A Framework for Aggregating Private and Public Web Archives

Mat Kelly
Old Dominion University
Web Science & Digital Libraries Research Group
Department of Computer Science
Norfolk, Virginia USA
mkelly@cs.odu.edu
The Web is Ephemeral
Web Archives to the Rescue: Typical Access

1. Go to archive.org in your browser
2. Enter the URL you want to see in the past in the form field
3. Submit your query
Web Archives to the Rescue: Typical Access

4. Locate the capture on the calendar or histogram view

5. Select the year/capture for the day

6. Repeat until you find the closest date and time
7. Finally, view the capture from March 9, 2013
Web Archiving - Live Web psu.edu
Web Archiving - Archival Capture
Web Archiving

Associate live Web URIs

With their archived representations
Web Archives provides access to the Web that was

What did psu.edu look like in the past?
Multiple archival efforts (3 of many)
More archives produces a more comprehensive picture
Even then, not everything is preserved

What did obscuresite.com look like in the past?

0 captures for obscuresite.com
User sees on live Web may not be what is captured

What did facebook.com look like in the past?
...And oftentimes that is for the best

Have you preserved my online banking?
Other times, we may want our content archived

Have you archived my daughter’s flickr.com photos?

0 captures for that URI
...especially when it has disappeared

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly
“Save this, but only for me.”

- **Screenshots** of Web pages are insufficient
  - Not interactive/representative, do not integrate, lose context otherwise provided in metadata
- Large-scale archives’ tools are open source
- Individuals can archive, but there are still technical barriers
Individuals, Too, Can Archive The Web

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly
Captures from Institutional and Personal Sources

at $t_A$

at $t_C$

at $t_{D-Z}$
Memento Facilitates this Aggregation

RFC7089
Today’s Memento Aggregation

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly
Today’s Memento Aggregation
Desire: Include Personal Archives
Desire: Include Other Non-Aggregated Archives
Rapidly Changing Pages May Not Be Comprehensively Captured

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly
Archiving More Archives Provides a Better Picture of the Web
Research Questions

**RQ1:** What sort of content is difficult to capture and replay for preservation from the perspective of a Web browser?

**RQ2:** How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?

**RQ3:** What issues exist for capturing and replaying content behind authentication?

**RQ4:** How can content that was captured behind authentication signal to Web archive replay systems that it requires special handling?

**RQ5:** How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?

**RQ6:** What kinds of access control do users who create private Web archives need to regulate access to their archives?
Research Questions

**RQ1:** What sort of content is difficult to capture and replay for preservation from the perspective of a Web browser?

**RQ2:** How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?

**RQ3:** What issues exist for capturing and replaying content behind authentication?

**RQ4:** How can content that was captured behind authentication signal to Web archive replay systems that it requires special handling?

**RQ5:** How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?

**RQ6:** What kinds of access control do users who create private Web archives need to regulate access to their archives?
Outline

- Introduction/Motivation
- Background
- Preliminary Research
- Proposed Framework
- Evaluation Plan
Outline

- Introduction/Motivation
- Background
- Preliminary Research
- Proposed Framework
- Evaluation Plan
Needed Association of Live-to-Archived Web

- Archived captures (mementos) are “aware” of what they represent from the live Web
- ...but we don't have a good way to establish the bidirectional linkage
Representations can be **Linked** in time
Background: Memento


Background: Memento Request Example

HTTP Request
- **Accept-Datetime:** Wed, 02 Aug 2017 23:15:00 GMT
- **GET:** http://web.archive.org/web/http://www.cnn.com

Request [cnn.com](http://www.cnn.com) at Sept 11, 2001 at 9am EST
Background: Memento Request Example

HTTP Request
- **Accept-Datetime**: Wed, 02 Aug 2017 23:15:00 GMT

Request cnn.com at Sept 11, 2001 at 9am EST

HTTP Response (302)
- **Memento-Datetime**: Wed, 02 Aug 2017 23:18:04 GMT
- **Link**: 

URI-T - timemap
URI-R - original
URI-G - timegate
URI-M - memento
Background: Dereferencing a TimeMap at URI-T

- Date-based pagination
- Other formats for TimeMap

Request URI-T
Role-based delegation and authentication

A familiar paradigm used for authentication on the live Web
Web users question trusting institutions to preserve private Web contents\(^1\)

OAuth 2.0\(^2\) facilitates authentication cohesion of entities

---

**RQ3:** What issues exist for capturing and replaying content behind authentication?

**RQ6:** What kinds of access control do users who create private Web archives need to regulate access to their archives?

---


HTTP Prefer

- HTTP negotiation already available via Accept-* headers
- *Prefer* syntax provide mechanism for client to specify preferences
  - ...with which servers may not comply

Memento Aggregation State of the Art
Memento Aggregation - MementoWeb

 Also available via CLI:

 $ curl http://timetravel.mementoweb.org/timemap/link/http://odu.edu
Memento Aggregation - MemGator

- Open Source Memento Aggregator - [github.com/oduwsdl/memgator](https://github.com/oduwsdl/memgator)
- Easy personal/local deployment
- Specify archive list on launch
  - Easily configurable JSON
  - Use default collection if not specified
- TimeMap Formats:
  - Link
  - JSON
  - CDXJ

CDXJ: An Alternative TimeMap Format

Link, CDXJ, and JSON TimeMaps: Multiple formats to express same information

Link syntax is not expandable: They were meant to be displayed in a constrained environment (in HTTP Link headers)

Bridges gap between live and archived Webs

Leverages Memento aggregator’s capability, returns TimeMaps

Indicates # of captures for a URI while you browse

Provides navigation of mementos while browsing live Web

Single-click submission of URI-R to multiple Web archives

Outline

- Introduction/Motivation
- Background
- Preliminary Research
- Proposed Framework
- Evaluation Plan
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly

Mat Kelly
Department of Computer Science
Old Dominion University
Norfolk, Virginia
m.kelly@cs.odu.edu

Michelo C. Weigle
Department of Computer Science
Old Dominion University
Norfolk, Virginia
m.weigle@cs.odu.edu

ABSTRACT

The Internet Archive’s Wayback Machine is the most common way that typical users interact with web archives. The Internet Archive uses the Heritrix web crawler to transform pages on the publicly available web into Web Archival Record (WARC) files, which can then be accessed using the Wayback Machine. Because Heritrix can only access the publicly available web, many personal pages (e.g., password-protected pages, social media pages) cannot be easily archived in a standard WARC format. We have created a Google Chrome extension for downloading a WARC file from any webpage. Using this tool, content that might have been otherwise lost in time can be archived in a standard format by any user. This tool provides a way for casual users to effortlessly preserve material that might be lost if the standard WARC format is extended to include “long term storage, maintenance, and access of personal critical assets that have emotional, intellectual, and historical value to individuals” [3].

RQ1: What sort of content is difficult to capture and replay for preservation from the perspective of a Web browser?

RQ2: How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?

Copyright is held by the author(s).
JWIC'12, June 10-14, 2012, Washington, DC, USA.
ACM 978-1-4503-1144-1/12/06.

WarCreate - Create Wayback-Consumable WARC Files from Any Webpage

Preserve everything you see!
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

ABSTRACT

We describe Mink, a new web browser extension that provides a different model for integration of the live and archived web. While a user browses the live web, Mink actively queries the archives and reports other instances of the page in the archives without requiring active querying by the user. Further, by querying the archives dynamically and asynchronously, a user can view the extent to which the currently viewed page on the live web has been archived and proactively submit a request to various archives using an overlay.

Categories and Subject Descriptors
H.3.7 [Online information services]: Architectures and Protocols

1. INTRODUCTION

To better integrate the past and live web, implementations of the MicroMint framework use the Mink browser extension to query the archives (using URIs and HTTP Accept-Datetime headers as parameters) to provide resources on the past web.

Bridge live and archived Web
Provides a seamless viewing experience

RQ2: How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?

RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly

Is the metric for missing resources applicable across Web?

RQ1: What sort of content is difficult to capture and replay for preservation from the perspective of a Web browser?
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

Abstract

We describe the mobile app Mobile Mink which extends Mink, a browser extension that integrates the live and archived web. Mobile Mink discovers mobile and desktop URLs and provides the user an aggregated TimeMap of both mobile and desktop mementos. Mobile Mink also allows users to submit mobile and desktop URLs for archiving at the Internet Archive and Archive.today. Mobile Mink helps to increase the archival coverage of the growing mobile web.

RQ2: How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?

Mobile Mink is an Android application that is currently in development and will be released for download in the Google Play app store. Much like its desktop browser parent, Mobile Mink provides the user to navigate between the past and present webs. Mobile Mink also allows the user to submit mobile and desktop URLs to be archived by archival services.

When using a web browser native to the Android operating system, the user is presented with a list of options, one of which is the Mobile Mink shortcut in the browser's options (Figure 2). Selecting the option of viewing mementos begins the process of discovering mobile and desktop URLs of the current URI-R. First, Mobile Mink identifies the URI-R of the currently viewed page. Mobile Mink identifies the URI-R as either a mobile or desktop URI. If Mobile Mink identifies an initial URI in the URI-R, Mobile Mink translates the URI to a mobile or desktop URI and performs a URI modification (i.e., removes all HTTPS protocols) of that URI.
Personal archives are more resilient when propagated

RQ4: How can content that was captured behind authentication signal to Web archive replay systems that it requires special handling?

RQ6: What kinds of access control do users who create private Web archives need to regulate access to their archives?
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2017 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly

RQ2: How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Archive Replay
17. JCDL 2018 - A Framework...

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archived Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Preliminary Research

1. JCDL 2012 - WARCcreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly

ABSTRACT
We compare and contrast three different ways to implement an archival replay banner. We propose an implementation that utilizes Custom Elements and adds some unique behaviors, not common in existing archival replay systems, to enhance the user experience. Our approach has a minimal user interface footprint and resource overhead while still providing rich interactivity and extended on-demand provenance information about the archived resources.

CCS CONCEPTS
- Information systems → Digital libraries and archives → Human-centered computing → User interface design.

KEYWORDS
Memento/Archival Replay

RQ1: What sort of content is difficult to capture and replay for preservation from the perspective of a Web browser?

RQ2: How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?
Preliminary Research

1. JCDL 2012 - WARCreate
2. TPDL 2013 - Change in Archivability
3. DLib 2013 - Method for Identifying
4. JCDL 2014 - Mink
5. JCDL 2014 - Archival Acid Test
6. JCDL 2014 - Not All Mementos
7. IJDL 2015 - Impact of JavaScript
8. IJDL 2015 - Not All Mementos
9. JCDL 2015 - Mobile Mink
10. JCDL 2016 - InterPlanetary Wayback (ipwb)
11. TPDL 2016 - ipwb extended
12. JCDL 2017 - WAIL Electron
13. JCDL 2017 - ServiceWorker Replay
14. JCDL 2017 - Impact of Canonicalization
15. JCDL 2018 - ArchiveNow
16. JCDL 2018 - Replay Banners
17. JCDL 2018 - A Framework...
Outline

● Introduction/Motivation
● Background
● Preliminary Research
● Proposed Framework
● Evaluation Plan
Proposed Framework
(for aggregating private and public Web archives)
Proposed Framework

- Archival negotiation beyond time
- Query precedence & short-circuiting
- Mementities
PROPOSED FRAMEWORK

Archival Negotiation Beyond Time
More Expressive TimeMaps

- Memento Quality (e.g., Damage)\(^1\)
- How Many Captures?\(^2\)
- How Many Are Identical?\(^2,3\)
- Other Attributes of Mementos...

\(^1\) Brunelle et al., JCDL 2014, IJDL 2015
\(^2\) Kelly et al., JCDL 2017
\(^3\) AlSum and Nelson, ECIR 2014
Additional TimeMap Attributes

Content-based Attributes

Derived Attributes

Access Attributes
TimeMap Enrichment: Content-Based Attributes

- Status Code\(^1\)
- Content-Digest
  - In WARC & CDX
  - Not all archives expose CDX

- Would allow more info about mementos without requiring comprehensive dereferencing

\(^1\) Kelly et al., “Impact of URI Canonicalization on Memento Count”, JCDL 2017, arXiv 1703.03302
TimeMap Enrichment: Derived Attributes

- Thumbnails (e.g., via SimHash)\(^1\)
  - Calculation based on root memento’s HTML
- Memento Damage (JCDL 2014, IJDL)\(^2\)
  - Requires dereferencing embedded resources

---

\(^1\) AlSum and Nelson, Thumbnail Summarization Techniques for Web Archives, ECIR 2014, pp. 299-310.

How to distinguish Private captures from Public captures in a TimeMap?

RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
TimeMap Enrichment - in a CDXJ TimeMap

Line breaks added for clarity, CDXJ records occupy a single line

19981212013921 {
  "rel": "memento",
  "datetime": "Tue, 16 Nov 2010 06:05:16 GMT",
  "status_code": 200,
  "digest": "sha1:LK26DRRQJ4WATC6LBVF3B3Z4P2CP5ZZ7",
  "damage": 0.24,
  "simhash": "6551110622422153488",
  "content-language": "en-US",
  "access": {
    "type": "Blake2b",
    "token": "c6ed419e74907d220c69858614d86...ef0a3a88a41"
  }
}

Content-based attributes

Derived Attributes

Access Attributes
TimeMap

+ Enrichment with Additional Attributes

“StarMap”
PROPOSED FRAMEWORK

Query Precedence
- and -
Short Circuiting
Query Precedence

- More control of querying in series and parallel

"Check my archive first, then Carol’s, then all public archives."

See Atkins’, “Paywalls in the Internet Archive”, March 2018
Query Precedence

- More control of querying in series and parallel

“Check my archive first, then Carol’s, then all public archives.”

See Atkins’, “Paywalls in the Internet Archive”, March 2018
Query Precedence

- More control of querying in series and parallel

“Check my archive first, then Carol’s, then all public archives.”

See Atkins’, “Paywalls in the Internet Archive”, March 2018
Query Precedence

- More control of querying in series and parallel

“Check my archive first, then Carol’s, then all public archives.”

See Atkins’, “Paywalls in the Internet Archive”, March 2018
Query Short-Circuiting

- May give priority to archive relevancy.
- Series halt when threshold met.

“Check private archives first. Iff you find no captures, only then check the public archives.”
"Check private archives first. Iff you find no captures, only then check the public archives.

- May give priority to archive relevancy.
- Series halt when threshold met.
PROPOSED FRAMEWORK

Mementies
Mementities

- Memento + Entity (*entity* term already overused)

Introduced in this Framework

Conventional Memento
PROPOSED FRAMEWORK

Mementities
Memento Meta-Aggregator (MMA)
MMA: Archive Selection

GET /archives/

archivesList.json
MMA: User-Driven Archival Specification
MMA Aggregation sources

MMA_α:
from MA_2, MA_1, and WA_6

MMA_β:
from WA_7 and WA_8

MMA_γ:
from MMA_β, MA_5, and WA_1
MMA Dynamics By-Example

- Personal Archive Aggregation
- MMA Chaining
- Client-Side Aggregation Preference
MMA Dynamics - Personal Archive Aggregation

ALICE'S CNN.com captures

Public videos

ALICE'S private captures

FB  bank  flickr

Personal Archive Aggregation
Alice Saves the Web

Personal Archive Aggregation
Alice Wants to See Her Captures Temporally Inline

Personal Archive Aggregation
Mementity Dynamics - Alice & Her Archives (WA_A)

Personal Archive Aggregation
Alice Deploys $\text{MMA}_A$
Carol Asks $\text{MMA}_A$ for CNN

MMA Chaining
MMA_A returns CNN Memento \{M_A, M_{IA}\}
Carol Wants to Aggregate Her Own Captures

\[ \text{CNN}(M(WA_C)) \]
Carol Creates $\text{MMA}_C$ to Access $\text{WA}_C$ and $\text{MMA}_A$
Carol Asks MMA\textsubscript{C} For CNN

MMA Chaining
MMA_A returns CNN Memento \{M_A, M_{IA}, M_C\}
Bob May Request $M(CNN)$ From $MMA_A$ or $MMA_C$

Client-Side Aggregation Preference
Bob Prefers to Exclude IA Captures

...and does not want to setup his own MMA
Bob Requests Supported Archives

GET /archives/
Bob Customizes the Set in the JSON
Bob Requests CNN for His Custom Set

Client-Side Aggregation Preference
MMA Complies or Ignores Preference

Client-Side Aggregation Preference
Hooray, Aggregation!
PROPOSED FRAMEWORK

Mementities
Hooray, Aggregation!

RQ4: How can content that was captured behind authentication signal to Web archive replay systems that it requires special handling?
Private Web Archive Adapter (PWAA)

- Auth Layer for to encourage Private Web archive aggregation
- Typical OAuth 2.0 flow
- Auth role cohesive to PWAA
- Persistent access through tokenization

RQ6: What kinds of access control do users who create private Web archives need to regulate access to their archives?
PWAA - Sharing Tokens

RQ6: What kinds of access control do users who create private Web archives need to regulate access to their archives?
PWAA - Previously Authorized
PWAA - Unauthorized Request
PWAA - Sharing Tokens

RQ6: What kinds of access control do users who create private Web archives need to regulate access to their archives?
Alice Passes Associative Token to MMA

RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
MMA requests URI-R...
...relays token where applicable

RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
RQ5: How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?
PROPOSED FRAMEWORK

Mementities
StarGate

- Content negotiation in Web archives **beyond time**
- “Star” ~ wildcard (*) → **any** dimension of negotiation
- Allow for queries like: *Only show me mementos…*
  - That are not redirects (**content-based attribute** HTTP Status ≠ 3XX)
  - Of a sufficient quality (**derived attribute** Memento Damage < 0.4)
  - Are from personal Web archives (**access attribute** indicate Facebook.com memento is not a login page)
Implicit Filtering via MMA or Directly (a la TG)
Negotiation in the Privacy Dimension
(via short circuiting)

Get URI-Ms for URI-R only from personal Web archives
privateOnly
Negotiation on Content-Based or Derived Attributes
(with response filtering)

Get URI-Ms for URI-R of good quality that are unique
\( M_D < 0.25, unique(simhash) \)

1. CONTENT-BASED ATTRIBUTE

2. DERIVED ATTRIBUTE

3. Abbreviated StarMap with filtering applied
Outline

- Introduction/Motivation
- Background
- Preliminary Research
- Proposed Framework
- Evaluation Plan
Framework Evaluation

- Evaluation of mementity design decisions
- Costs of more expressive TimeMaps (StarMaps) and Link header enrichment
- Evaluation through implementation
Evaluation of Mementity Design Decisions

- Effectiveness in resolving initial use cases and access patterns
- “It was there yesterday, where did it go?”
- “Save this, but only for me.”
- “I want to share this but control who can see it.”
Costs of more expressive TimeMaps (StarMaps) and Link header enrichment

- **Computational:**
  - Mostly server-side, potential to further leverage client

- **Temporal**
  - Required on variant generation

- **Spatial**
  - Permutation variant storage

- **Access**
  - Variant negotiation

**EVALUATION:** Design/Practicality • Enrichment costs • Implementation
Evaluation Through Implementation

Extend for client-side archival specification

Exhibit features of an MMA

Regulate access to Private Web archives

Facilitate archival negotiation in more dimensions

EVALUATION: Design/Practicality • Enrichment costs • Implementation
A Framework for Aggregating Private and Public Web Archives

Mat Kelly
Old Dominion University
Web Science & Digital Libraries Research Group
Department of Computer Science
Norfolk, Virginia USA
mkelly@cs.odu.edu

Seminar, Penn State University
February 14, 2019
Backup Slides
Research Questions

**RQ1:** What sort of content is difficult to capture and replay for preservation from the perspective of a Web browser?

**RQ2:** How do Web browser APIs compare in potential functionality to the capabilities of archival crawlers?

**RQ3:** What issues exist for capturing and replaying content behind authentication?

**RQ4:** How can content that was captured behind authentication signal to Web archive replay systems that it requires special handling?

**RQ5:** How can Memento aggregators indicate that private Web archive content requires special handling to be replayed, despite being aggregated with publicly available Web archive content?

**RQ6:** What kinds of access control do users who create private Web archives need to regulate access to their archives?
User Access Patterns

- Pattern 1: Single archive access
- Pattern 2: Aggregation of multiple Web archives
- Pattern 3: Aggregator chaining
- Pattern 4: Aggregation with authentication
- Pattern 5: Aggregation including a hybrid public-private archive
- Pattern 6: Aggregation with filtering via MMA interaction
- Pattern 7: Aggregation with filtering via SG interaction

Pre-existing archival usage

Contribution beyond proposal
CDXJ: An Alternative TimeMap Format

![CDXJ TimeMap](image)

**Link (RFC 7089) TimeMap**

- Original URI (URI-R)
- Other TimeMaps (URI-Ts)
- TimeGate (URI-G)
- Relative Relations

Private & Public Archives May Differ for the Same URI
Should Public Archives *Really* Capture the Private Web?

A Framework for Aggregating Public and Private Web Archives
February 14, 2019
Mat Kelly