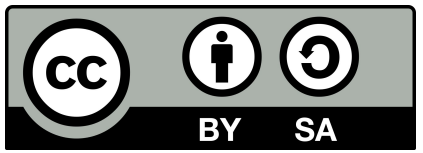
Facilitating A More Complete, Accessible Web History

Mat Kelly, Ph.D.
College of Computer and Informatics
Drexel University
mkelly@drexel.edu

Guest lecture for
INFO 821: Foundations of Information Science
May 2, 2023

https://matkelly.com/info821

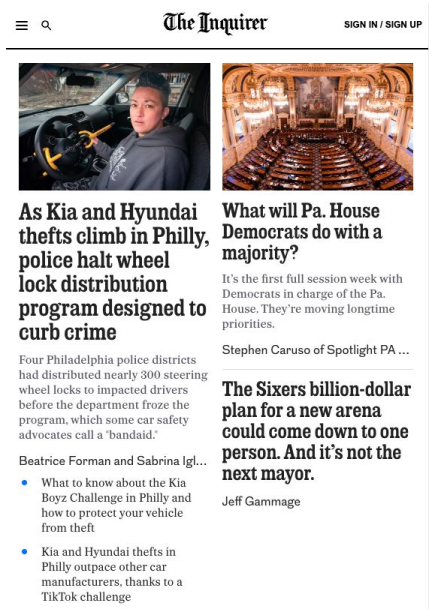


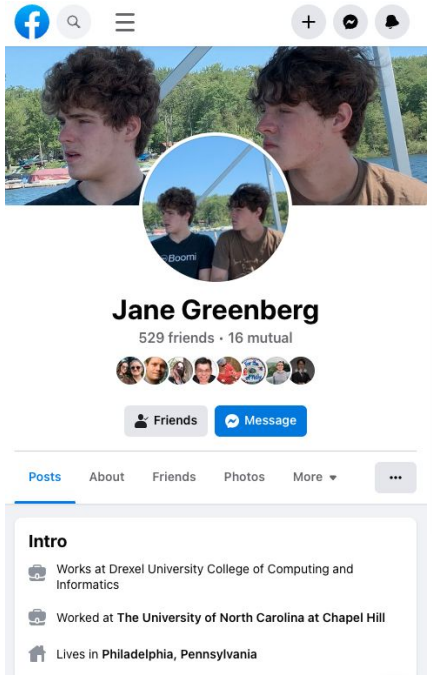
Mat Kelly, PhD

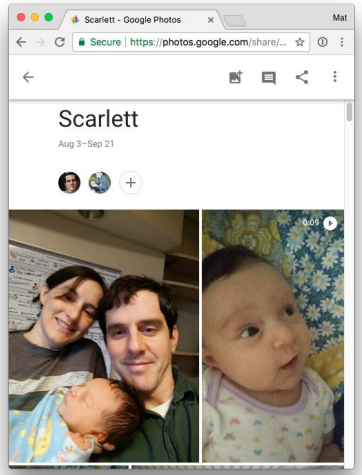
- Assistant Professor at Drexel CCI, Dept. of Information Science
- PhD 2019, Old Dominion University, Computer Science
- MS 2012, Old Dominion University, Computer Science
- BS 2006, University of Florida, Computer Science

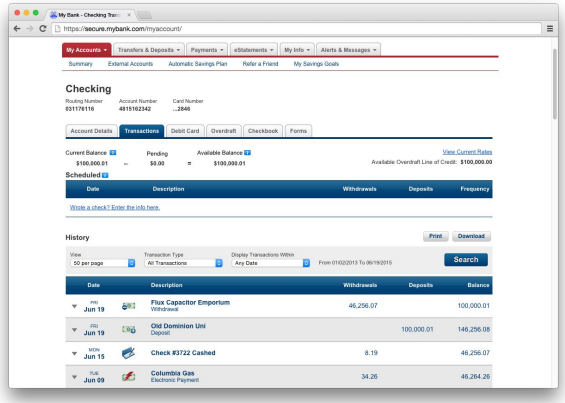
Research Focus - Web archiving

- Save the Web, it's important
- The Web has gotten increasingly complex!
- Should *everything* be saved? What about our private stuff?





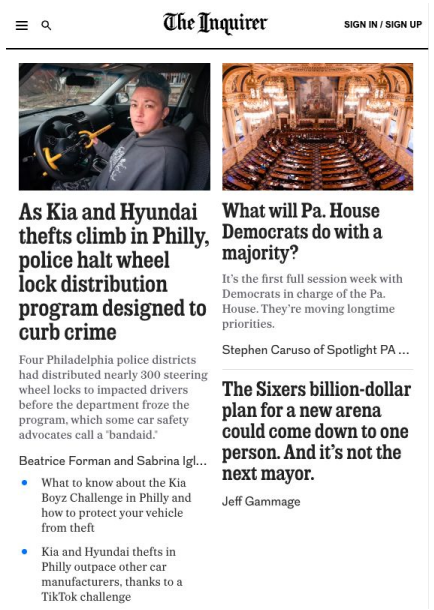




https://matkelly.com/info821

The Public Web is dynamic

- JavaScript may embed resources at runtime
  - e.g., fetch dynamics based on user interaction
- Consistent delta in web browsers and web archiving tools
- These tools don't have your creds, ergo private content never saved



https://matkelly.com/info821

Topical History

- “The scientist builds in order to study, the engineer studies in order to build.”
- Programmer turned researcher
- Driven my data liberation and interfacing incompatible systems

Frederick P. Brooks, Jr.

Fred Brooks is the first recipient of the **ACM Allen Newell Award***—an honor to be presented annually to an individual whose career contributions have bridged computer science and other disciplines. Brooks was honored for a breadth of career contributions within computer science and engineering and his interdisciplinary contributions to visualization methods for biochemistry. Here, we present his acceptance lecture delivered at SIGGRAPH 94.*

The Computer Scientist as

Toolsmith II

It is a special honor to receive an award named for Allen Newell. Allen was one of the fathers of computer science. He was especially important as a visionary and a leader in developing artificial intelligence (AI) as a subdiscipline, and in enunciating a vision for it.

What a man *is* is more important than what he does professionally, however, and it is Allen's humble, honorable, and self-giving character that makes it a double honor to be a Newell awardee. I am profoundly grateful to the awards committee.

Rather than talking about one particular research area, I should like to stay in the spirit of the Newell Award by sharing some lifetime reflections on the computer science enterprise, reflections which naturally reflect my convictions about the universe. The title and opening section of this talk were first formulated for a 1977 speech [1]. Let me reiterate the points, since many of you were barely born then.

In some quarters and at some times, computer graphics has been seen as a left-handed stepchild of computer science. Another view of computer science sees it as a discipline focused on problem-solving systems, and in this view computer graphics is very near the center of the discipline.

A Discipline Misnamed

When our discipline was newborn, there was the usual perplexity as to its proper name. We at Chapel Hill, following, I believe, Allen Newell and Herb Simon, settled on "computer science" as our department's name. Now, with the benefit of three decades' hindsight, I believe that to have been a mistake. If we understand why, we will better understand our craft.

What Is a Science?

Webster says science is "a branch of study concerned with the observation and classification of facts, especially with the establishment and quantitative formulation of verifiable general laws." [2]

This puts it pretty well—a science is concerned with the *discovery* of facts and laws.

A folk adage of the academic profession says, "Any-

Data Liberation

- Often (nowadays) by services to *allow* you to download your own data
- We know APIs are restrictive, fail, incomplete, etc.
- Web is similar:
  - What you experienced, you should be able to re-experience
  - Requires replication beyond the bits

ArchiveFacebook (2010)

- User-driven data liberation of *their* content on FB
- Firefox extension
- Open source
- Resultant data stored locally
- Caveats
  - Stored on file system, not “archived”
  - Limited platform (Firefox)
  - FB didn’t care for the name (i.e., C&D)
  - Extensions platform changed (XUI → WebExtensions)

ArchiveFacebook (2010-?)

- User-driven data liberation of *their* content on FB
- Firefox extension
- Open source
- Resultant data stored locally
- Caveats
  - Stored on file system, not “archived”
  - Limited platform (Firefox)
  - FB didn’t care for the name (i.e., C&D)
  - Extensions platform changed (XUI → WebExtensions)

WARC format

- International Standard (ISO 28500:2017) format for storing web archives
- Transactional record
  - retains HTTP requests, responses, metadata, crawl info, etc.
- All large web archiving efforts use format (e.g., Internet Archive)
- Generated by archival crawlers as they “visit” a live web page

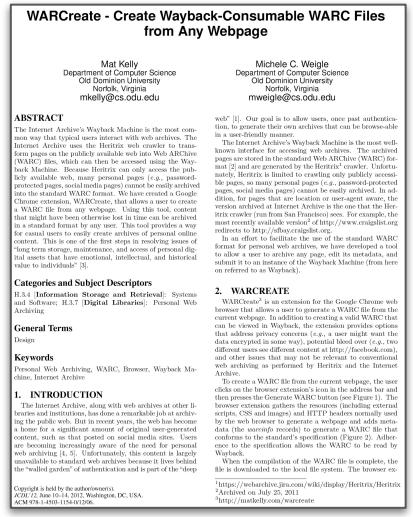
```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://mrc.cci.drexel.edu/
WARC-Date: 2023-05-01T21:48:51Z
WARC-Record-ID: <urn:uuid:be5034a7-dc37-61e2-7e14-6f8eb0d739f0>
Content-Type: application/http; msgtype=response
Content-Length: 92713

HTTP/1.1 200 OK
Server: nginx/1.18.0 (Ubuntu)
Date: Mon, 01 May 2023 21:48:51 GMT
Content-Type: text/html; charset=UTF-8
Transfer-Encoding: chunked
Connection: keep-alive
Link: <https://mrc.cci.drexel.edu/wp-json/>; rel="https://api.w.org/"
Link: <https://mrc.cci.drexel.edu/wp-json/wp/v2/pages/26>; rel="alterr
Link: <https://wp.me/P9jp5m-q>; rel=shortlink

<!DOCTYPE html><html lang="en-US" class=""><head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
<link rel="profile" href="http://gmpg.org/xfn/11">

<title>Metadata Research Center</title>
<meta name="robots" content="max-image-preview:large">
<link rel="dns-prefetch" href="//fonts.googleapis.com">
```

WARCreate (2012)



WARCreate - Create Wayback-Consumable WARC Files from Any Webpage

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

Michele C. Weigle
Department of Computer Science
Old Dominion University
Norfolk, Virginia
mweigle@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 ARChive (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 into the standard WARC format. We have created a Google Chrome extension, WARCreate, that allows a user to create 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 easily create archives of personal online content. This is one of the first steps in resolving issues of "long term storage, maintenance, and access of personal digital assets that have emotional, intellectual, and historical value to individuals" [3].

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.3.7 [**Digital Libraries**]: Personal Web Archiving

General Terms

Design

Keywords

Personal Web Archiving, WARC, Browser, Wayback Machine, Internet Archive

1. INTRODUCTION

The Internet Archive, along with web archives at other libraries and institutions, has done a remarkable job at archiving the public web. But in recent years, the web has become a home for a significant amount of original user-generated content, such as that posted on social media sites. Users are becoming increasingly aware of the need for personal web archiving [4, 5]. Unfortunately, this content is largely unavailable to standard web archives because it lives behind the "walled gardens" of authentication and is part of the "deep web" [1]. Our goal is to allow users, once past authentication, to generate their own archives that can be browse-able in a user-friendly manner.

The Internet Archive's Wayback Machine is the most well-known interface for accessing web archives. The archived pages are stored in the standard Web ARChive (WARC) format [2] and are generated by the Heritrix[1] crawler. Unfortunately, Heritrix is limited to crawling only publicly accessible pages, so many personal pages (e.g., password-protected pages, social media pages) cannot be easily archived. In addition, for pages that are location or user-agent aware, the version archived at Internet Archive is the one that the Heritrix crawler (run from San Francisco) sees. For example, the most recently available version[2] of http://www.craigslist.org redirects to http://sfbay.craigslist.org.

In an effort to facilitate the use of the standard WARC format for personal web archives, we have developed a tool to allow a user to archive any page, edit its metadata, and submit it to an instance of the Wayback Machine (from here on referred to as Wayback).

2. WARCREATE

WARCreate[3] is an extension for the Google Chrome web browser that allows a user to generate a WARC file from the current webpage. In addition to creating a valid WARC that can be viewed in Wayback, the extension provides options that address privacy concerns (e.g., a user might want the data encrypted in some way), potential bleed over (e.g., two different users see different content at http://facebook.com), and other issues that may not be relevant to conventional web archiving as performed by Heritrix and the Internet Archive.

To create a WARC file from the current webpage, the user clicks on the browser extension's icon in the address bar and then presses the Generate WARC button (see Figure 1). The browser extension gathers the resources (including external scripts, CSS and images) and HTTP headers normally used by the web browser to generate a webpage and adds metadata (the *warcinfo* records) to generate a WARC file that conforms to the standard's specification (Figure 2). Adherence to the specification allows the WARC to be read by Wayback.

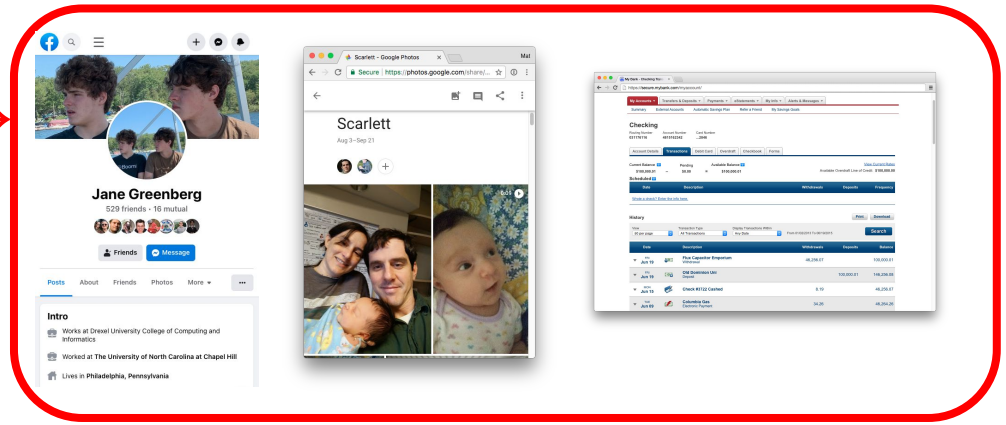
When the compilation of the WARC file is complete, the file is downloaded to the local file system. The browser ex-

[1] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
[2] Archived on July 25, 2011
[3] http://matkelly.com/warcreate

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

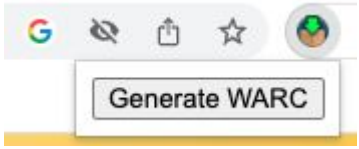
#preserving #linking #using #sharing
JCDL 2012
ACM/IEEE - CS Joint Conference on Digital Libraries

- Google Chrome extension
- "Create WARC files from any webpage"
- What you see is what you get
  - No delegation to a crawl
  - One-off archiving
- Could capture pages beyond those accessible to an institutional crawler
  - e.g., →



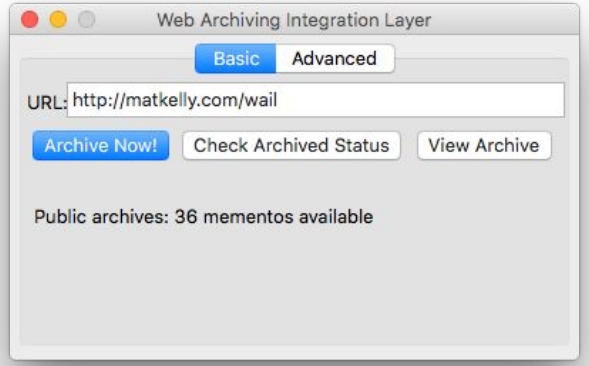
Seeding a Research Trajectory

- WARCreate was a driver for questions
  - What can(not) be captured?
  - If we could capture the previously uncapturable, where should we store it?
  - Should these captures be exhibited temporally inline?
  - What about privacy?
- Tools were hard to configure, aspiring personal web archivists would rather rely on simpler, yet effective interfaces



WARCreate's 1-button interface

WAIL (2013)



- Web Archiving Integration Layer
- Desktop app
- Bundled hard-to-configure web archiving tools into a simpler interface
  - Heritrix - institutional grade archival crawler
  - OpenWayback - archival replay system, interprets WARCs, makes usable

bundles:

and more!

Access is Fundamental to Preservation

The Memento Framework

- HTTP Framework for Time-Based Access to Resource States
- RFC 7089 (A Recognized Standard)

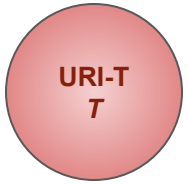
- Provides way to associate live Web URIs (URI-Rs)
  - https://drexel.edu
- With URIs of archived Web pages (URI-Ms)

  - http://web.archive.org/web/20110320142207/http://www.drexel.edu/

  - http://archive.is/dCqK

TimeMaps & TimeGates

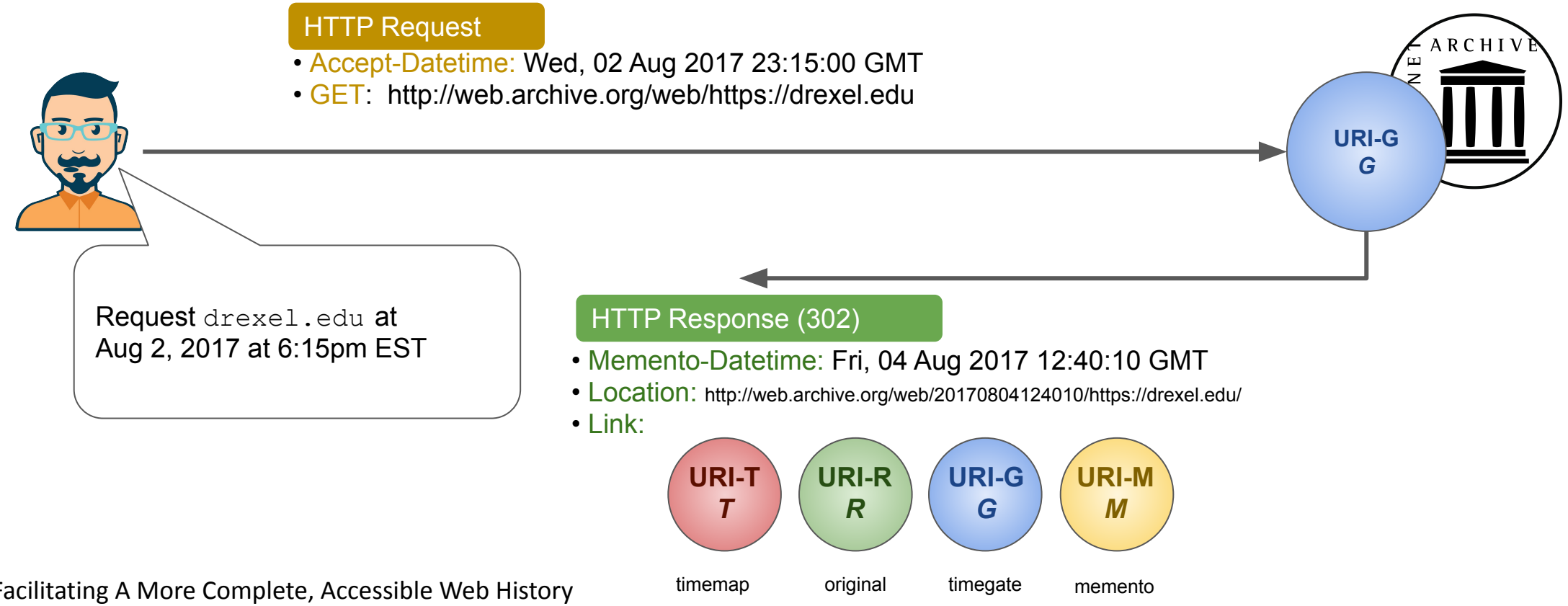
- TimeMaps – listing of URI-R, URI-Ms, and associated metadata
  - e.g., relative relation, datetime



- TimeGate – endpoint for requesting a URI-R *R* at time *t*
  - Enables content negotiation of the Web in the dimension of **time**

Content Negotiation in Time

- "Time Travel for the Web": using a TimeGate

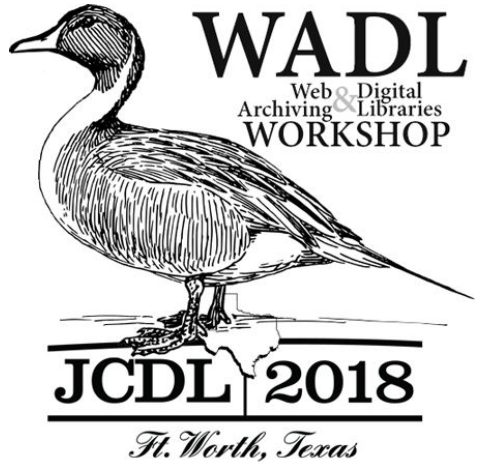


Which Archives are Queried?

- Archival sources are set server-side
- Client has no control over archival sources
- “You’ll get what you’ll get and you’ll be happy”
  - > a barrier in improving the picture of the past Web

{ Internet Archive | UK Web Archive – preserving uk web | archive.today – webpage capture }

more info, see:



Mat Kelly, Sawood Alam, Michael L. Nelson, and Michele C. Weigle, “**Client-Assisted Memento Aggregation Using the Prefer Header**,” Presented at the *Web Archiving and Digital Libraries Workshop (WADL 2018)*, June 2018.

TimeMaps Show Limited Information

WHERE? • URI-M
  - e.g., https://web.archive.org/web/20090512213206/http://www.drexel.edu/

WHEN? • Datetime – (RFC1123 - Requirements for Internet Hosts)
  - e.g., `Tue, 12 May 2009 21:32:06 GMT`

WHAT? • Link Relation (RFC5988 - Web Linking)
  - e.g., `rel="first memento"`

Same TimeMap Metadata in Multiple Formats

```
...
<http://localhost:8080/20101116060516/http://facebook.com/>; rel="memento";
datetime="Tue, 16 Nov 2010 06:05:16 GMT",
...
```

Memento entry in Link (RFC 7089) TimeMap

Memento entry in CDXJ TimeMap

```
...
20101116060516 {
  "uri": "http://localhost:8080/20101116060516/http://facebook.com/",
  "rel": "memento",
  "datetime": "Tue, 16 Nov 2010 06:05:16 GMT",
}
...
```

Memento (URI-M)

Relative Relations

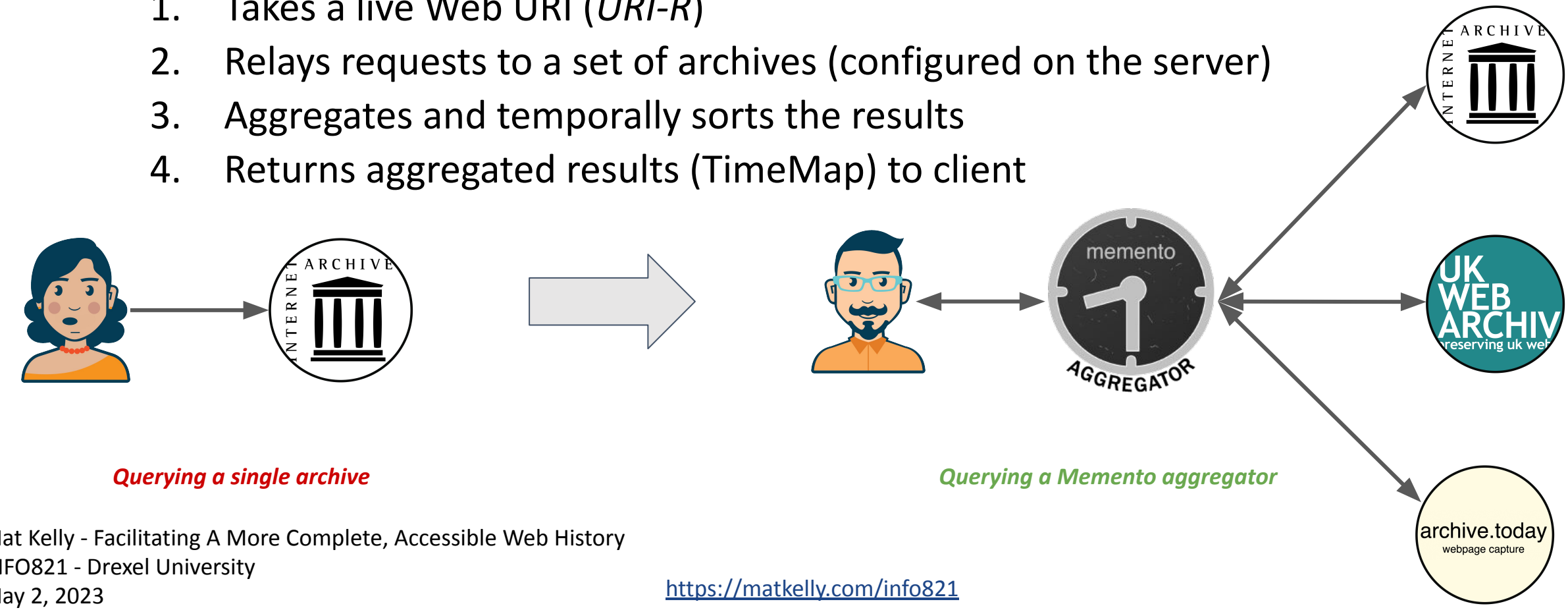
Memento-Datetime

Memento Aggregation

filling temporal gaps by using multiple sources

Memento aggregators are “queryable” Web services that:

1. Takes a live Web URI (*URI-R*)
2. Relays requests to a set of archives (configured on the server)
3. Aggregates and temporally sorts the results
4. Returns aggregated results (TimeMap) to client

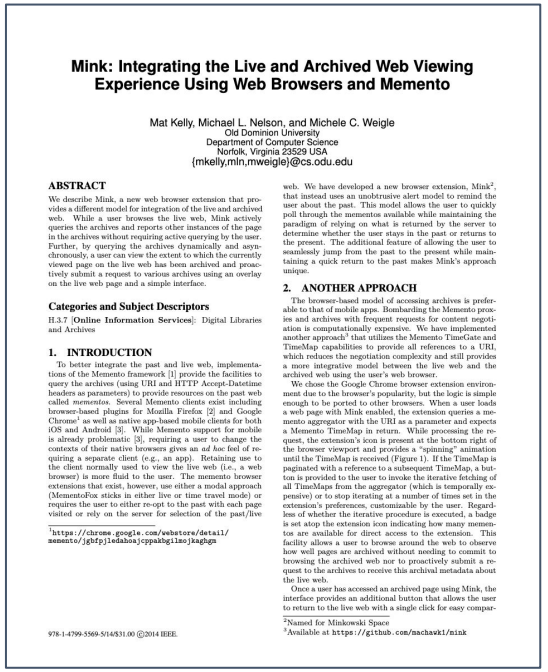


Querying a single archive

Querying a Memento aggregator

Mink (2014)

- Chrome extension
- A unified experience, view the extent of history available for a web page as you browse
- One-click submission to multiple web archives
  - Recall: appeal of simple interfaces



Mink: Integrating the Live and Archived Web Viewing Experience Using Web Browsers and Memento

Mat Kelly, Michael L. Nelson, and Michele C. Weigle
Old Dominion University
Department of Computer Science
Norfolk, Virginia 23529 USA
{mkelly,mln,mweigle}@cs.odu.edu

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 on the live web page and a simple interface.

Categories and Subject Descriptors

H.3.7 [**Online Information Services**]: Digital Libraries and Archives

1. INTRODUCTION

To better integrate the past and live web, implementations of the Memento framework [1] provide the facilities to query the archives (using URI and HTTP Accept-Datetime headers as parameters) to provide resources on the past web called *mementos*. Several Memento clients exist including browser-based plugins for Mozilla Firefox [2] and Google Chrome[1] as well as native app-based mobile clients for both iOS and Android [3]. While Memento support for mobile is already problematic [3], requiring a user to change the contexts of their native browsers gives an *ad hoc* feel of requiring a separate client (e.g., an app). Retaining use to the client normally used to view the live web (i.e., a web browser) is more fluid to the user. The memento browser extensions that exist, however, use either a modal approach (MementoFox sticks in either live or time travel mode) or requires the user to either re-opt to the past with each page visited or rely on the server for selection of the past/live web. We have developed a new browser extension, Mink[2], that instead uses an unobtrusive alert model to remind the user about the past. This model allows the user to quickly poll through the mementos available while maintaining the paradigm of relying on what is returned by the server to determine whether the user stays in the past or returns to the present. The additional feature of allowing the user to seamlessly jump from the past to the present while maintaining a quick return to the past makes Mink's approach unique.

2. ANOTHER APPROACH

The browser-based model of accessing archives is preferable to that of mobile apps. Bombarding the Memento proxies and archives with frequent requests for content negotiation is computationally expensive. We have implemented another approach[3] that utilizes the Memento TimeGate and TimeMap capabilities to provide all references to a URI, which reduces the negotiation complexity and still provides a more integrative model between the live web and the archived web using the user's web browser.

We chose the Google Chrome browser extension environment due to the browser's popularity, but the logic is simple enough to be ported to other browsers. When a user loads a web page with Mink enabled, the extension queries a memento aggregator with the URI as a parameter and expects a Memento TimeMap in return. While processing the request, the extension's icon is present at the bottom right of the browser viewport and provides a "spinning" animation until the TimeMap is received (Figure 1). If the TimeMap is paginated with a reference to a subsequent TimeMap, a button is provided to the user to invoke the iterative fetching of all TimeMaps from the aggregator (which is temporally expensive) or to stop iterating at a number of times set in the extension's preferences, customizable by the user. Regardless of whether the iterative procedure is executed, a badge is set atop the extension icon indicating how many mementos are available for direct access to the extension. This facility allows a user to browse around the web to observe how well pages are archived without needing to commit to browsing the archived web nor to proactively submit a request to the archives to receive this archival metadata about the live web.

Once a user has accessed an archived page using Mink, the interface provides an additional button that allows the user to return to the live web with a single click for easy compar-

[1] https://chrome.google.com/webstore/detail/memento/jgbfpjledahoajcppahbglimojkagbgn

[2] Named for Minkowski Space

[3] Available at https://github.com/machawk1/mink

978-1-4799-5569-5/14/$31.00 ©2014 IEEE.

JCDL 2014

Problems Remained

1. How to temporally blend private/personal captures with extensive history of the public web
2. URI-R is not enough to distinguish private/public captures
   - Among other variants
3. Machine dies, efforts for naught

InterPlanetary Wayback (ipwb) (2016)

- Integrated WARCs with the InterPlanetary File System (IPFS)
- Allowed personally captured WARCs to be more resilient in time, shared P2P
- Content addressing rather than lookup by URI-R
  - Facilitated efficient storage, integrity, deduplication, etc.

Did not curb issues of PII within WARCs

apple.com over time

InterPlanetary Wayback: The Permanent Web Archive

Sawood Alam, Mat Kelly, and Michael L. Nelson
Old Dominion University, Department of Computer Science, Norfolk VA, 23529, USA
{salam,mkelly,mln}@cs.odu.edu

JCDL 2016 NEWARK

TPDL2016 HANNOVER

https://matkelly.com/info821

Archival Content Negotiation in Dimensions Beyond Time

For `drexel.edu`, show me…

- only unique captures (1 URI-M per hash variant)
- an efficiently thumbnails summary (use SimHash for thumbnail generation)
- only capture where the quality is > 0.24 (w/ a custom metric)

- Any of the above in combination or with an additional datetime parameter

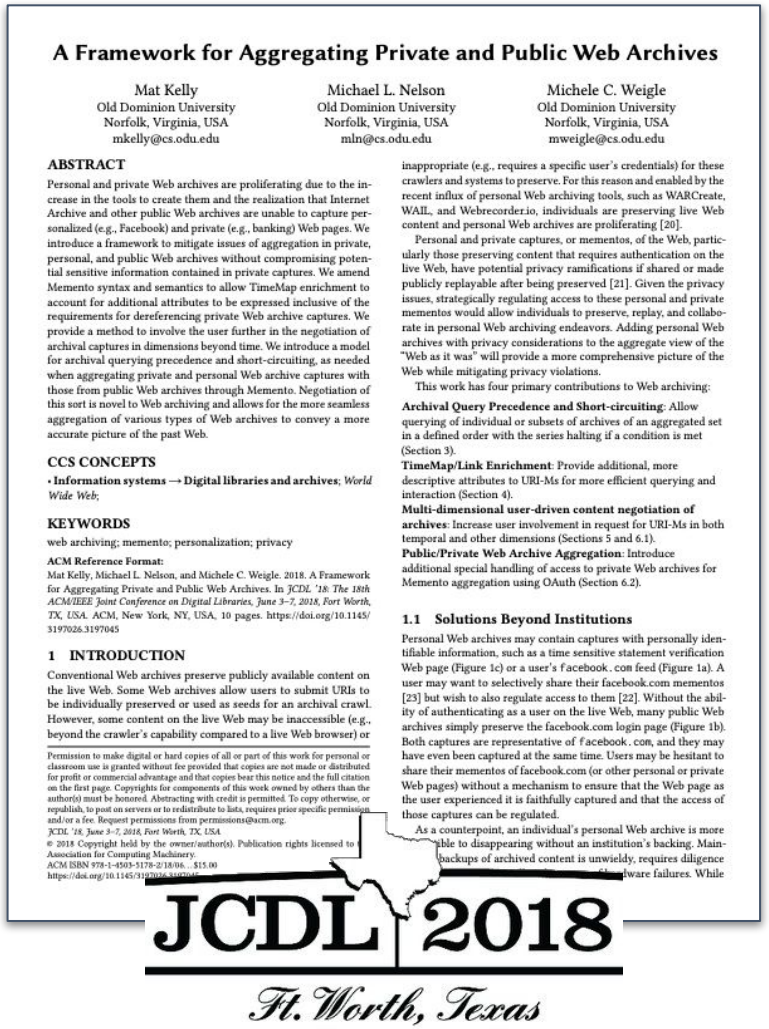
(note the potential for combinatorial complexity)

A Framework for Aggregating Private and Public Web Archives

Provided a hierarchical approach at supplementing the set of Web archives aggregated

Regulate access to Private Web archives

Facilitate archival negotiation in more dimensions



A Framework for Aggregating Private and Public Web Archives

Mat Kelly, Old Dominion University, Norfolk, Virginia, USA, mkelly@cs.odu.edu
Michael L. Nelson, Old Dominion University, Norfolk, Virginia, USA, mln@cs.odu.edu
Michele C. Weigle, Old Dominion University, Norfolk, Virginia, USA, mweigle@cs.odu.edu

ABSTRACT

Personal and private Web archives are proliferating due to the increase in the tools to create them and the realization that Internet Archive and other public Web archives are unable to capture personalized (e.g., Facebook) and private (e.g., banking) Web pages. We introduce a framework to mitigate issues of aggregation in private, personal, and public Web archives without compromising potential sensitive information contained in private captures. We amend Memento syntax and semantics to allow TimeMap enrichment to account for additional attributes to be expressed inclusive of the requirements for dereferencing private Web archive captures. We provide a method to involve the user further in the negotiation of archival captures in dimensions beyond time. We introduce a model for archival querying precedence and short-circuiting, as needed when aggregating private and personal Web archive captures with those from public Web archives through Memento. Negotiation of this sort is novel to Web archiving and allows for the more seamless aggregation of various types of Web archives to convey a more accurate picture of the past Web.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; *World Wide Web*;

KEYWORDS

web archiving; memento; personalization; privacy

ACM Reference Format:

Mat Kelly, Michael L. Nelson, and Michele C. Weigle. 2018. A Framework for Aggregating Private and Public Web Archives. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3197026.3197045

1 INTRODUCTION

Conventional Web archives preserve publicly available content on the live Web. Some Web archives allow users to submit URIs to be individually preserved or used as seeds for an archival crawl. However, some content on the live Web may be inaccessible (e.g., beyond the crawler's capability compared to a live Web browser) or inappropriate (e.g., requires a specific user's credentials) for these crawlers and systems to preserve. For this reason and enabled by the recent influx of personal Web archiving tools, such as WARCreate, WAIL, and Webrecorder.io, individuals are preserving live Web content and personal Web archives are proliferating [20].

Personal and private captures, or mementos, of the Web, particularly those preserving content that requires authentication on the live Web, have potential privacy ramifications if shared or made publicly replayable after being preserved [21]. Given the privacy issues, strategically regulating access to these personal and private mementos would allow individuals to preserve, replay, and collaborate in personal Web archiving endeavors. Adding personal Web archives with privacy considerations to the aggregate view of the "Web as it was" will provide a more comprehensive picture of the Web while mitigating privacy violations.

This work has four primary contributions to Web archiving:

Archival Query Precedence and Short-circuiting: Allow querying of individual or subsets of archives of an aggregated set in a defined order with the series halting if a condition is met (Section 3).

TimeMap/Link Enrichment: Provide additional, more descriptive attributes to URI-Ms for more efficient querying and interaction (Section 4).

Multi-dimensional user-driven content negotiation of archives: Increase user involvement in request for URI-Ms in both temporal and other dimensions (Sections 5 and 6.1).

Public/Private Web Archive Aggregation: Introduce additional special handling of access to private Web archives for Memento aggregation using OAuth (Section 6.2).

1.1 Solutions Beyond Institutions

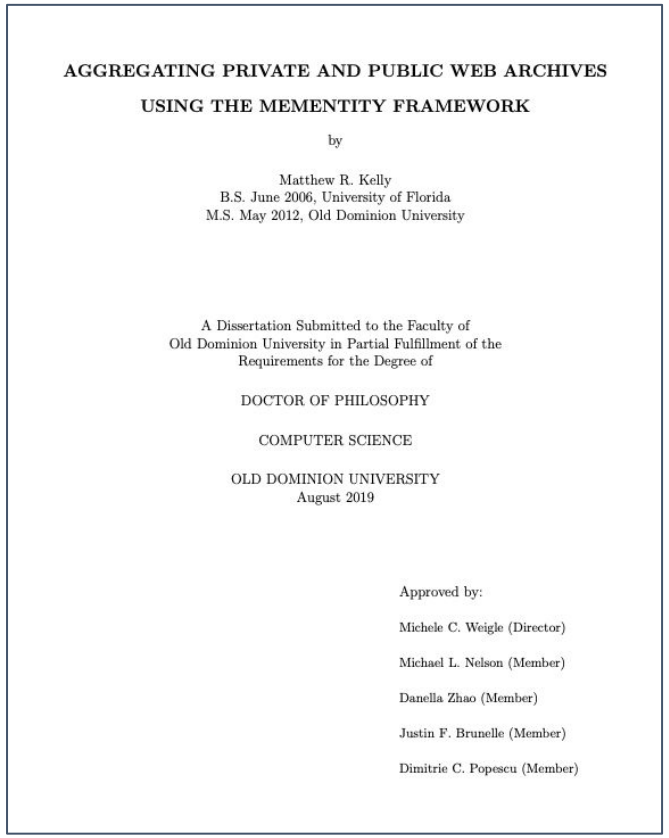
Personal Web archives may contain captures with personally identifiable information, such as a time sensitive statement verification Web page (Figure 1c) or a user's facebook.com feed (Figure 1a). A user may want to selectively share their facebook.com mementos [23] but wish to also regulate access to them [22]. Without the ability of authenticating as a user on the live Web, many public Web archives simply preserve the facebook.com login page (Figure 1b). Both captures are representative of facebook.com, and they may have even been captured at the same time. Users may be hesitant to share their mementos of facebook.com (or other personal or private Web pages) without a mechanism to ensure that the Web page as the user experienced it is faithfully captured and that the access of those captures can be regulated.

As a counterpoint, an individual's personal Web archive is more [illegible] to disappearing without an institution's backing. Main- [illegible] backups of archived content is unwieldy, requires diligence [illegible] ware failures. While

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
JCDL '18, June 3–7, 2018, Fort Worth, TX, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-5178-2/18/06…$15.00
https://doi.org/10.1145/3197026.3197045

Mat Kelly, Michael L. Nelson, and Michele C. Weigle, "A Framework for Aggregating Private and Public Web Archives," In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Fort Worth, Texas, June 2018, pp. 273–282.

Defended, Graduated, Joined Drexel (2019)



AGGREGATING PRIVATE AND PUBLIC WEB ARCHIVES
USING THE MEMENTITY FRAMEWORK

by

Matthew R. Kelly
B.S. June 2006, University of Florida
M.S. May 2012, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2019

Approved by:

Michele C. Weigle (Director)

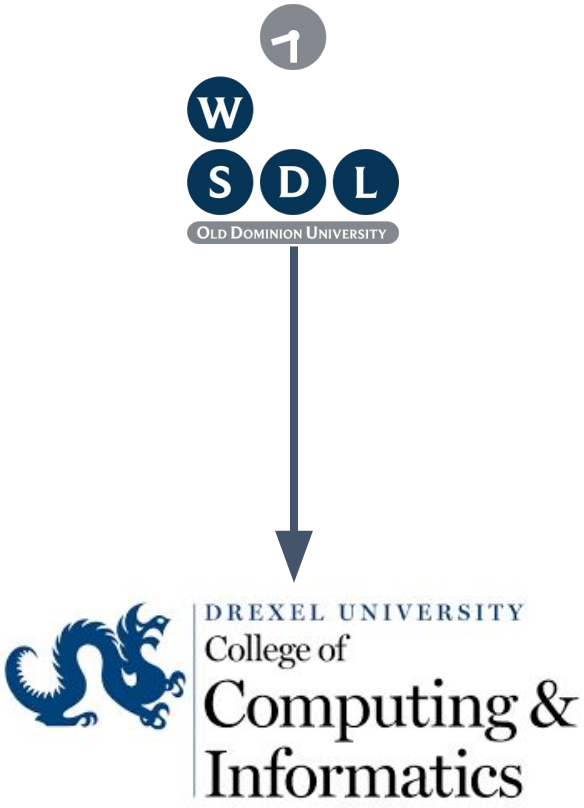
Michael L. Nelson (Member)

Danella Zhao (Member)

Justin F. Brunelle (Member)

Dimitrie C. Popescu (Member)

https://matkelly.com/dissertation



Outstanding Questions*

(points needing further research)

- Memento introduces linking and inter-resource relations
  - intentionally open-ended for further extension and exploration
- Web archives are very large but largely centralized
- Much web archive research has been in usage and not creation, enrichment, enhancing access, etc.
- Private Web archives are rarely considered but often considered the most important (i.e., users' personal) Web content
- Semantics, asynchronous generation, caching/storage
- Further leveraging client-side querying for information retrieval

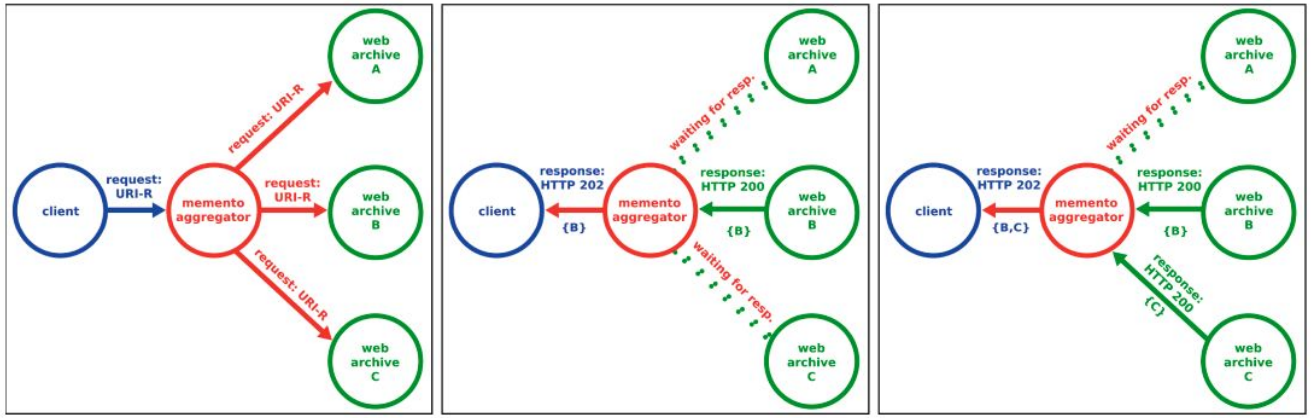
* Since last presenting in Jane's INFO821

Threads Abound, Some Loose, Some Newfound

- Distributed Persistent Identifiers (ARKs)
  - Nuances akin to distributed aggregators (MMA), resilience (ipwb)
- Research Using (un-)Archived Content
  - HBCU Faculty Migration (as evidenced on the past web)
  - Missing Web advertisements as lost historical context
- Inter-domain term evolution
  - Temporal aspect, but in undefined dimensions
- Complexities of a more complete web history…

Complexities of Advanced Aggregation

- Suppose you run an aggregator to curate your narrative's web sources
- Queries may, in-turn query other aggregators
  - Some of which may have privileged access to private archives
- Waiting for all to respond is inefficient, can be pipelined



Aggregator Reuse and Extension for Richer Web Archive Interaction

Mat Kelly[0000-0001-9233-0172]

Drexel University, Philadelphia, PA 19104, USA
mkelly@drexel.edu
https://matkelly.com

Abstract. Memento aggregators enable users to query multiple web archives for captures of a URI in time through a single HTTP endpoint. While this one-to-many access point is useful for researchers and end-users, aggregators are in a position to provide additional functionality to end-users beyond black box style aggregation. This paper identifies the state-of-the-art of Memento aggregation, abstracts its processes, highlights shortcomings, and offers systematic enhancements.

1 Introduction

Web archives act as a historical record of the web. The Internet Archive (IA) possesses the largest number of web archive holdings. These holdings are accessible through a set of interfaces to the Wayback Machine. Beyond IA, other web archives exhibit focused collection efforts, often providing unique captures within IA's temporal and spatial (i.e., URL [7]) voids [17]. A common usage pattern in accessing IA's captures is to request the archive's web site at archive.org, submit a URL of interest by providing it in a text input field, then selecting a date and time from the set of available captures for that URL in the past. This pattern may differ between web archives' respective web interfaces. Memento [27] provides the standards-based interoperable means, dynamics, syntax, and semantics for representing identifiers for archival captures (mementos) from a set of web archives. Each archive that supports the Memento Framework provides an HTTP endpoint for retrieving mementos from their respective archival holdings. Users can send a request for all captures of a URL to a variety of supporting archives through a single endpoint by an accessible tool that performs the logic of querying and combining results from multiple sources—a Memento aggregator.

Memento aggregators typically have reference to a set of endpoints to web archives that implement the Memento Framework. An aggregator may express this through a URI "template" like Fig. 1 or as a URI with an implicit append operation of a URI-R [27]. Upon receiving a request from a client with a parameterized URL (e.g., the URI-R applied to the template URI), an aggregator relays the argument received in this request as parameters for subsequent requests to each archive. When the aggregator receives a sufficient response,[1] as dictated

[1] This criteria is implementation-specific and may be associated with a temporal threshold, memento count, etc.

Y.-H. Tseng et al. (Eds.): ICADL 2022, LNCS 13636, pp. 313–328, 2022.
https://doi.org/10.1007/978-3-031-21756-2_25

ICADL 2022

Mat Kelly, "Aggregator Reuse and Extension for Richer Web Archive Interaction," In Proceedings of the 24th International Conference on Asia-Pacific Digital Libraries (ICADL 2022), Hanoi, Vietnam, November 30–December 2, 2022, pp. 313–328.
https://matkelly.com/info821

Client-side Replay Tweaks for Truer Experience

- Mindset: archived web page has to be manipulated (e.g., links) to ensure archived resources representations resolve
- This is often done server-side, presenting a manipulation of the bits of the historical record.
  - Typically a bad thing, but the experience is better
  - An imperfect process, dynamically loaded resources are hard to rewrite
- Client-side JavaScript techs (e.g., ServiceWorkers) can improve the experience without manipulating the bits server-side
  - …and allow some inaccessible resources to be resolved!

Has already been implemented in high-fidelity replay system and informs current archival replay practice

John Berlin, Mat Kelly, Michael L. Nelson, and Michele C. Weigle, "To Re-experience the Web: A Framework for the Transformation and Replay of Archived Web Pages," ACM Transactions on the Web (TWEB), Just Accepted. March 2023.

Facilitating A More Complete, Accessible Web History

Mat Kelly, Ph.D.
College of Computer and Informatics
Drexel University
mkelly@drexel.edu

John Berlin, Mat Kelly, Michael L. Nelson, and Michele C. Weigle, “To Re-experience the Web: A Framework for the Transformation and Replay of Archived Web Pages,” ACM Transactions on the Web (TWEB), Just Accepted. March 2023.
Mat Kelly, “Aggregator Reuse and Extension for Richer Web Archive Interaction,” In Proceedings of the 24th International Conference on Asia-Pacific Digital Libraries (ICADL 2022), Hanoi, Vietnam, November 30–December 2, 2022, pp. 313–328.
Mat Kelly, Michael L. Nelson, and Michele C. Weigle, “A Framework for Aggregating Private and Public Web Archives,” In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Fort Worth, Texas, June 2018, pp. 273–282.

DREXEL UNIVERSITY College of Computing & Informatics

https://matkelly.com/info821