The Memento Infrastructure to Support Research Using Web Archive Collections

http://mementoweb.org/

Martin Klein
Los Alamos National Laboratory
@mart1nkle1n
https://orcid.org/0000-0003-0130-2097

Herbert Van de Sompel
Los Alamos National Laboratory
@hvdsomp
https://orcid.org/0000-0002-0715-6126
Past Web Archive Landscape
Slide from Michael L. Nelson:
https://www.slideshare.net/phonedude/weaponized-web-archives-provenance-laundering-of-short-order-evidence
Current Web Archive Landscape

Perma.cc

Library of Congress

archive.is

archive.is - webpage capture

The National Archives

STANFORD UNIVERSITY LIBRARIES

ARQUIVO.PT

Vefsafn.is

Los Alamos National Laboratory

Fedora Commons

MediaWiki

WordPress

unicode-math

nli

UK Web Archive

cc

WebCite

@mart1nkle1n @hvdsomp

RESAW Workshop 2018, Porto, Portugal, 13 Sep 2018
Status Quo on Web Archive-Based Research

“These URLs were checked against the Internet Archive … “

“… few encountered links were actually available in the Internet Archive”

“We illustrate the challenges using data extracted from .... a collection of all Web pages from the … top level domain crawled by the Internet Archive …”

Fortunately the Internet Archive appear to have a good record of the account...
Getting ... is important because it continually changes, and the Internet Archive’s crawler will only get the initial page of results.
Merit of Using Multiple Web Archives

AISum et al.
“Profiling web archive coverage for top-level domain and content language”
https://doi.org/10.1007/s00799-014-0118-y

Lulwah M. Alkwai et al.
“Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages”
https://doi.org/10.1145/3041656
Memento

http://mementoweb.org/

#memento
Memento

2 components to Memento

1. Retrieve archival snapshot of URI-R at time t
   - Datetime negotiation

2. Retrieve a list of all archival snapshots of URI-R - TimeMap
   - Version history

• Responses to the above reflect perspective of who you ask
  • Individual web archive (IA, Arquivo.pt, etc)
  • Individual resource versioning system (W3C Wiki, Specs)
  • Aggregator (all Memento-compliant web archives)
Memento Implementations for Humans 1/2

http://timetravel.mementoweb.org/
http://timetravel.mementoweb.org/list/20171210034550/https://buddah.projects.history.ac.uk/
Memento Implementations for Humans 2/2

Memento Implementations for Humans 2/2

Memento Implementations for Humans 2/2

Memento Implementations for Humans 2/2

Memento Implementations for Humans 2/2

Memento for Machines

• Request the best Memento from a compliant web archive


• Request a TimeMap from a compliant web archive

  curl http://compliant.archive/timemap/<URI-R>

• Request a TimeMap from the Memento Aggregator

  curl http://labs.mementoweb.org/timemap/format/<URI-R>
Memento APIs for Machines

• Redirect to best Memento
  
  http://timetravel.mementoweb.org/memento/YYYY<MM|DD|HH|MM|SS>/URI

• Provide a JSON description of a Memento
  
  http://timetravel.mementoweb.org/api/json/YYYY<MM|DD|HH|MM|SS>/URI

Overview of APIs & archives’ endpoints:
http://timetravel.mementoweb.org/guide/api/
http://labs.mementoweb.org/aggregator_config/archivelist.xml

Overview of related tools:
https://github.com/machawk1/awesome-memento
• Assess “Content Drift” in scholarly communication

• How much has a web resource that was referenced in a scholarly article changed since the publication of the article?

• We don’t know what “was there” at the time of publication. We only know what “is there” now.

→ Devised a novel approach to assess content drift based on available Mementos

S. Jones et al. “Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content” PLOS ONE 2017 https://doi.org/10.1371/journal.pone.0167475
Novel Approach to Assess Content Drift

publication date of URI referencing paper

Memento Pre of referenced URI
Step 1: Find Mementos

- $t-1$
- $t$
- $t+1$

Memento Pre of referenced URI

Memento Post of referenced URI

Publication date of URI referencing paper
Step 2: Select Representative Mementos

- Compare for textual similarity
- Publication date of URI referencing paper
- Memento Pre of referenced URI
- Memento Post of referenced URI
Step 2: Select Representative Mementos

- Textually similar
- Publication date of URI referencing paper
- Representative Memento Pre of referenced URI
- Representative Memento Post of referenced URI
Step 3: Dereference Live Web Version of URI
Step 4: Representative Memento vs. Live Version

- **Representative Memento Pre of referenced URI**
- **Representative Memento Post of referenced URI**
- **Live Web referenced URI**

**Comparison:**
- **Publication date of URI referencing paper**
- **Consultation date of URI referencing paper**

**Textual Similarity:**
- Compare for textual similarity
- **Textually similar**
Content Drift & Link Rot Over Time - arXiv
Questions asked:

- Can we create event collections by focused crawling online-available web archives?

- How do event collections created from the archived web compare to those created from the live web?

- How does the amount of time passed since the event affect the collections built from the live and the archived web?

- How do event collections built from the archived web compare to manually curated collections?
Experiment & Methodology

- Topics limited to terror attacks and mass shootings in the U.S.
- From different times in the past
- Focused crawl of:
  a) 22 archives, simultaneously, via Memento aggregator infrastructure
  b) the live web
- Take content and temporal relevance into account, equally weighted
- Use events’ Wikipedia page - around the time of the event - and its references as input for focused crawler
  - URIs as seeds
  - Content for relevance assessment

TUC, 01/08/2011 – URIs per Level

Web Archive Crawl

Live Web Crawl

M. Klein et al. “Focused Crawl of Web Archives to Build Event Collections” in WebSci 2018
https://doi.org/10.1145/3201064.3201085
TUC, 01/08/2011 – Web Archive Contributions

- web.archive.org 75%
- wayback.archive-it.org 14%
- webarchive.loc.gov 7%
- web.archive.bibalex.org 2%
- archive.is 2%

M. Klein et al. “Focused Crawl of Web Archives to Build Event Collections” in WebSci 2018
https://doi.org/10.1145/3201064.3201085
Limitations

• Memento is about URI and datetime
  • No search and TDM across archives

• Dark archives & personal archives commonly not accessible
  • Technical and political issues

• NZL, AUS, WebCitation, Wikipedia not Memento-compliant
  (we have the tools)
The Memento Infrastructure to Support Research Using Web Archive Collections

http://mementoweb.org/

Martin Klein
Los Alamos National Laboratory
@mart1nkle1n
https://orcid.org/0000-0003-0130-2097

Herbert Van de Sompel
Los Alamos National Laboratory
@hvdsomp
https://orcid.org/0000-0002-0715-6126