot consistently navigate in the past within an archive
  • Cannot navigate across archives

• Architecturally:
  • 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
Memento Wants to Integrate the Past and Current Web

cannot navigate the past →

cannot get to the past from the now

time

past Web @ t1
past Web @ t2
past Web @ t3

the current Web
The Memento Framework:

Accessing Resource Versions with and without Memento
Memento
Adding Time to the Web

Memento wants to make it as straightforward to access the Web of the past as it is to access the current Web.

If you know the URI of a Web resource, the technical framework proposed by Memento allows you to see a version of that resource as it existed at some date in the past, by entering that URI in your browser like you always do and by specifying the desired date in a browser plug-in. Or you can actually browse the Web of the past by selecting a date and clicking away. Whatever you land upon will be versions of Web resources as they were around the selected date. Obviously, this will only work if previous versions are available somewhere on the Web. But if they are, and if they are on servers that support the Memento framework, you will get to them.

At this point, there aren't any formal technical specifications detailing the Memento framework but we will get to that. For now, the information on this site should provide quite a good insight into how Memento is trying to change the Web by adding a time dimension to its most common protocol, HTTP. The Guide to Memento is a good entry point.
The Memento Framework:

Protocol to Integrate Past and Current Web

Overview
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
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
HTTP GET on URI-A

Client

GET A HTTP/1.1

Server

HTTP/1.1 200 OK
Content-Type: text/html
Content-Language: en
GET with conneg on URI-T – Server Choice – 302 Found – Step 1

GET T HTTP/1.1
Accept: text/html, application/pdf;q=0.8
Accept-Language: en-US, fr;q=0.7, de;q=0.5

HTTP/1.1 302 Found
TCN: choice
Vary: negotiate, accept, accept-language
Location: A
Content-Type: text/html
Content-Language: en
Alternates: {"A" 1.0 {type text/html}
{langua: en}}, {"B" ...}, {"C" ...}

Client

Server
GET with conneg on URI-T – Server Choice – 302 Found – Step 2

GET A HTTP/1.1
Accept: text/html, application/pdf; q=0.8
Accept-Language: en-US, fr;q=0.7, de;q=0.5

HTTP/1.1 200 OK
Content-Location: A
Content-Type: text/html
Content-Language: en

server

client

T

A

B

C

[HTML]

[PDF]

[PDF]

[Flag US]

[Flag US]

[Flag France]
GET with conneg on URI-T – Server List – 406 Not Acceptable
Original Resources and Mementos
Bridge from Present to Past
Bridge from Past to Present
Memento Framework

Original Resource

Memento TimeGate

Archived & Version Resources

URI-R
- HTTP Link original
- HTTP Link timegate

URI-G
- Content Negotiation Accept-Datetime = Ti
- Content Negotiation Accept-Datetime = Tj

URI-M1
- Apr 10 2001 representation

URI-M2
- Aug 15 2007 representation
Memento Client Server Interaction

**Browser**

---

**Original Resource**

---

**TimeGate**

---

**Memento Archive**

---

**Diagram**

1. **Browser** asks: Do you have a preferred TimeGate?
2. **Original Resource** (R) responds: Yes, G // No, use a default
3. **Browser** asks: Where is the archived copy for the time that I want?
4. **TimeGate** (G) responds: It's at M // I don't know, please try another TimeGate
5. **Browser** asks: Please give me the archived copy
6. **Memento Archive** (M) responds: Here it is
Memento HTTP Flow

HEAD R, Accept-Datet ime

Link ➔ G

GET G, Accept-Datet ime

302 ➔ M, Vary, Link ➔ M,R,T

GET M, Accept-Datet ime

200, Memento-Datet ime, Link ➔ M,R,T,G
The Memento Framework:

Protocol to Integrate Past and Current Web

Interesting Cases
Multiple Archives

HTTP Link timegate

Content Negotiation
Accept-Datetime = Tm

Content Negotiation
Accept-Datetime = Tn

Content Negotiation
Accept-Datetime = Ti

Content Negotiation
Accept-Datetime = Tj

URI-M1

URI-M2

URI-M3

URI-M4

Original Resource

Memento TimeGate

Archived & Version Resources
Original Resource Gone
Original Resource’s Server Gone

Original Resource

Memento TimeGate

Archived & Version Resources

I know a timegate

HTTP GET
Accept-Datetime = T_j

HTTP GET
Accept-Datetime = T_j

URI-M_1

URI-M_2

GONE

GONE

April 10, 2001 representation

August 15, 2007 representation
Original Resource Provides no Link
The Memento Framework:

Protocol to Integrate Past and Current Web

HTTP Headers
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

**Memento-Datetime: Mon, 12 Oct 2009 14:20:33 GMT**
HTTP Response Headers: Vary

• Vary
  o Returned by TimeGate
  o Similar to regular content negotiation
  o 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

HTTP Response Headers: Link

- Link
  - Returned by Original Resource, TimeGate and Mementos
  - Various new Relation Types are introduced:
    - "original"
    - "timegate"
    - "memento"
    - "timemap"
  - Rules about usage of these Links for Original Resources, TimeGates, and Mementos
  - HTTP Link Header: RFC 5988
    See: https://datatracker.ietf.org/doc/rfc5988/

Memento HTTP Flow

HEAD R, Accept-Datetime

Link ➔ G
timegate

GET G, Accept-Datetime

302 ➔ M, Vary, Link ➔ M,R,T
memento,original,timemap

GET M, Accept-Datetime

200, Memento-Datetime, Link ➔ M,R,T,G
memento,original,timemap,timegate
The Memento Framework:

Protocol to Integrate Past and Current Web

HTTP Interactions
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
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
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://arxiv.example.net/web/20000915112826/http://a.example.org>; rel="original",
     <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",
Content-Length: 0
Content-Type: text/plain; charset=UTF-8
Connection: close
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
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/2000915112826/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:

Discovery to Support Integration of Past and Current Web
Discovery to the Memento Framework

• Important:
  o To support discovery of TimeGates when servers of Original Resources do not provide “timegate” Links (which is currently the case for 99.99999 % of Web servers);
  o To allow emergence of aggregating TimeGates that cache metadata about Mementos in multiple archives (avoiding performance issues caused by on-the-fly negotiating with multiple remote TimeGates).

• Approaches to support discovery of:
  o TimeGates;
  o TimeMaps (and hence Mementos and TimeGates);
  o Mementos.
The Memento Framework:
Discovery to Support Integration of Past and Current Web

TimeGate Discovery
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

```plaintext
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
```

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 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/
```
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:

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.

```xml
<http://cnn.com/>;rel="original",
```
TimeMaps: Link Attributes

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

- **Existing Attributes for Links**
  - "rel"  The type of relationship
  - "type"  The (mime) format 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
TimeMaps: Link Attributes

• New Attribute for Mementos:
  • datetime The Memento-Datetime

• Proposed new Attributes for Mementos:
  • license License associated with Memento
  • embargo Time after which Memento is available

<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/.../cnn.com/>;rel="memento";
  license="http://archive.org/license/1";
  datetime="...";embargo="Mon, 20 Jul 2011 00:00:00 GMT",
...
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)
- mcurl (in development)
- Paper in Code4Lib Journal
  http://journal.code4lib.org/articles/4979
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

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Result</strong></td>
<td><strong>URI/Date Returned</strong></td>
</tr>
<tr>
<td>✓ TimeGate returns status 302</td>
<td></td>
</tr>
<tr>
<td>✓ TimeGate has Location header for Memento</td>
<td><a href="http://memento.waybackmachine.org/memento/20100422222407/http://www.ietf.org/">http://memento.waybackmachine.org/memento/20100422222407/http://www.ietf.org/</a></td>
</tr>
<tr>
<td>✓ TimeGate acknowledges Accept-Datetime in Vary header</td>
<td></td>
</tr>
<tr>
<td>✓ Link Header has link with rel=original</td>
<td><a href="http://www.ietf.org/">http://www.ietf.org/</a></td>
</tr>
<tr>
<td>✓ Link Header has link with rel=timebundle</td>
<td><a href="http://memento.waybackmachine.org/lot/timebundle/http://www.ietf.org/">http://memento.waybackmachine.org/lot/timebundle/http://www.ietf.org/</a></td>
</tr>
<tr>
<td>✓ Link Header has link with rel=mmap</td>
<td><a href="http://memento.waybackmachine.org/lot/timemap/link/http://www.ietf.org/">http://memento.waybackmachine.org/lot/timemap/link/http://www.ietf.org/</a></td>
</tr>
<tr>
<td>✓ TimeMap Link has application/link-format type</td>
<td></td>
</tr>
<tr>
<td>✓ Link Header has one or more links with rel=memento</td>
<td><a href="http://memento.waybackmachine.org/memento/20100219084310/http://www.ietf.org/">http://memento.waybackmachine.org/memento/20100219084310/http://www.ietf.org/</a></td>
</tr>
<tr>
<td>✓ Selected Memento in Link Header with rel=memento</td>
<td></td>
</tr>
<tr>
<td>✓ Memento has parsable datetime attribute</td>
<td>Fri, 18 Feb 2010 08:43:10 GMT</td>
</tr>
<tr>
<td>✓ Memento has parsable datetime attribute</td>
<td>Wed, 05 Nov 1996 11:49:54 GMT</td>
</tr>
</tbody>
</table>

- **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
The Memento Framework:

Resource Versioning
current version @t₀

URI-R

ₜ₀ ———— ₜ₁ ———— ₜ₂ — 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
Memento Framework

Original Resource: http://dbpedia.org/resource/France
Time-Series Analysis across DBpedia Versions

Data collected through HTTP Navigation

Depending on implementation of an archival solution, path [1] or [2] is available.
[1] is a replay of an archived 303 that takes the client back to an information resource on the current web,
where the Memento cycle can be repeated. [2] is a 303 directly to the appropriate Memento that has the
description of the Real World Object for the desired point in time.
The Memento Framework:

Persistent Web Annotations
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 ▼
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.
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
URI as Access Point to a Page

http://weather.example.com/oxaca

page

HTML @now
URI as Access Point to Page and Data

http://weather.example.com/oxaca

thing

HTTP 303
media-type conneg
language conneg

page
data

HTML @now
RDF XML @now
URI as Access Point to Current and Past Pages and Data

http://weather.example.com/oxaca

HTTP Link
timegate

HTTP 303
media-type conneg
language conneg

page
data

HTML
@now
RDF
XML
@now

page
data
page
data

HTML
@t1
RDF
XML
@t1
HTML
@t2
RDF
XML
@t2
The Memento Framework:

Transactional Archives
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, …).
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
- Check out http://theresourcedepot.org
References


Memento wants to make navigating the Web’s Past Easy

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