Saving the Web

Mat Kelly
mkelly@drexel.edu
Assistant Professor, Information Science
Drexel University College of Computing and Informatics

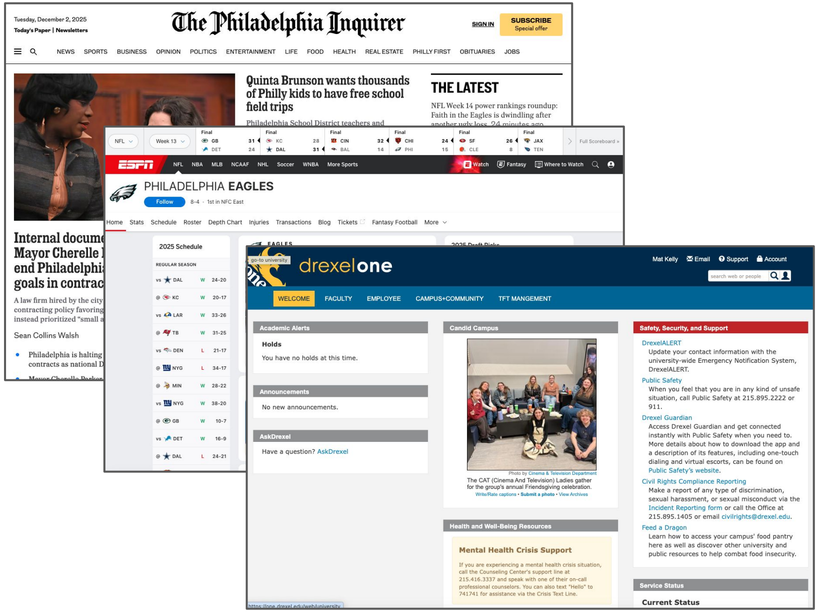
INFO591 Week 10
December 4, 2025

Supported by:
INSTITUTE of Museum and Library SERVICES

Slides: https://bit.ly/info591

https://matkelly.com

The Web



Tuesday, December 2, 2025
Today's Paper | Newsletters

The Philadelphia Inquirer

SIGN IN SUBSCRIBE Special offer

NEWS SPORTS BUSINESS OPINION POLITICS ENTERTAINMENT LIFE FOOD HEALTH REAL ESTATE PHILLY FIRST OBITUARIES JOBS

Quinta Brunson wants thousands of Philly kids to have free school field trips

Philadelphia School District teachers and

THE LATEST

NFL Week 14 power rankings roundup: Faith in the Eagles is dwindling after another ugly loss. 24 minutes ago

Internal docume
Mayor Cherelle
end Philadelphi
goals in contrac

A law firm hired by the city
contracting policy favoring
instead prioritized "small a

Sean Collins Walsh

- Philadelphia is halting
contracts as national D
- Mayor Cherelle Parker

NFL Week 13 Final GB 31 DET 24 Final KC 28 DAL 31 Final CIN 32 BAL 14 Final CHI 24 PHI 15 Final SF 26 CLE 8 Final JAX TEN Full Scoreboard »

ESPN NFL NBA MLB NCAAF NHL Soccer WNBA More Sports Watch Fantasy Where to Watch

PHILADELPHIA EAGLES
Follow 8-4 · 1st in NFC East

Home Stats Schedule Roster Depth Chart Injuries Transactions Blog Tickets Fantasy Football More

2025 Schedule

REGULAR SEASON

vs	DAL	W	24-20
@	KC	W	20-17
vs	LAR	W	33-26
@	TB	W	31-25
vs	DEN	L	21-17
@	NYG	L	34-17
@	MIN	W	28-22
vs	NYG	W	38-20
@	GB	W	10-7
vs	DET	W	16-9
@	DAL	L	24-21

EAGLES

2025 Draft Picks

go-to university drexelone

Mat Kelly Email Support Account
search web or people

WELCOME FACULTY EMPLOYEE CAMPUS+COMMUNITY TFT MANGEMENT

Academic Alerts

Holds

You have no holds at this time.

Announcements

No new announcements.

AskDrexel

Have a question? AskDrexel

Candid Campus

Photo by Cinema & Television Department

The CAT (Cinema And Television) Ladies gather for the group's annual Friendsgiving celebration.

Write/Rate captions • Submit a photo • View Archives

Health and Well-Being Resources

Mental Health Crisis Support

If you are experiencing a mental health crisis situation, call the Counseling Center's support line at 215.416.3337 and speak with one of their on-call professional counselors. You can also text "Hello" to 741741 for assistance via the Crisis Text Line.

Safety, Security, and Support

DrexelALERT
Update your contact information with the university-wide Emergency Notification System, DrexelALERT.

Public Safety
When you feel that you are in any kind of unsafe situation, call Public Safety at 215.895.2222 or 911.

Drexel Guardian
Access Drexel Guardian and get connected instantly with Public Safety when you need to. More details about how to download the app and a description of its features, including one-touch dialing and virtual escorts, can be found on Public Safety's website.

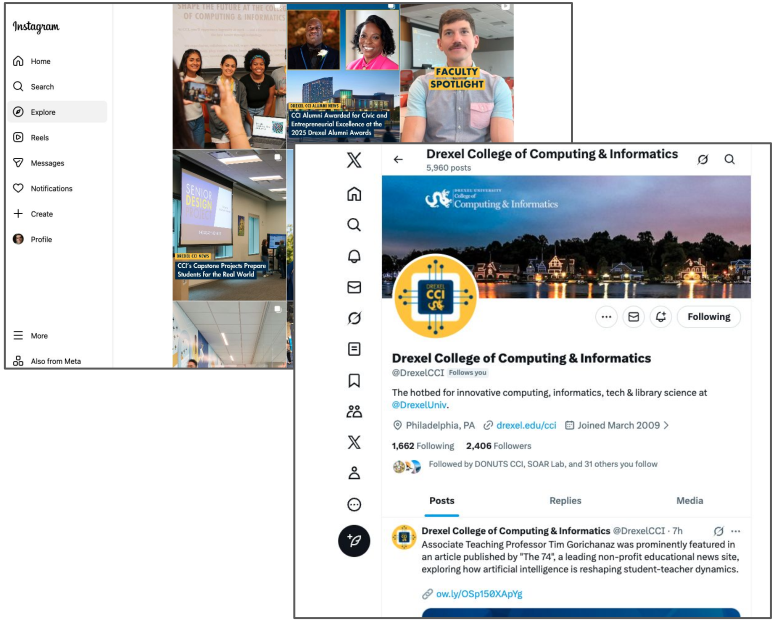
Civil Rights Compliance Reporting
Make a report of any type of discrimination, sexual harassment, or sexual misconduct via the Incident Reporting form or call the Office at 215.895.1405 or email civilrights@drexel.edu.

Feed a Dragon
Learn how to access your campus' food pantry here as well as discover other university and public resources to help combat food insecurity.

Service Status

Current Status

https://one.drexel.edu/web/university



Instagram

Home
Search
Explore
Reels
Messages
Notifications
Create
Profile
More
Also from Meta

SHAPE THE FUTURE AT THE CO... OF COMPUTING & INFORMATICS

DREXEL CCI ALUMNI NEWS
CCI Alumni Awarded for Civic and Entrepreneurial Excellence at the 2025 Drexel Alumni Awards

FACULTY SPOTLIGHT

SENIOR DESIGN

DREXEL CCI NEWS
CCI's Capstone Projects Prepare Students for the Real World

Drexel College of Computing & Informatics
5,960 posts

Drexel University College of Computing & Informatics

Following

Drexel College of Computing & Informatics
@DrexelCCI Follows you

The hotbed for innovative computing, informatics, tech & library science at @DrexelUniv.

Philadelphia, PA drexel.edu/cci Joined March 2009

1,662 Following 2,406 Followers

Followed by DONUTS CCI, SOAR Lab, and 31 others you follow

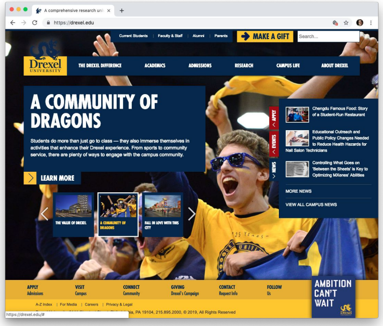
Posts Replies Media

Drexel College of Computing & Informatics @DrexelCCI · 7h
Associate Teaching Professor Tim Gorichanaz was prominently featured in an article published by "The 74", a leading non-profit educational news site, exploring how artificial intelligence is reshaping student-teacher dynamics.

ow.ly/OSp150XApYg

The Web's Importance

- Information access
- Personal interfacing (e.g., social media)
- Cultural artifact
- Personal knowledge reference



A COMMUNITY OF DRAGONS

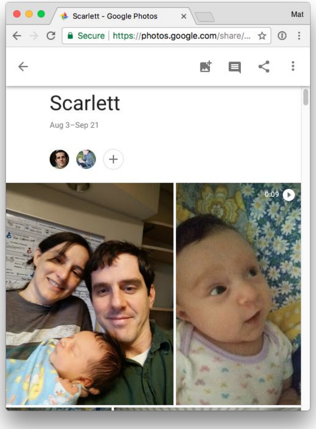
Students do more than just go to class — they also immerse themselves in activities that enhance their Drexel experience. From sports to community service, there are plenty of ways to engage with the campus community.

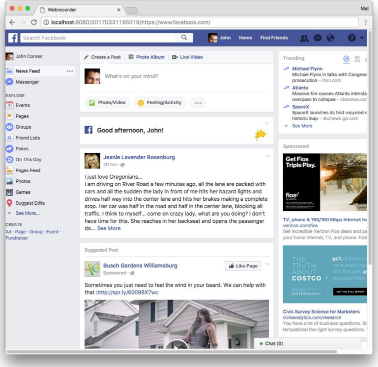
Chengdu Famous Food: Story of a Student-Run Restaurant

Educational Outreach and Public Policy Changes Needed to Reduce Health Hazards for Nail Salon Technicians

Controlling What Goes on 'Between the Sheets' Is Key to Optimizing MXenes' Abilities

ia, PA 19104, 215.895.2000, © 2019, All Rights Reserved





Good afternoon, John!

Jeanie Lavender Rosenburg
20 hrs ·

I just love Oregonians...
I am driving on River Road a few minutes ago, all the lane are packed with cars and all the sudden the lady in front of me hits her hazard lights and drives half way into the center lane and hits her brakes making a complete stop. Her car was half in the road and half in the center lane, blocking all traffic. I think to myself... come on crazy lady, what are you doing? I don't have time for this. She reaches in her backseat and opens the passenger do... See More

Busch Gardens Williamsburg
Sponsored ·

Sometimes you just need to feel the wind in your beard. We can help with that :http://spr.ly/60098X7wc

Michael Flynn
Michael Flynn in talks with Congres
prosecution – msn.com

Atlanta
Massive fire causes Atlanta intersta
overpass to collapse – cbsnews.cor

SpaceX
SpaceX launches its first recycled r
historic leap – abcnews.go.com

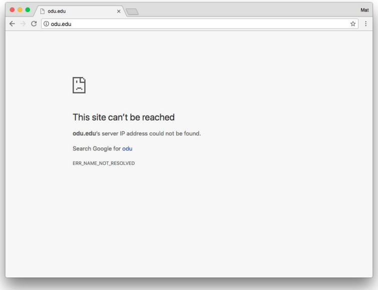
TV, phone & 150/150 Mbps Internet fo
verizon.com/fios
Get incredible Verizon Fios deals and pa
your home internet, TV, and phone. Fast

91% of America
consumers hav
of it, but only 4
shopped there

Civis Survey Science for Marketers
civisanalytics.com/research
You have a lot of business questions. Sc
templatized the right survey questions.

The Web is Ephemeral

- No guarantee that information or services available today will be available tomorrow
- We are reliant on this service
  - Inclusive of mobile apps
- Links rot quickly
- User-generated content is especially fragile
  - Apps, social media posts, and comments vanish as services evolve

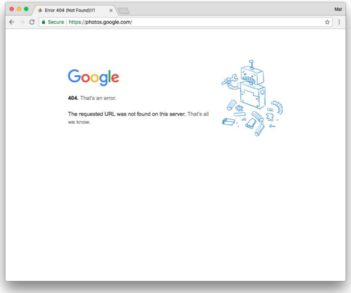


This site can't be reached

odu.edu's server IP address could not be found.

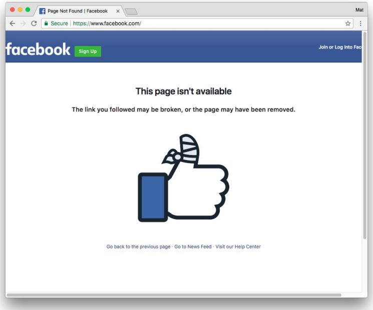
Search Google for odu

ERR_NAME_NOT_RESOLVED



404. That's an error.

The requested URL was not found on this server. That's all we know.



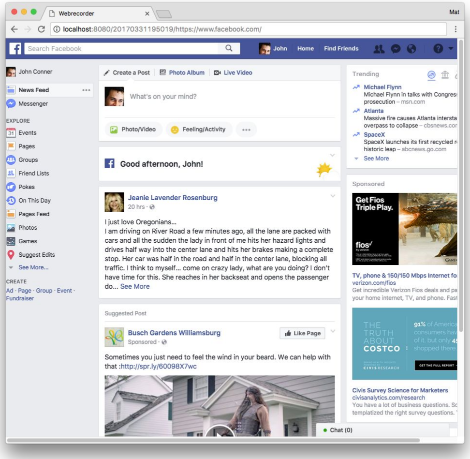
This page isn't available

The link you followed may be broken, or the page may have been removed.

Go back to the previous page · Go to News Feed · Visit our Help Center

Our Web

- Born digital data
- Personalized Content
- Private Content
- Unique content
- Typically not preserved
    - Should it be?



Good afternoon, John!

Jeanie Lavender Rosenburg
20 hrs ·

I just love Oregonians...
I am driving on River Road a few minutes ago, all the lane are packed with cars and all the sudden the lady in front of me hits her hazard lights and drives half way into the center lane and hits her brakes making a complete stop. Her car was half in the road and half in the center lane, blocking all traffic. I think to myself... come on crazy lady, what are you doing? I don't have time for this. She reaches in her backseat and opens the passenger do... See More

Busch Gardens Williamsburg
Sponsored ·

Sometimes you just need to feel the wind in your beard. We can help with that :http://spr.ly/60098X7wc

Michael Flynn
Michael Flynn in talks with Congres
prosecution – msn.com

Atlanta
Massive fire causes Atlanta intersta
overpass to collapse – cbsnews.cor

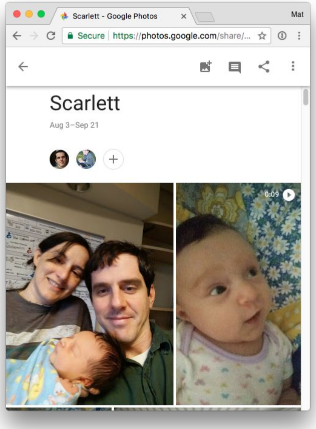
SpaceX
SpaceX launches its first recycled r
historic leap – abcnews.go.com

See More

TV, phone & 150/150 Mbps Internet fo
verizon.com/fios
Get incredible Verizon Fios deals and pa
your home internet, TV, and phone. Fast

THE TRUTH ABOUT COSTCO
91% of America consumers hav of it, but only 4 shopped there
GET THE FULL REPORT
CIVIS RESEARCH

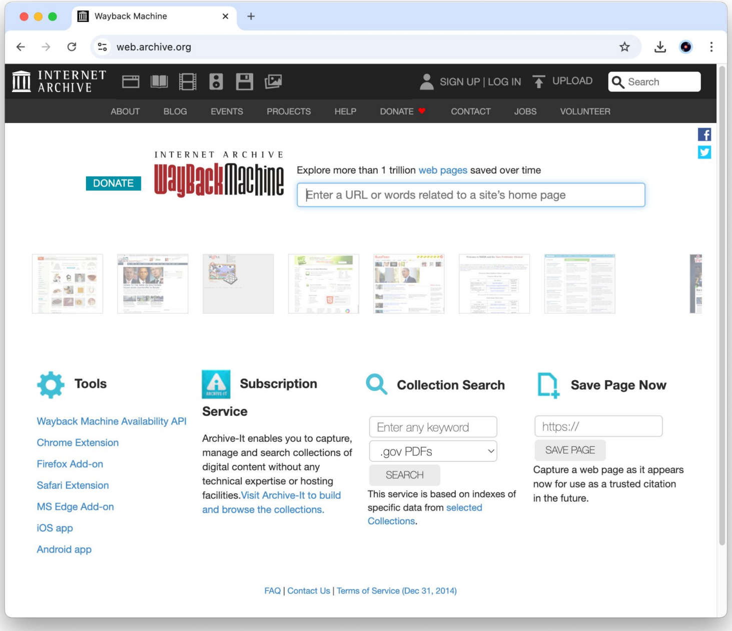
Civis Survey Science for Marketers
civisanalytics.com/research
You have a lot of business questions. Sc
templatized the right survey questions.



But Can We Save It?

The Internet Archive

- Captures the web at 150 TB / day; currently 175 PB of web history (Dec 2025)
- Also: music recordings, radio broadcasts, podcasts, software, images, etc.



Explore more than 1 trillion web pages saved over time

Enter a URL or words related to a site's home page

Tools

Wayback Machine Availability API
Chrome Extension
Firefox Add-on
Safari Extension
MS Edge Add-on
iOS app
Android app

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Collection Search

Enter any keyword
.gov PDFs
SEARCH

This service is based on indexes of specific data from selected Collections.

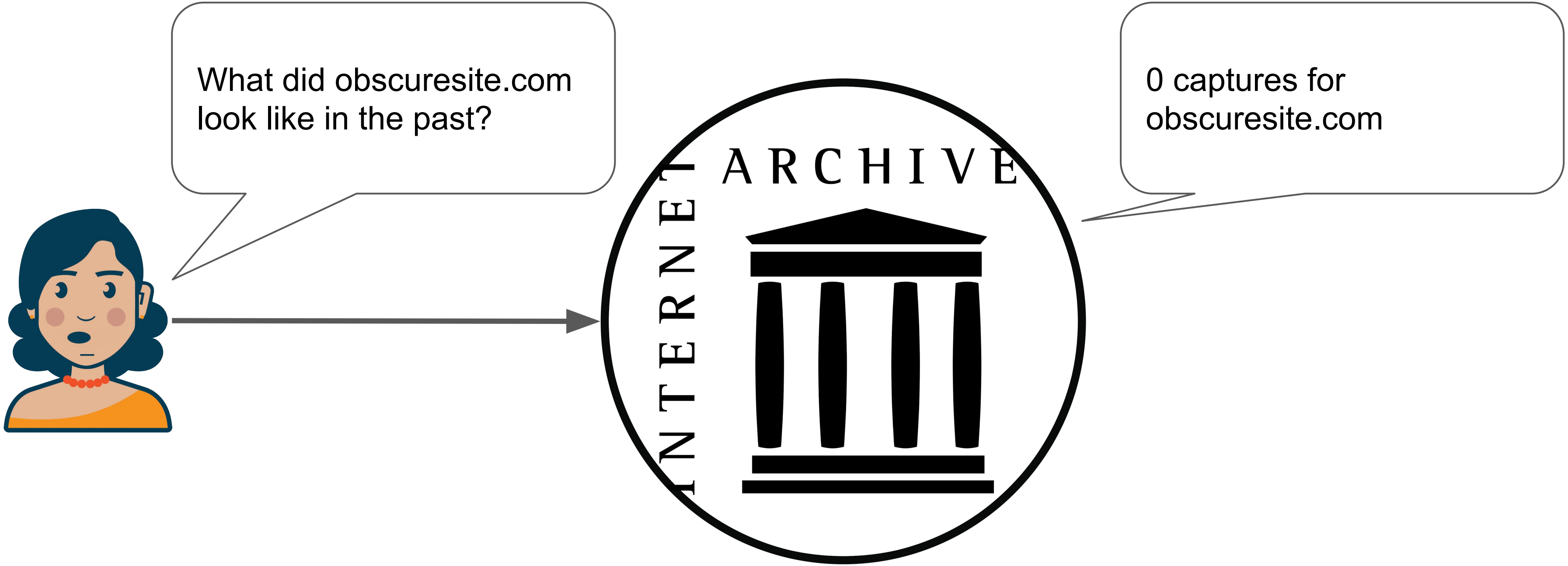
Save Page Now

https://
SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

Not everything is preserved

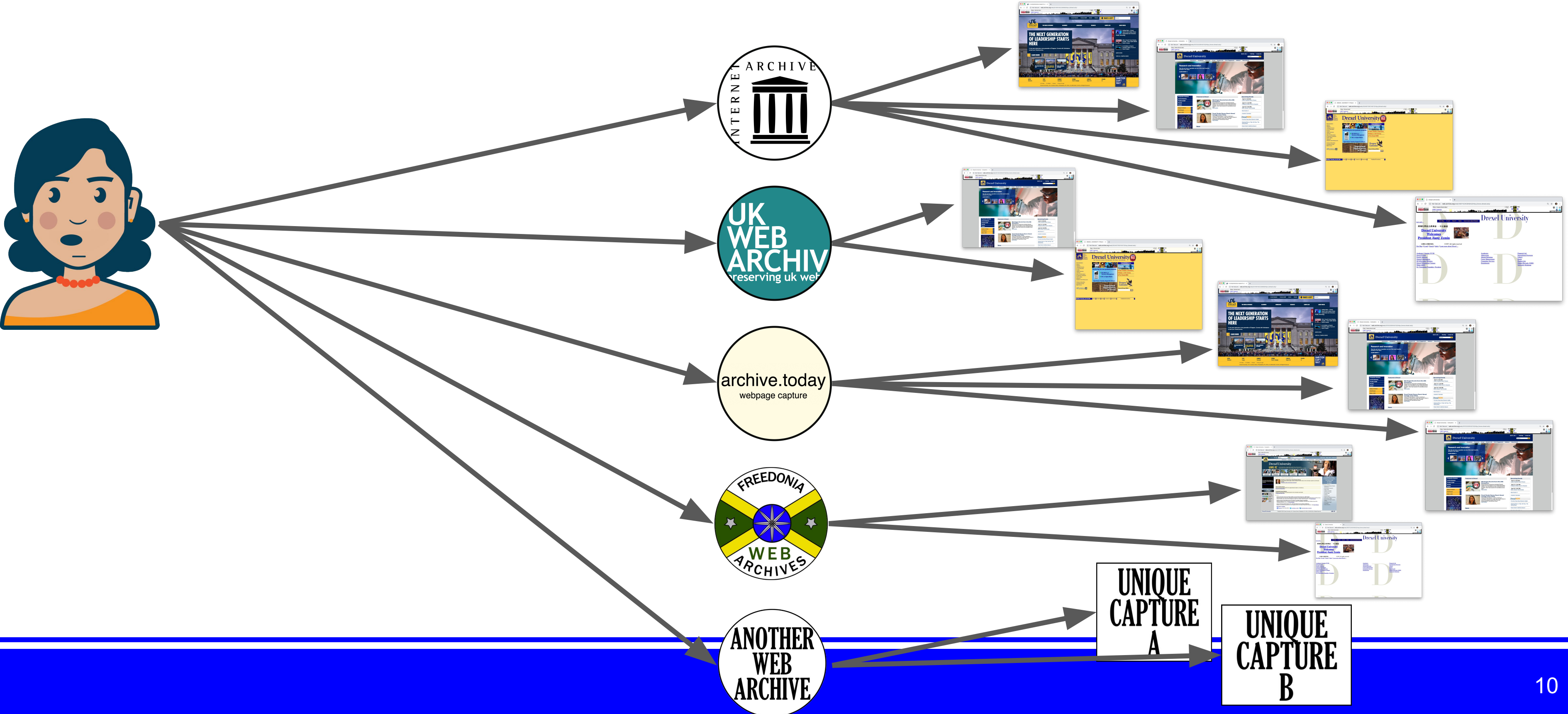


What did obscuresite.com look like in the past?

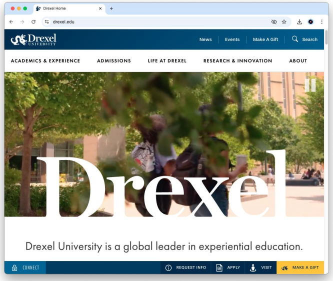
0 captures for obscuresite.com

Multiple archival efforts (3 of many)

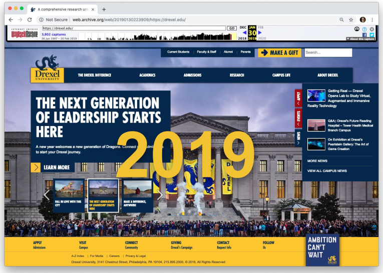
More archives produces a more comprehensive picture

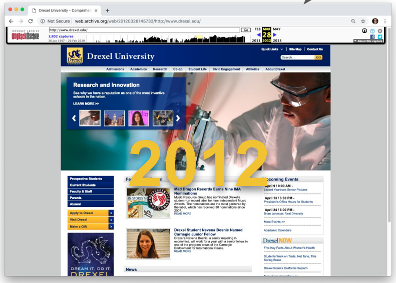


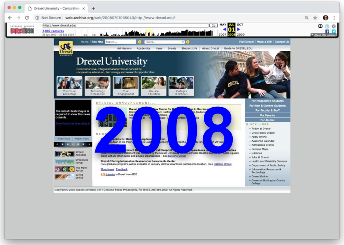
Web Archiving

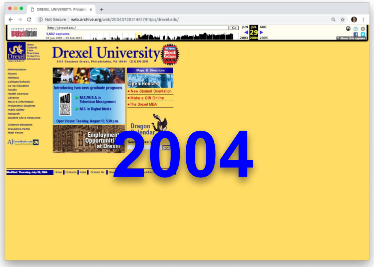


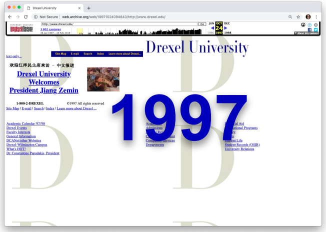
Associate live Web URIs











With their archived representations

Web Archiving: not just “Save As…”

- Metadata
- Provenance
- Standards
- Big Data
- Composite Resources (cf. analog/conventional archiving)
- Technical Nuances
  - Indexing
  - Deduplication
  - Resilience
  - Handling Dynamicism
- Preservation is a race against time — the moment something appears online, it can begin disappearing

Web Archiving Metadata

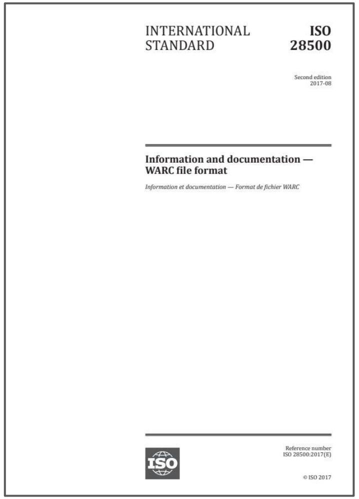
- The how, when and why web content was collected, preserved, and made accessible
- High-level descriptors
- Fields in crawls
- Derivative metadata formats
- Types:
  - Technical (format, checksums)
  - Descriptive (URL, pub date, site title)
  - Preservation (relationship w/ other captures, date/time, crawler used)
  - Administrative (curatorial decision, access restrictions)

Web Archiving Provenance

- Provenance documents where archived content came from, how it was collected, and when.
- Establishes the authenticity and integrity of archived web resources
  - Was it really as depicted? Who did it? When?
- Often realized through metadata
- Essential for reproducibility
  - Others can understand or repeat the capture process
- Reveals biases or gaps
  - what was included, excluded, or technically unreachable
- Clarifies chain of custody when content is transformed, migrated, or redacted over time

Web Archiving Standards

- WARC: ISO 28500:2017
  - Storage format for web archives
  - Container that contains transactional information and metadata
- Memento: IETF RFC7089
  - Provides mechanisms for interacting with the past web



INTERNATIONAL STANDARD

ISO 28500

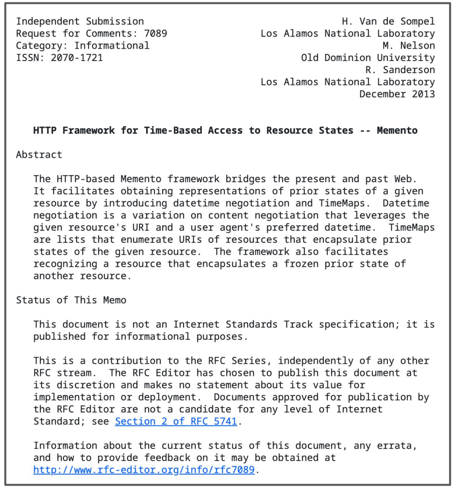
Second edition
2017-08

Information and documentation — WARC file format

Information et documentation — Format de fichier WARC

Reference number
ISO 28500:2017(E)

© ISO 2017

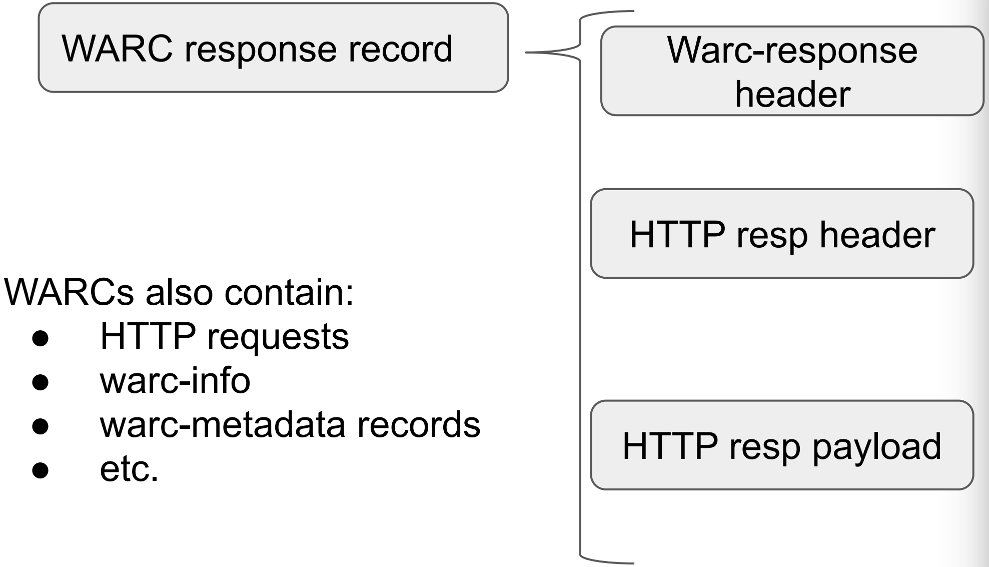


```
Independent Submission                                   H. Van de Sompel
Request for Comments: 7089                 Los Alamos National Laboratory
Category: Informational                                         M. Nelson
ISSN: 2070-1721                                   Old Dominion University
                                                             R. Sanderson
                                           Los Alamos National Laboratory
                                                            December 2013


   HTTP Framework for Time-Based Access to Resource States -- Memento

Abstract

   The HTTP-based Memento framework bridges the present and past Web.
   It facilitates obtaining representations of prior states of a given
   resource by introducing datetime negotiation and TimeMaps.  Datetime
   negotiation is a variation on content negotiation that leverages the
   given resource's URI and a user agent's preferred datetime.  TimeMaps
   are lists that enumerate URIs of resources that encapsulate prior
   states of the given resource.  The framework also facilitates
   recognizing a resource that encapsulates a frozen prior state of
   another resource.

Status of This Memo

   This document is not an Internet Standards Track specification; it is
   published for informational purposes.

   This is a contribution to the RFC Series, independently of any other
   RFC stream.  The RFC Editor has chosen to publish this document at
   its discretion and makes no statement about its value for
   implementation or deployment.  Documents approved for publication by
   the RFC Editor are not a candidate for any level of Internet
   Standard; see Section 2 of RFC 5741.

   Information about the current status of this document, any errata,
   and how to provide feedback on it may be obtained at
   http://www.rfc-editor.org/info/rfc7089.
```

Web Archiving Standards: WARC

- ISO-standardized file format (ISO 28500) used for storing web crawls and related digital content
- Uses records (e.g., response, request, metadata, revisit, resource) to represent different capture types
- Preserves original HTTP request/response pairs, supporting authenticity and replay
- Designed for long-term digital preservation: stable, extensible, and format-agnostic

WARC Anatomy



WARC response record

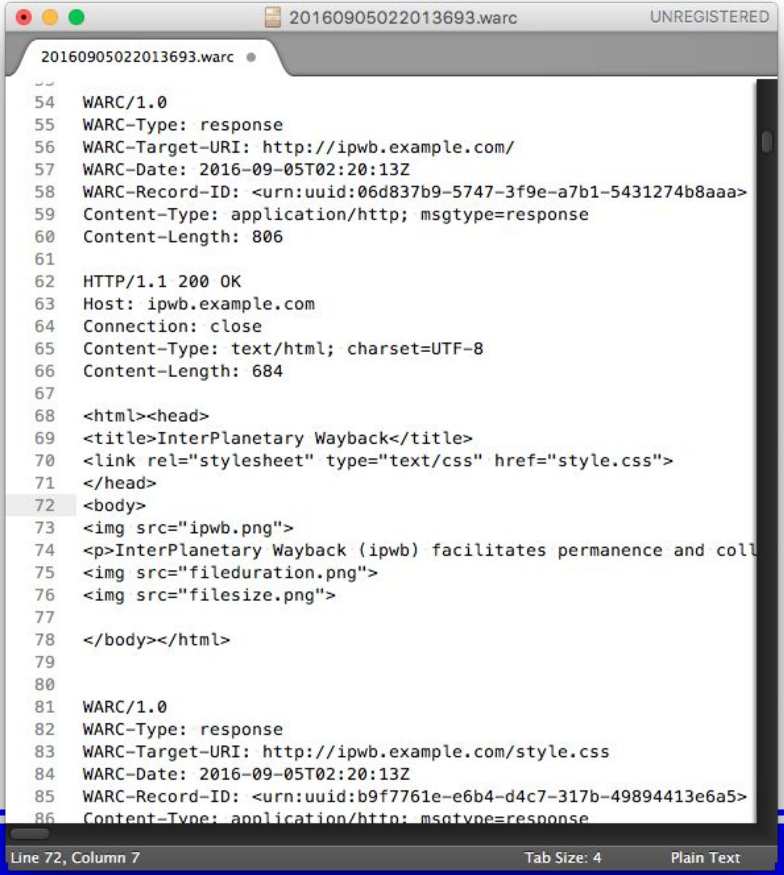
Warc-response header

HTTP resp header

HTTP resp payload

WARCs also contain:

- HTTP requests
- warc-info
- warc-metadata records
- etc.



```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://ipwb.example.com/
WARC-Date: 2016-09-05T02:20:13Z
WARC-Record-ID: <urn:uuid:06d837b9-5747-3f9e-a7b1-5431274b8aaa>
Content-Type: application/http; msgtype=response
Content-Length: 806

HTTP/1.1 200 OK
Host: ipwb.example.com
Connection: close
Content-Type: text/html; charset=UTF-8
Content-Length: 684

<html><head>
<title>InterPlanetary Wayback</title>
<link rel="stylesheet" type="text/css" href="style.css">
</head>
<body>
<img src="ipwb.png">
<p>InterPlanetary Wayback (ipwb) facilitates permanence and coll
<img src="fileduration.png">
<img src="filesize.png">

</body></html>


WARC/1.0
WARC-Type: response
WARC-Target-URI: http://ipwb.example.com/style.css
WARC-Date: 2016-09-05T02:20:13Z
WARC-Record-ID: <urn:uuid:b9f7761e-e6b4-d4c7-317b-49894413e6a5>
Content-Type: application/http; msgtype=response
```

Web Archiving Standards: Memento

- Standardizes semantics and syntax for time on the web
- Content negotiation in the temporal dimension
- IETF RFC7089 (standard)

URI-R

Original Resource
Resource that exists or used to exist; we are interested in a past state of it

`Link: <`**`URI-R`**`>;type="`**`original`**`"`
points to Original Resource

URI-M

Memento
Resource that encapsulates a past state of the Original Resource

`Link: <`**`URI-M`**`>;type="`**`memento`**`"`
points to Memento

URI-G

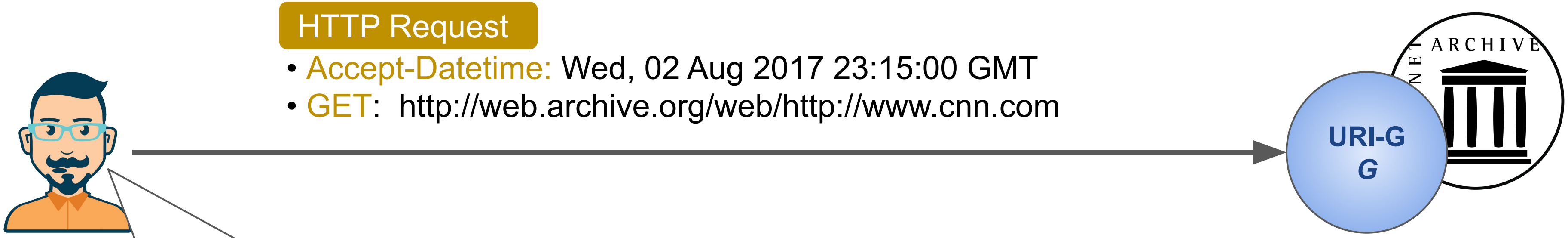
TimeGate
Resource that "decides", based on a given datetime, which is the temporally best Memento for an Original Resource

`Link: <`**`URI-TG`**`>;type="`**`timegate`**`"`
points to TimeGate

Background: Memento Request Example

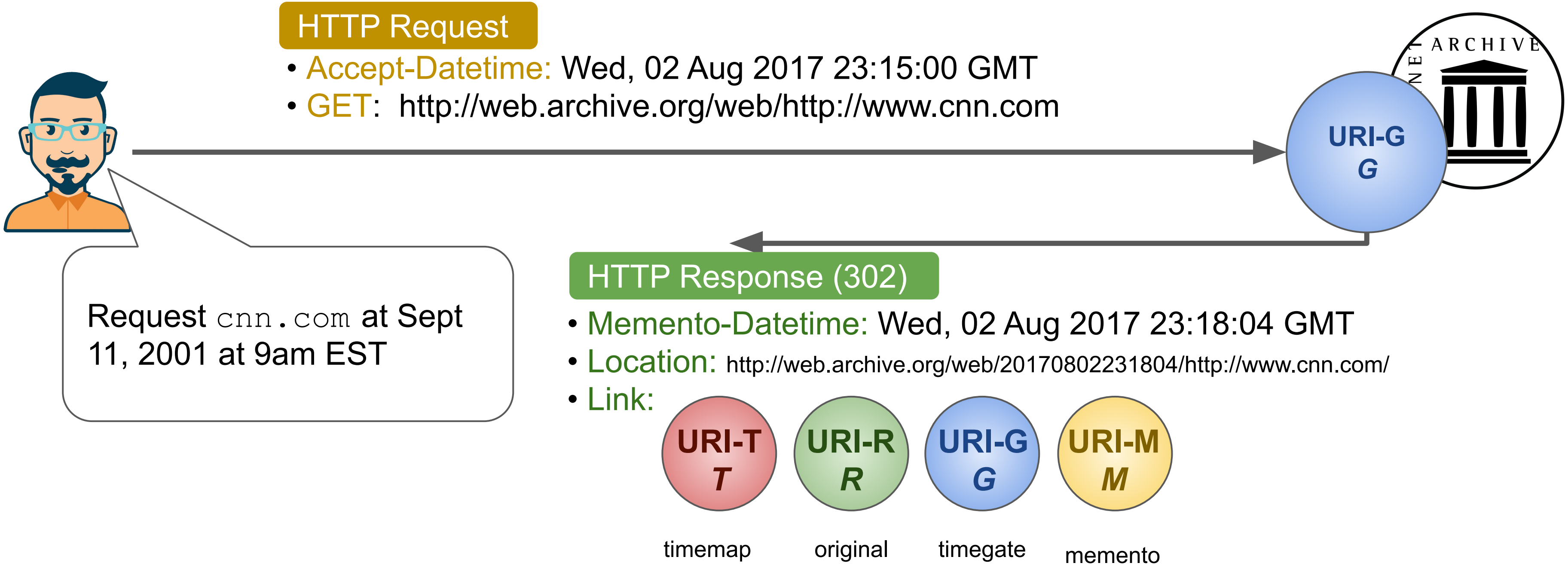
HTTP Request

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



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

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` at Sept 11, 2001 at 9am EST

HTTP Response (302)

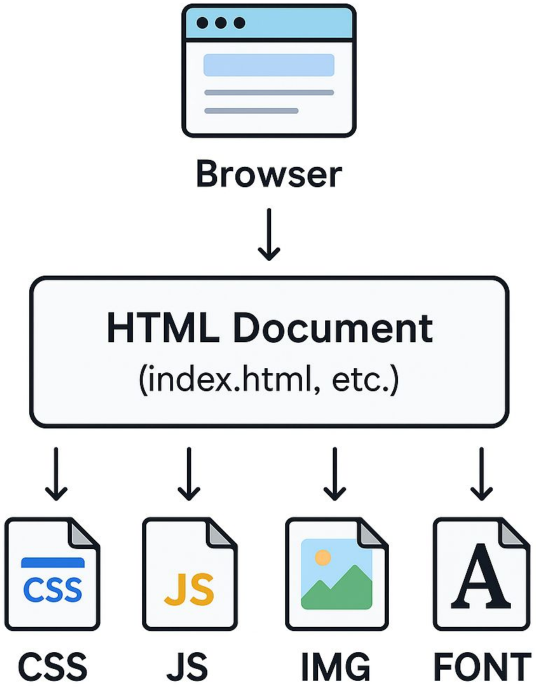
- Memento-Datetime: Wed, 02 Aug 2017 23:18:04 GMT
- Location: http://web.archive.org/web/20170802231804/http://www.cnn.com/
- Link:

Web Archiving (Big) Data

- Entire sites, documents, media, files
- Massive, complex, heterogeneous datasets, often measured in petabytes.
- Volume (billions of URLs harvested, longitudinal captures over the years)
- Variety (structured, semi-structured, unstructured content; media, scripts)
- Velocity (continuous capture of rapidly changing web content)

Composite Resources

- HTML → CSS, images, media → deeper embedded resources
- Content may appear dynamically based on user interaction

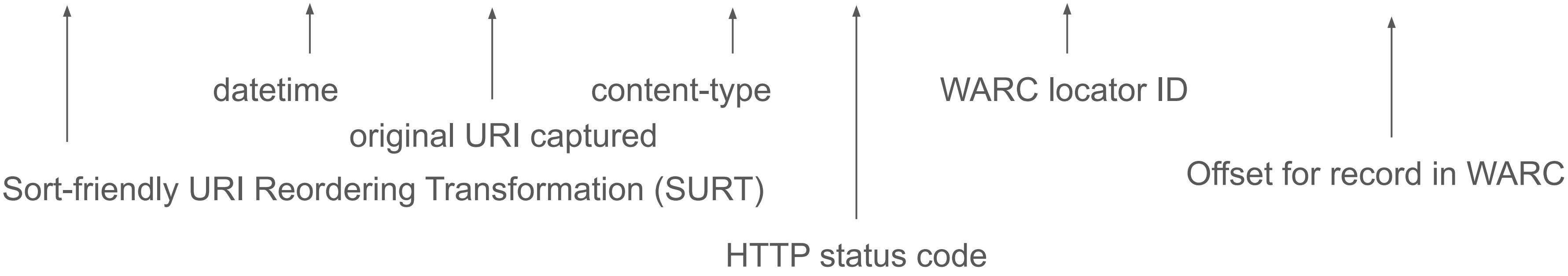


Indexing WARCs

- Fast lookup of a URL w/o scanning terabytes of WARC data
- Reply systems (e.g., Wayback, pywb) for reconstructing a page quickly
- Efficient queries over billions of archived documents
- Without indexing, random access is impossible
- Single WARC file can hold thousands of individual resources
- Index formats:
    - CDX (classic, well-defined fields)
    - CDXJ (modern, more flexible, supports extra metadata fields)
    - Database indexes (performance at scale, full-text search and complex queries)

Index Examples

```
edu,drexel)/ 19970626040823 http://www.drexel.edu:80/ text/html 200 NA4IS3YV4BOPOE3KRG6VMNQLFFTZMWJG 1814
edu,drexel)/ 20220214060114 https://www.drexel.edu/ text/html 301 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ 459
edu,drexel)/ 20230709040752 https://drexel.edu/ text/html 200 XMW7AUSKNSWO2WMOML3TH2I3HVRVVHGD 30966
edu,drexel)/ 20251005193011 https://drexel.edu/ text/html 200 H5GRERCRQSZOWAEUV5APC7U67W2ZO7LQ 32444
```



datetime

content-type

WARC locator ID

original URI captured

Offset for record in WARC

Sort-friendly URI Reordering Transformation (SURT)

HTTP status code

Deduplication

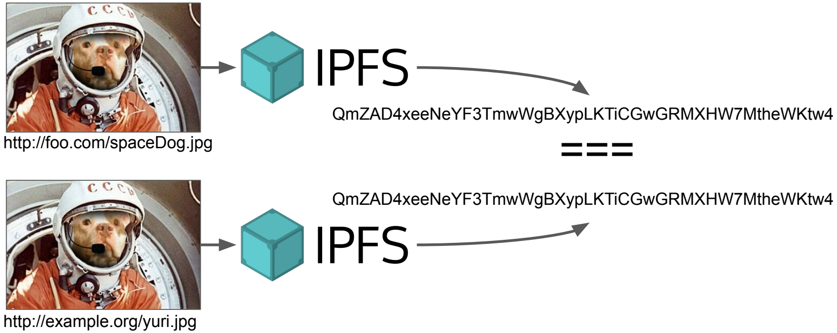
- Identify and handle duplicate or near-duplicate content
- Storage reduced, replay accurate, efficient indexing, avoid unnecessary redundancy
- Faster crawling, more efficient indexing, improved data quality with reduced noise, better preservation integrity
- Can be performed at crawl time or during replay
- Types:
    - URL-level (same page at different URLs)
    - Payload (identical content at different hosts)
    - Temporal (repeated captures unchanged over time)
    - Structural (template pages w/ minimal diffs)

Screenshots of apple.com of the past

Resilience of Web Archives

- Wholesale or piecemeal copies of archives with means to check integrity
- How can you verify what is being purported as history is truly representative?
- Does authenticity matter if used for personal reference?

ipwb
interplanetary wayback


IPFS
QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4
http://foo.com/spaceDog.jpg
===
QmZAD4xeeNeYF3TmwWgBXypLKTiCGwGRMXHW7MtheWKtw4
IPFS
http://example.org/yuri.jpg

https://github.com/oduwsdl/ipwb

Handling Dynamicism

- The delta between a crawler and a browser causes resources to be missed
- JavaScript-driven, dynamic sites are hard to capture.

On the Change in Archivability of Websites Over Time

Mat Kelly, Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA, 23529, USA
{mkelly,jbrunelle,mweigle,mln}@cs.odu.edu

Abstract. As web technologies evolve, web archivists work to keep up so that our digital history is preserved. Recent advances in web technologies have introduced client-side executed scripts that load data without a referential identifier or that require user interaction (e.g., content loading when the page has scrolled). These advances have made automating methods for capturing web pages more difficult. Because of the evolving schemes of publishing web pages along with the progressive capability of web preservation tools, the *archivability* of pages on the web has varied over time. In this paper we show that the archivability of a web page can be deduced from the type of page being archived, which aligns with that page's accessibility in respect to dynamic content. We show concrete examples of when these technologies were introduced by referencing mementos of pages that have persisted through a long evolution of available technologies. Identifying these reasons for the inability of these web pages to be archived in the past in respect to accessibility serves as a guide for ensuring that content that has longevity is published using good practice methods that make it available for preservation.

Keywords: Web Archiving, Digital Preservation

1 Introduction

The web has gone through a gradient yet demarcated series of phases in which interactivity has become more fluid to the end-user. Early websites were static. Adoption of JavaScript allowed the components on a web page to respond to users' actions or be manipulated in ways that made the page more usable. Ajax [9] combines multiple web technologies to give web pages the ability to perform operations asynchronously. The adoption of Ajax by web developers facilitated the fluidity of user interaction on the web. Through each phase in the progression of the web, the ability to preserve the content displayed to the user has also progressed but in a less linear trend.

A large amount of the difficulty in web archiving stems from the crawler's insufficient ability to capture content related to JavaScript. Because JavaScript is executed on the client side (i.e., within the browser after the page has loaded), it should follow that the archivability could be evaluated using a consistent replay medium. The medium used to archive (normally a web crawler tailored for archiving, e.g., Heritrix [21]) is frequently different from the medium used to replay the archive (henceforth, the *web browser*, the predominant means of

doi:10.1007/978-3-642-40501-3_5

Int J Digit Libr (2016) 17:95–117
DOI 10.1007/s00799-015-0140-8

CrossMark

The impact of JavaScript on archivability

Justin F. Brunelle · Mat Kelly · Michele C. Weigle · Michael L. Nelson

Received: 7 November 2013 / Revised: 12 January 2015 / Accepted: 14 January 2015 / Published online: 25 January 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract As web technologies evolve, web archivists work to adapt so that digital history is preserved. Recent advances in web technologies have introduced client-side executed scripts (Ajax) that, for example, load data without a change in top level Universal Resource Identifier (URI) or require user interaction (e.g., content loading via Ajax when the page has scrolled). These advances have made automating methods for capturing web pages more difficult. In an effort to understand why mementos (archived versions of live resources) in today's archives vary in completeness and sometimes pull content from the live web, we present a study of web resources and archival tools. We used a collection of URIs shared over Twitter and a collection of URIs curated by Archive-It in our investigation. We created local archived versions of the URIs from the Twitter and Archive-It sets using WebCite, wget, and the Heritrix crawler. We found that only 4.2 % of the Twitter collection is perfectly archived by all of these tools, while 34.2 % of the Archive-It collection is perfectly archived. After studying the quality of these mementos, we identified the practice of loading resources via JavaScript (Ajax) as the source of archival difficulty. Further, we show that resources are increasing their use of JavaScript to load embedded resources. By 2012, over half (54.5 %) of pages use JavaScript to load embedded resources. The number of embedded resources loaded via JavaScript has increased by 12.0 % from 2005 to 2012. We also show that JavaScript is responsible for 33.2 % more missing resources in 2012 than in 2005. This shows that JavaScript is responsible for an increasing proportion of the embedded resources unsuccessfully loaded by mementos. JavaScript is also responsible for 52.7 % of all missing embedded resources in our study.

Keywords Web architecture · Web archiving · Digital preservation

1 Introduction

How well can we archive the web? This is a question that is becoming increasingly important and more difficult to answer. Additionally, this question has significant impact on web users [40,43] and commercial and government compliance [52,53,66].

The web has gone through a gradient of changes fueled by increasing user demand for interactivity. Early websites were relatively static, while continued adoption of web technologies has made the pages personalized and more interactive. JavaScript, which executes on the client, provides additional features for the web user, enabling or increasing interactivity, client-side state changes, and personalized representations. These additional features offer an enhanced browsing experience for the user.

JavaScript has enabled a wide-scale migration from web *pages* to web *applications*. This migration continued with the introduction of Ajax (first introduced in 2005 [23]), which combined multiple technologies to give web pages the ability to perform asynchronous client–server interactions after the HTML is loaded. The first wide-scale implementation of Ajax was in Google Maps in 2005, but Ajax was officially added as a standard in 2006 [70]. While archival tools per-

J. F. Brunelle (✉) · M. Kelly · M. C. Weigle · M. L. Nelson
Department of Computer Science, Old Dominion University,
Norfolk, VA 23529, USA
e-mail: jbrunelle@cs.odu.edu

M. Kelly
e-mail: mkelly@cs.odu.edu

M. C. Weigle
e-mail: mweigle@cs.odu.edu

M. L. Nelson
e-mail: mln@cs.odu.edu

Springer

doi:10.1007/s00799-015-0140-8

The Archival Acid Test: Evaluating Archive Performance on Advanced HTML and JavaScript

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

When preserving web pages, archival crawlers sometimes produce a result that varies from what an end-user expects. To quantitatively evaluate the degree to which an archival crawler is capable of comprehensively reproducing a web page from the live web into the archives, the crawlers' capabilities must be evaluated. In this paper, we propose a set of metrics to evaluate the capability of archival crawlers and other preservation tools using the Acid Test concept. For a variety of web preservation tools, we examine previous captures within web archives and note the features that produce incomplete or unexpected results. From there, we design the test to produce a quantitative measure of how well each tool performs its task.

Categories and Subject Descriptors

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

General Terms

Experimentation, Standardization, Verification

Keywords

Web Crawler, Web Archiving, Digital Preservation

1. INTRODUCTION

Since much of our cultural discourse occurs on the web, web archiving is necessary for posterity. The goal of web archiving is to capture web pages so they can be "replayed" at a later date. Web archiving tools access these pages on the live web in a manner similar to tools used by search engines (crawlers) and preserve the pages in a format that allows the data and contextual information about the crawl to be re-experienced. These "archival crawlers" take different approaches in digital preservation and thus their capability and scope vary.

Because archival crawlers attempt to duplicate what a user would see if he accessed the page on the live web, variance from what is preserved and what would have been seen compromises the integrity of the archive. The functional difference between archival crawlers and web browsers causes this sort of unavoidable discrepancy in the archives, but it is difficult to evaluate how good of a job the crawler did if the information no longer exists on the live web. By examining what sort of web content is inaccurately represented or missing from the web archives, it would be useful to evaluate the capability of archival crawlers (in respect to that of web browsers that implement the latest technologies) to determine what might be missing from their functional repertoire.

Web browsers exhibited this deviation between each other in the early days of Web Standards. A series of "Acid Tests" that implemented the Web Standards allowed each browser to visually and functionally render a web page and produce an evaluation of how well the browser conformed to the standards. In much the same way, we have created an "Archival Acid Test" to implement features of web browsers in a web page. While all standards-compliant browsers will correctly render the live page, this is not always the case when the archived version of the page is rendered. This difference can be used to highlight the features that archival crawlers are lacking compared to web browsers and thus emphasize the deviations that will occur in web archives compared to what a user would expect from a digitally preserved web page.

2. RELATED WORK

Web archives are generated by a variety of tools in a variety of formats. An ISO standard format utilized by institutional and personal web archivists alike is the Web ARChive (WARC) format [1]. WARC files allow HTTP communication that occurred during a crawl as well as payload, metadata and other archival features to be encoded in a single or an extensibly defined set of WARC files.

Heritrix paved the way for Internet Archive (IA) to utilize their open source Heritrix to create ARC and WARC files from web crawls while capturing all resources necessary to replay a web page [2]. Other tools have since added WARC creation functionality [3, 4, 5]. Multiple software platforms exist that can replay WARCs but IA's Wayback Machine (and its open source counterpart[1]) is the de facto standard.

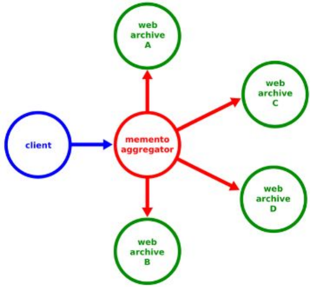
Multiple services exist that allow users to submit URIs for preservation. IA recently began offering a "Save Page Now" feature co-located with their web archive browsing inter-

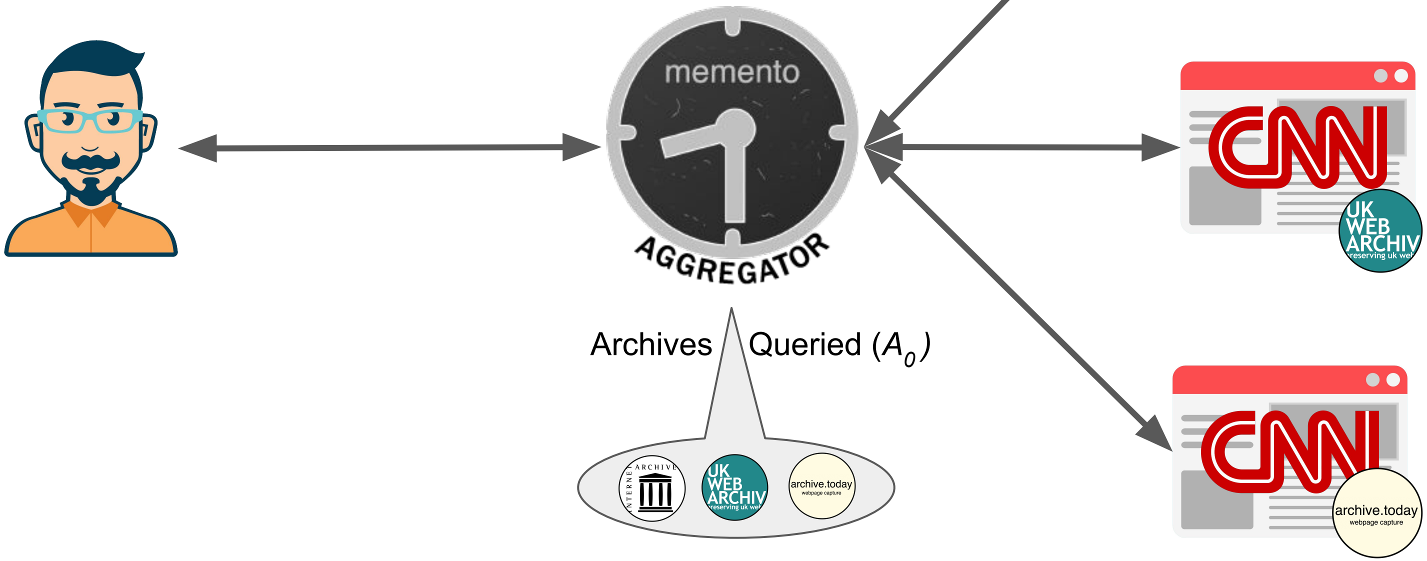
[1] https://github.com/iipc/openwayback

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

doi:10.1109/JCDL.2014.6970146

Web Archive Aggregation

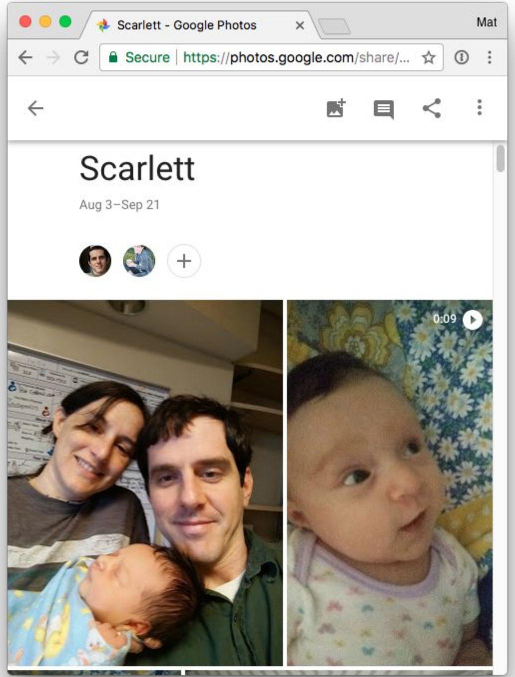


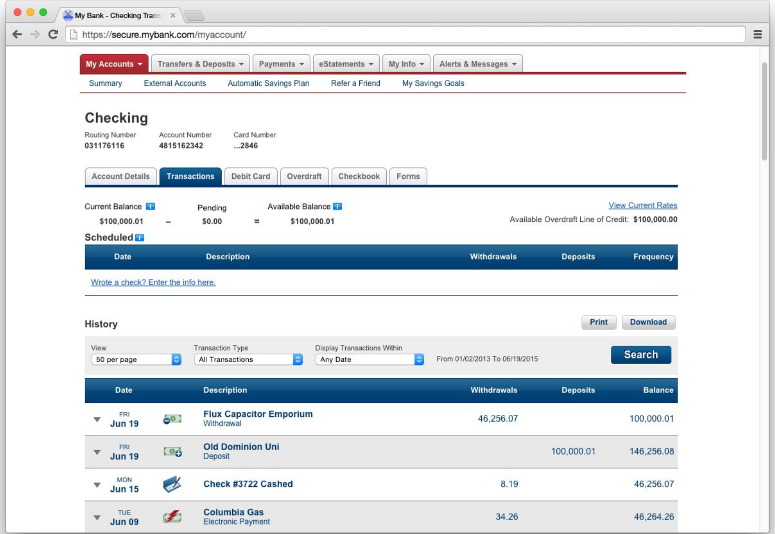


Archives Queried (A_0)

Where Mat Fits

You Are Responsible for Saving What’s Important





Checking

Routing Number	Account Number	Card Number
031176116	4815162342	...2846

Current Balance		Pending		Available Balance
$100,000.01	–	$0.00	=	$100,000.01

View Current Rates

Available Overdraft Line of Credit: $100,000.00

Scheduled

Date	Description	Withdrawals	Deposits	Frequency

Wrote a check? Enter the info here.

History

Print Download

View: 50 per page · Transaction Type: All Transactions · Display Transactions Within: Any Date · From 01/02/2013 To 06/19/2015 · Search

Date	Description	Withdrawals	Deposits	Balance
FRI Jun 19	Flux Capacitor Emporium Withdrawal	46,256.07		100,000.01
FRI Jun 19	Old Dominion Uni Deposit		100,000.01	146,256.08
MON Jun 15	Check #3722 Cashed	8.19		46,256.07
TUE Jun 09	Columbia Gas Electronic Payment	34.26		46,264.26

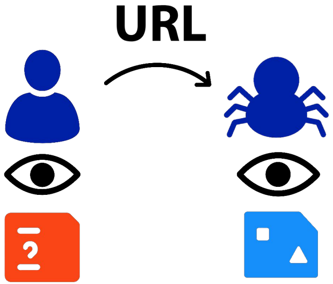
Most Web Archiving Tools Cater to Archiving the Public Web

- The face of society, Zeitgeist
- Easier to vet – captures from multiple institutions give legitimacy of the record
- Institutional self-interest
  - e.g., Stanford University Archives, UK Web Archive

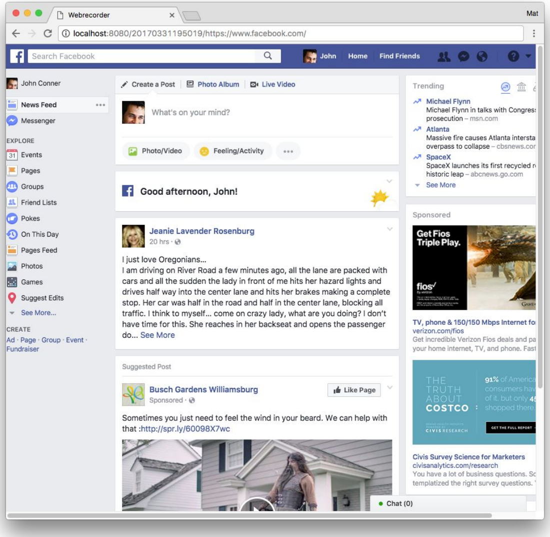
UKWA
UK WEB ARCHIVE

Stanford | University Libraries
Stanford Web Archive Portal

The Fallacious Key-Based Approach



- A URL is not enough
- Personalized Representations
- Dynamic web sites
- Context lost



localhost:8080/20170331195019/https://www.facebook.com/

Good afternoon, John!

Jeanie Lavender Rosenburg
20 hrs ·

I just love Oregonians...
I am driving on River Road a few minutes ago, all the lane are packed with cars and all the sudden the lady in front of me hits her hazard lights and drives half way into the center lane and hits her brakes making a complete stop. Her car was half in the road and half in the center lane, blocking all traffic. I think to myself... come on crazy lady, what are you doing? I don't have time for this. She reaches in her backseat and opens the passenger do... See More

Suggested Post

Busch Gardens Williamsburg
Sponsored ·

Sometimes you just need to feel the wind in your beard. We can help with that :http://spr.ly/60098X7wc

Trending

Michael Flynn
Michael Flynn in talks with Congres prosecution – msn.com

Atlanta
Massive fire causes Atlanta intersta overpass to collapse – cbsnews.cor

SpaceX
SpaceX launches its first recycled r historic leap – abcnews.go.com

See More

Sponsored

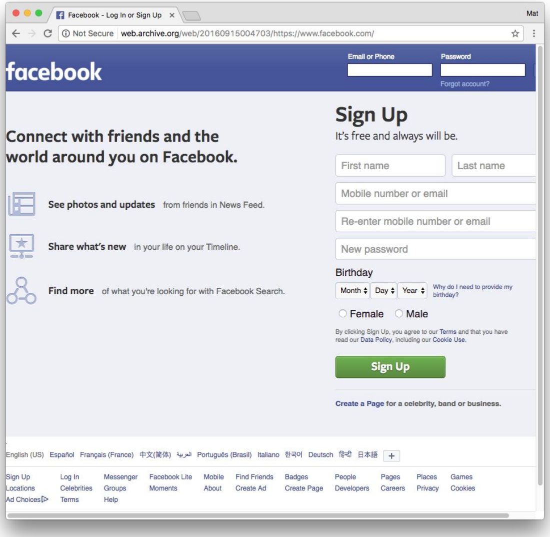
Get Fios Triple Play.

TV, phone & 150/150 Mbps Internet fo
verizon.com/fios
Get incredible Verizon Fios deals and pa your home internet, TV, and phone. Fast

THE TRUTH ABOUT COSTCO : 91% of America consumers hav of it, but only 4 shopped there.

Civis Survey Science for Marketers
civisanalytics.com/research
You have a lot of business questions. Sc templatized the right survey questions.

facebook.com capture with browser-based archiving tools



web.archive.org/web/20160915004703/https://www.facebook.com/

Connect with friends and the world around you on Facebook.

See photos and updates from friends in News Feed.

Share what's new in your life on your Timeline.

Find more of what you're looking for with Facebook Search.

Sign Up

It's free and always will be.

Birthday

Why do I need to provide my birthday?

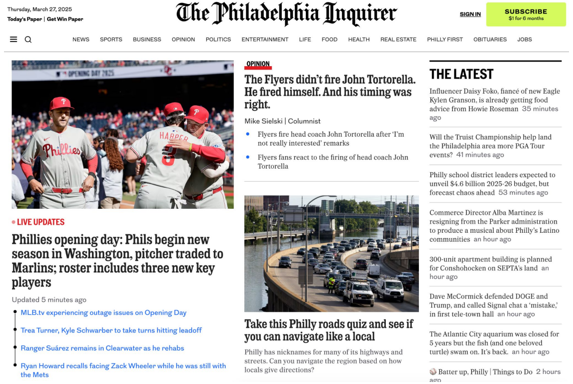
By clicking Sign Up, you agree to our Terms and that you have read our Data Policy, including our Cookie Use.

Create a Page for a celebrity, band or business.

facebook.com capture from Internet Archive

Can We Save The Web We See From Our Perspective?

- Repurpose user's daily driver profile as crawler basis
- Permutate attributes of a user to represent a "persona", producing a web experience closer to that of an actual web user cf. crawler
- Avoid clean slate crawling and delegation to a user-agnostic crawler
- Scale?



Thursday, March 27, 2025
Today's Paper | Get Win Paper

The Philadelphia Inquirer

● LIVE UPDATES

Phillies opening day: Phils begin new season in Washington, pitcher traded to Marlins; roster includes three new key players

Updated 5 minutes ago

- MLB.tv experiencing outage issues on Opening Day
- Trea Turner, Kyle Schwarber to take turns hitting leadoff
- Ranger Suárez remains in Clearwater as he rehabs
- Ryan Howard recalls facing Zack Wheeler while he was still with the Mets

OPINION

The Flyers didn't fire John Tortorella. He fired himself. And his timing was right.

Mike Sielski | Columnist

- Flyers fire head coach John Tortorella after 'I'm not really interested' remarks
- Flyers fans react to the firing of head coach John Tortorella

Take this Philly roads quiz and see if you can navigate like a local

Philly has nicknames for many of its highways and streets. Can you navigate the region based on how locals give directions?

THE LATEST

Influencer Daisy Foko, fiancé of new Eagle Kylen Granson, is already getting food advice from Howie Roseman 35 minutes ago

Will the Truist Championship help land the Philadelphia area more PGA Tour events? 41 minutes ago

Philly school district leaders expected to unveil $4.6 billion 2025-26 budget, but forecast chaos ahead 53 minutes ago

Commerce Director Alba Martinez is resigning from the Parker administration to produce a musical about Philly's Latino communities an hour ago

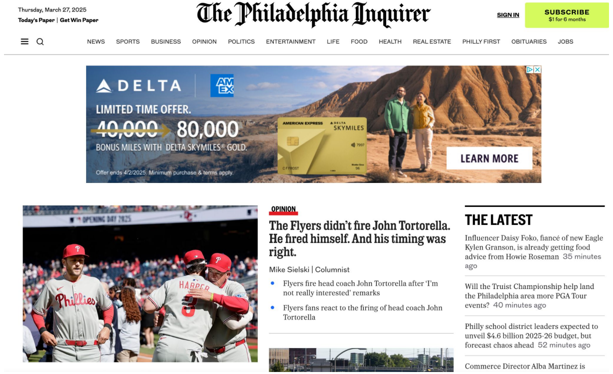
300-unit apartment building is planned for Conshohocken on SEPTA's land an hour ago

Dave McCormick defended DOGE and Trump, and called Signal chat a 'mistake,' in first tele-town hall an hour ago

The Atlantic City aquarium was closed for 5 years but the fish (and one beloved turtle) swam on. It's back. an hour ago

⚾ Batter up, Philly | Things to Do 2 hours ago

vs



Thursday, March 27, 2025
Today's Paper | Get Win Paper

The Philadelphia Inquirer

OPINION

The Flyers didn't fire John Tortorella. He fired himself. And his timing was right.

Mike Sielski | Columnist

- Flyers fire head coach John Tortorella after 'I'm not really interested' remarks
- Flyers fans react to the firing of head coach John Tortorella

THE LATEST

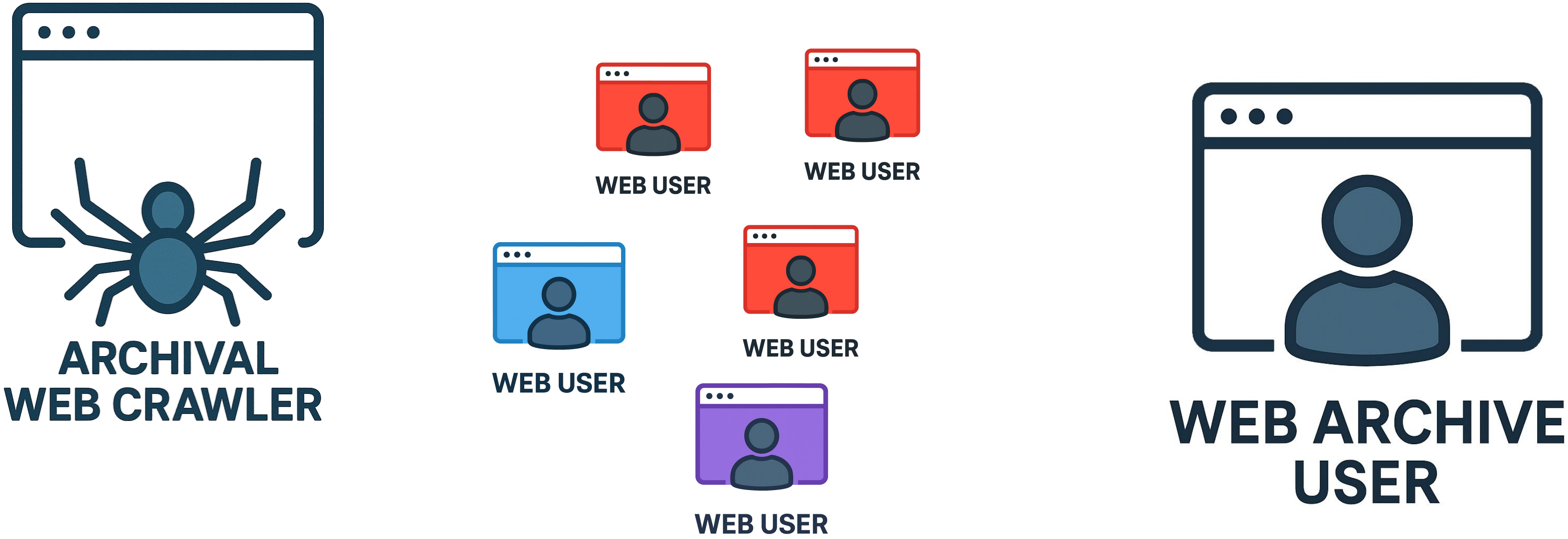
Influencer Daisy Foko, fiancé of new Eagle Kylen Granson, is already getting food advice from Howie Roseman 35 minutes ago

Will the Truist Championship help land the Philadelphia area more PGA Tour events? 40 minutes ago

Philly school district leaders expected to unveil $4.6 billion 2025-26 budget, but forecast chaos ahead 52 minutes ago

Commerce Director Alba Martinez is

The Past Web Saved Is Not The Web That Was



ARCHIVAL WEB CRAWLER

WEB USER

WEB USER

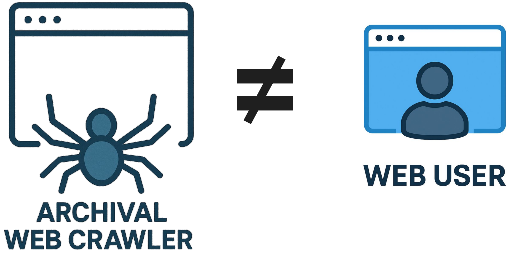
WEB USER

WEB USER

WEB USER

WEB ARCHIVE USER

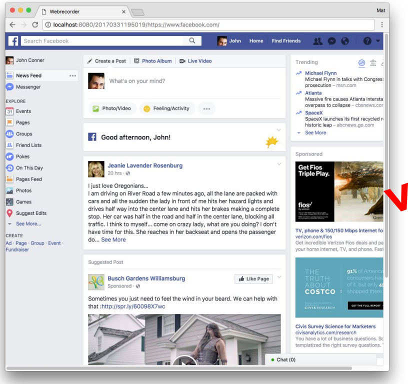
Crawlers Preserve A False Web (Premise)



- Archival crawlers preserve a version of the web inconsistent with web users' experience, a web that actually wasn't
- Customization, personalization based on user history is not *canonical*
- Crawlers (rightfully) see a clean/agnostic version of web sites, devoid of any individuals' experience, PII
- Ergo, what crawlers preserve is a version of the web inconsistent with what a user would have seen at that time
- False history? Nature of experience

By-Value Archiving with Posthoc Metadata Ascription

- Most true to the original form
- Capture complex content
- Personalized representations archivable
- Easier to fabricate content (not a good thing)
    - Requires a means of vetting authenticity



Good afternoon, John!

Jeanie Lavender Rosenburg
20 hrs ·

I just love Oregonians...
I am driving on River Road a few minutes ago, all the lane are packed with cars and all the sudden the lady in front of me hits her hazard lights and drives half way into the center lane and hits her brakes making a complete stop. Her car was half in the road and half in the center lane, blocking all traffic. I think to myself... come on crazy lady, what are you doing? I don't have time for this. She reaches in her backseat and opens the passenger do... See More

Busch Gardens Williamsburg
Sponsored ·

Sometimes you just need to feel the wind in your beard. We can help with that :http://spr.ly/60098X7wc

Michael Flynn
Michael Flynn in talks with Congres prosecution – msn.com

Atlanta
Massive fire causes Atlanta intersta overpass to collapse – cbsnews.cor

SpaceX
SpaceX launches its first recycled r historic leap – abcnews.go.com

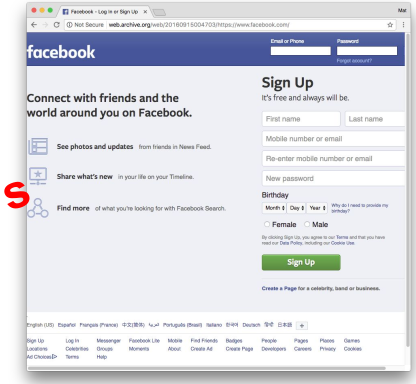
See More

TV, phone & 150/150 Mbps Internet fo
verizon.com/fios
Get incredible Verizon Fios deals and pa your home internet, TV, and phone. Fast

91% of America consumers hav of it, but only 4 shopped there.

Civis Survey Science for Marketers
civisanalytics.com/research
You have a lot of business questions. So templatized the right survey questions.

vs



Connect with friends and the world around you on Facebook.

See photos and updates from friends in News Feed.

Share what's new in your life on your Timeline.

Find more of what you're looking for with Facebook Search.

Sign Up
It's free and always will be.

Birthday

○ Female ○ Male

By clicking Sign Up, you agree to our Terms and that you have read our Data Policy, including our Cookie Use.

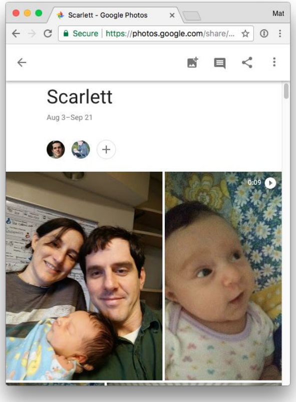
Create a Page for a celebrity, band or business.

NATIONAL ENDOWMENT FOR THE HUMANITIES

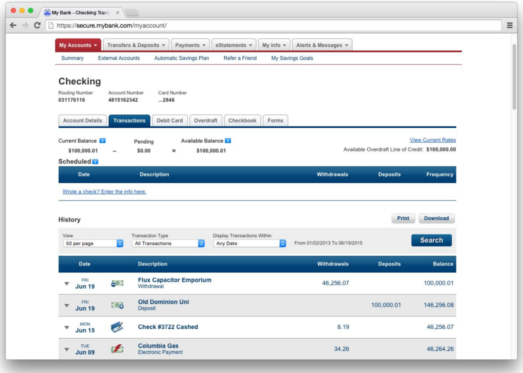
"Archive What I See Now: Bringing Institutional Web Archiving Tools to the Individual Researcher", National Endowment for the Humanities Digital Humanities Implementation Grant (HK-50181-14)

Archiving Private and Personal Web Content

- The things we care about may not belong in an institutional web archiving
- Bearer of our own important information
- Stewardship, ownership, copyright
- Integration into conventional public web archive corpora



Historical capture of Google Photos



Checking

Routing Number 031176116 · Account Number 4815162342 · Card Number ...2846

Current Balance $100,000.01 – Pending $0.00 = Available Balance $100,000.01

View Current Rates

Available Overdraft Line of Credit: $100,000.00

Scheduled

Date	Description	Withdrawals	Deposits	Frequency
	Wrote a check? Enter the info here.			

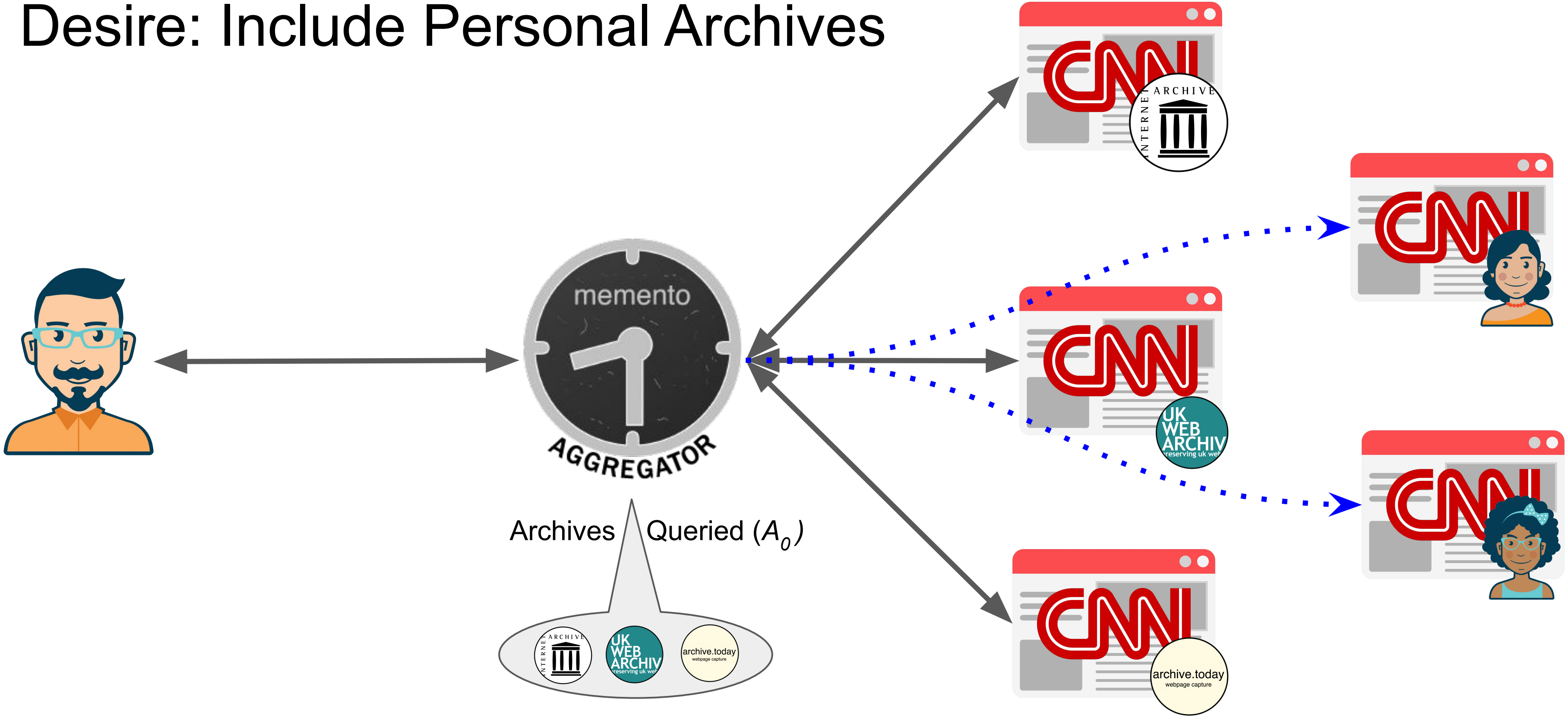
History

View: 50 per page · Transaction Type: All Transactions · Display Transactions Within: Any Date · From 01/02/2013 To 06/19/2015

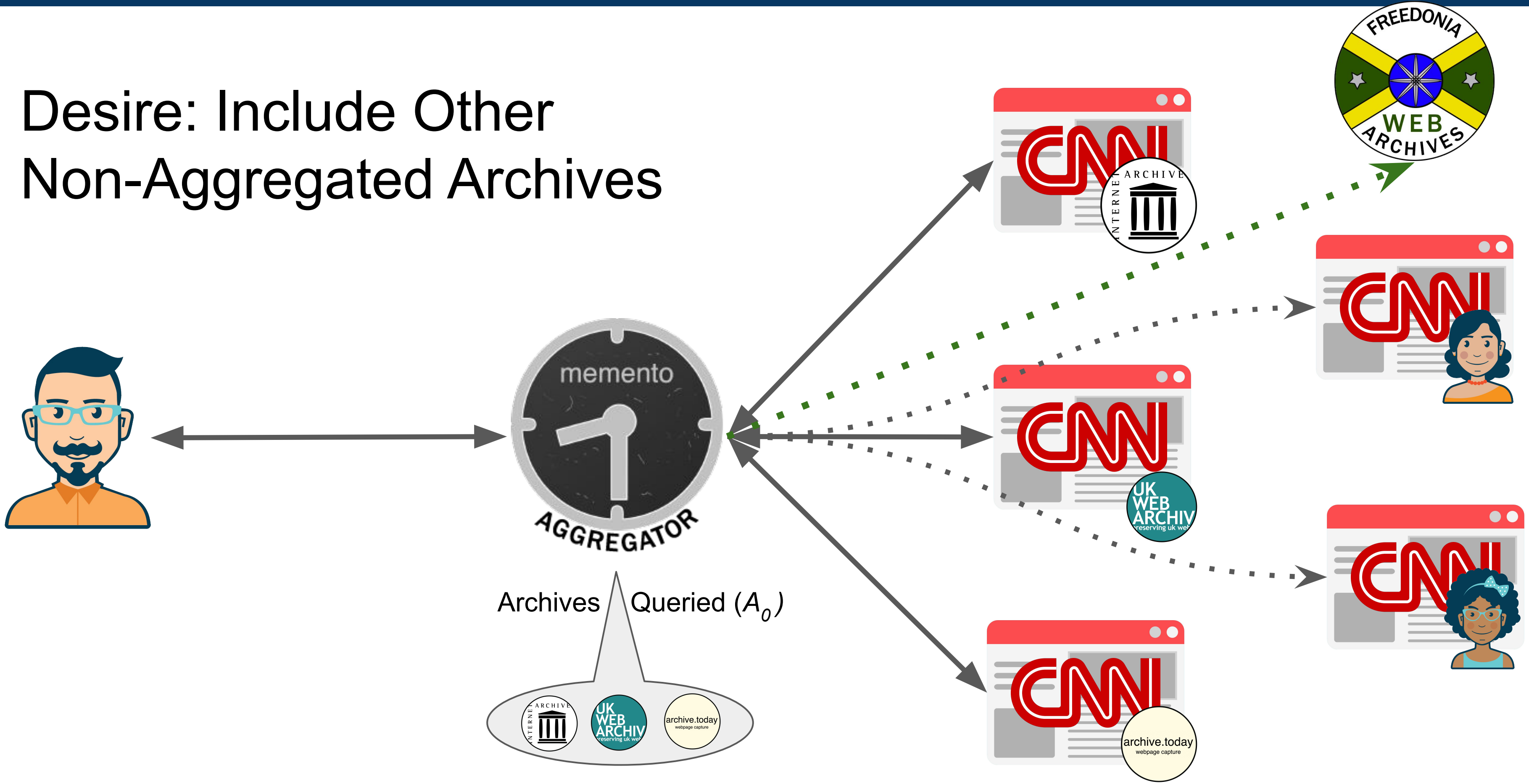
Date	Description	Withdrawals	Deposits	Balance
FRI Jun 19	Flux Capacitor Emporium Withdrawal	46,256.07		100,000.01
FRI Jun 19	Old Dominion Uni Deposit		100,000.01	146,256.08
MON Jun 15	Check #3722 Cashed	8.19		46,256.07
TUE Jun 09	Columbia Gas Electronic Payment	34.26		46,264.26

Historical capture of Online Banking Interface

Desire: Include Personal Archives

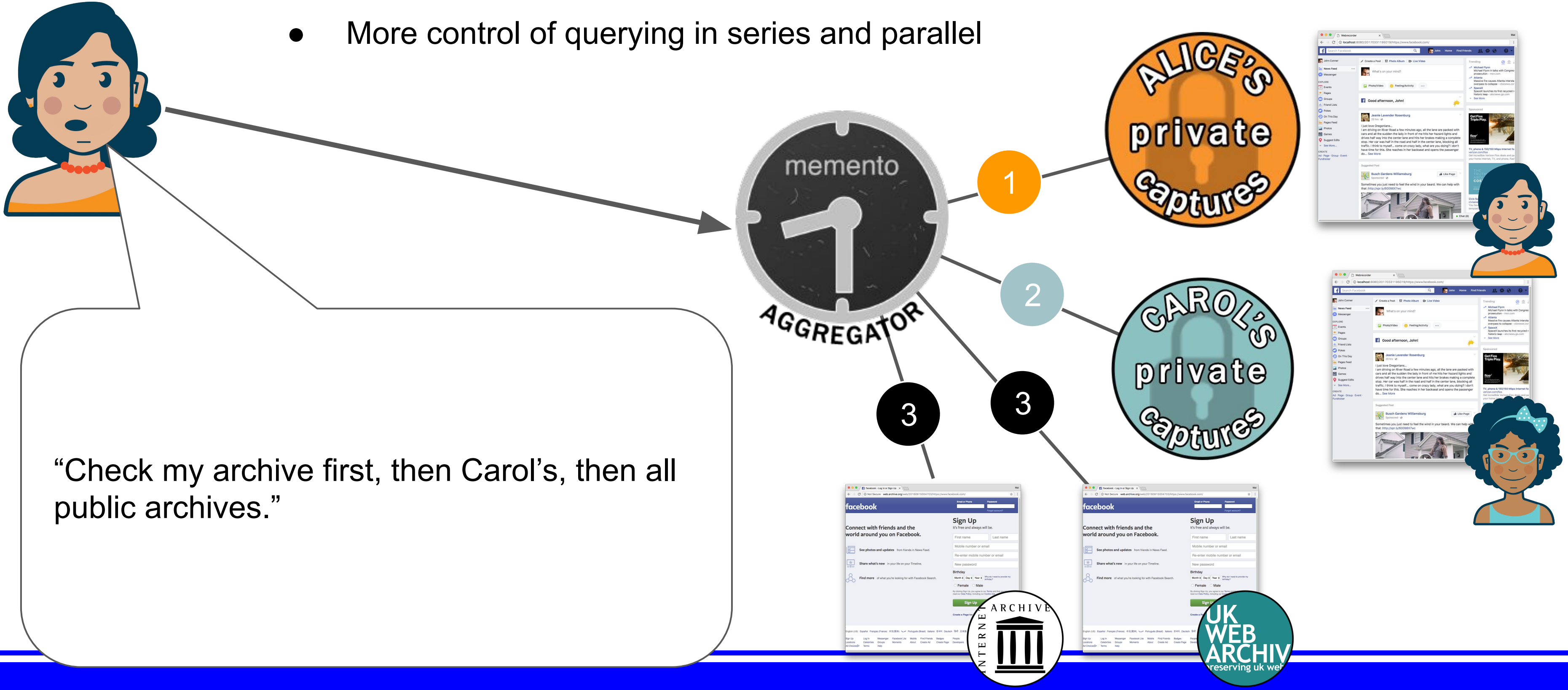

memento
AGGREGATOR
Archives Queried (A_0)
INTERNET ARCHIVE
UK WEB ARCHIVE preserving uk web
archive.today webpage capture
CNN

Desire: Include Other Non-Aggregated Archives



Query Precedence

- More control of querying in series and parallel

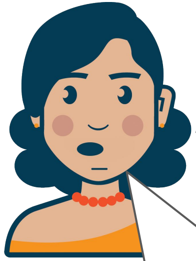


"Check my archive first, then Carol's, then all public archives."

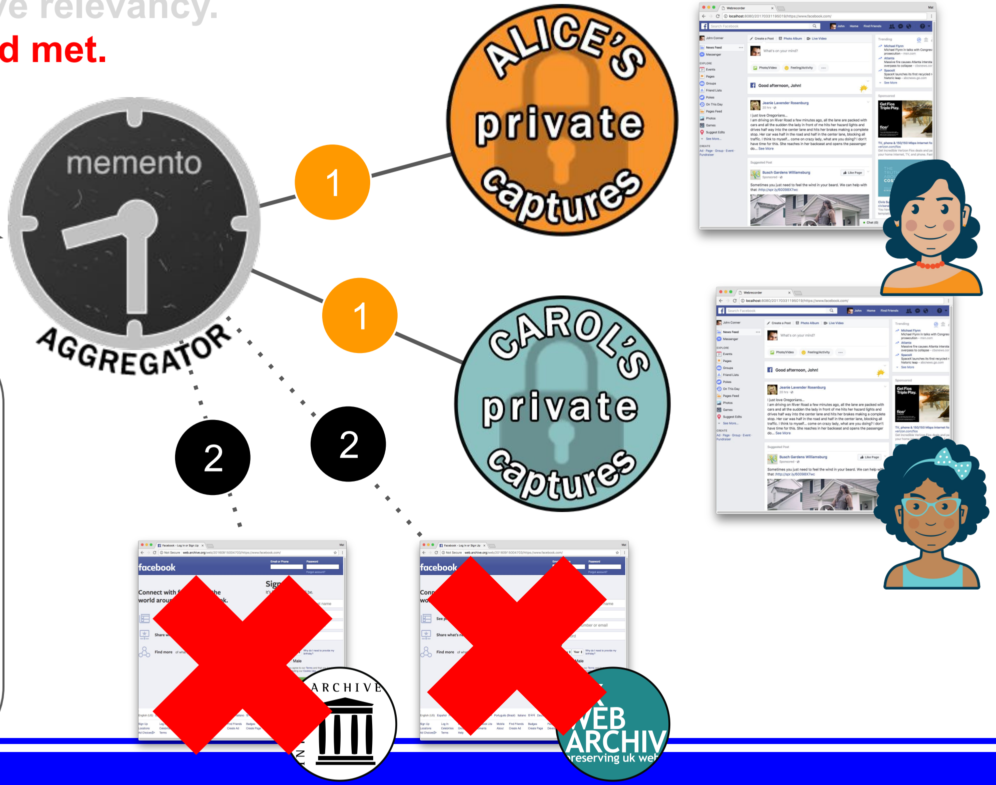
See Atkins', "Paywalls in the Internet Archive", March 2018

Query Short-Circuiting

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

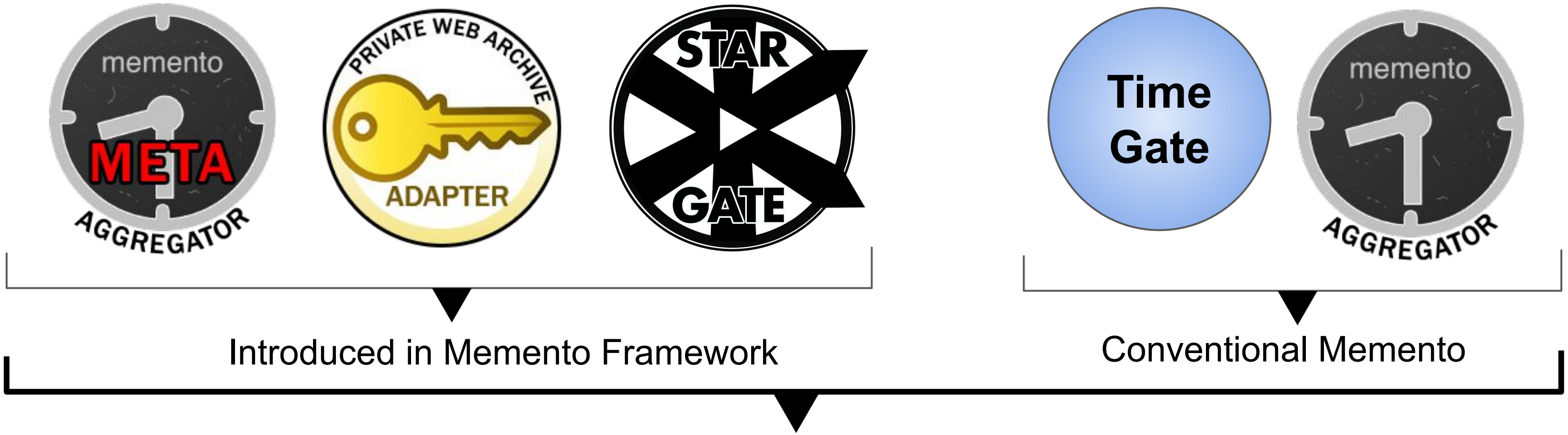


“Check private archives first. Iff you find no captures, only *then* check the public archives.



Mementities

- Memento + Entity (*entity* term already overused in web parlance)



Introduced in Memento Framework

Conventional Memento

Mementities

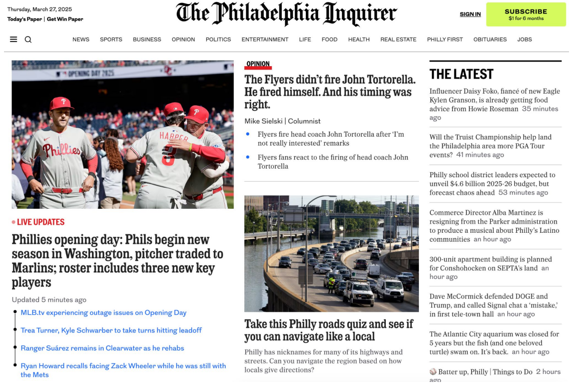
Originally introduced in Kelly et al. 2018

Future Access

- Should these private photos/bank statements be accessible to anyone?
  - Strawman answer: no, but how can this be systematically implemented to be both sophisticated for security and privacy and usable by the non-technical
- How do we balance ability to access in the future with privacy and security?
- Should a page's public/private access now persist into the future?
- Reducing accessibility now can be at odds with any accessibility in the future

Can We Save The Web We See From Our Perspective?

- Repurpose user’s daily driver profile as crawler basis
- Permutate attributes of a user to represent a “persona”, producing a web experience closer to that of an actual web user cf. crawler
- Avoid clean slate crawling and delegation to a user-agnostic crawler
- Scale?



Thursday, March 27, 2025
Today's Paper | Get Win Paper

The Philadelphia Inquirer

SIGN IN

SUBSCRIBE
$1 for 6 months

NEWS SPORTS BUSINESS OPINION POLITICS ENTERTAINMENT LIFE FOOD HEALTH REAL ESTATE PHILLY FIRST OBITUARIES JOBS

● LIVE UPDATES

Phillies opening day: Phils begin new season in Washington, pitcher traded to Marlins; roster includes three new key players

Updated 5 minutes ago

- MLB.tv experiencing outage issues on Opening Day
- Trea Turner, Kyle Schwarber to take turns hitting leadoff
- Ranger Suárez remains in Clearwater as he rehabs
- Ryan Howard recalls facing Zack Wheeler while he was still with the Mets

OPINION

The Flyers didn’t fire John Tortorella. He fired himself. And his timing was right.

Mike Sielski | Columnist

- Flyers fire head coach John Tortorella after ‘I’m not really interested’ remarks
- Flyers fans react to the firing of head coach John Tortorella

Take this Philly roads quiz and see if you can navigate like a local

Philly has nicknames for many of its highways and streets. Can you navigate the region based on how locals give directions?

THE LATEST

Influencer Daisy Foko, fiancé of new Eagle Kylen Granson, is already getting food advice from Howie Roseman 35 minutes ago

Will the Truist Championship help land the Philadelphia area more PGA Tour events? 41 minutes ago

Philly school district leaders expected to unveil $4.6 billion 2025-26 budget, but forecast chaos ahead 53 minutes ago

Commerce Director Alba Martinez is resigning from the Parker administration to produce a musical about Philly’s Latino communities an hour ago

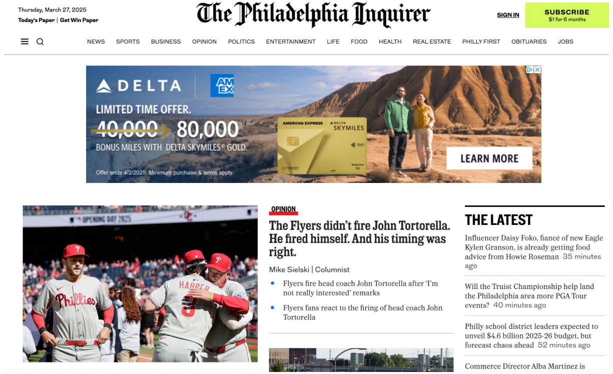
300-unit apartment building is planned for Conshohocken on SEPTA’s land an hour ago

Dave McCormick defended DOGE and Trump, and called Signal chat a ‘mistake,’ in first tele-town hall an hour ago

The Atlantic City aquarium was closed for 5 years but the fish (and one beloved turtle) swam on. It’s back. an hour ago

⚾ Batter up, Philly | Things to Do 2 hours ago

vs



Thursday, March 27, 2025
Today's Paper | Get Win Paper

The Philadelphia Inquirer

SIGN IN

SUBSCRIBE
$1 for 6 months

NEWS SPORTS BUSINESS OPINION POLITICS ENTERTAINMENT LIFE FOOD HEALTH REAL ESTATE PHILLY FIRST OBITUARIES JOBS

OPINION

The Flyers didn’t fire John Tortorella. He fired himself. And his timing was right.

Mike Sielski | Columnist

- Flyers fire head coach John Tortorella after ‘I’m not really interested’ remarks
- Flyers fans react to the firing of head coach John Tortorella

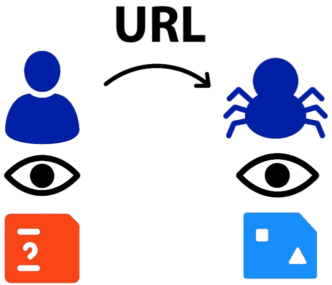
THE LATEST

Influencer Daisy Foko, fiancé of new Eagle Kylen Granson, is already getting food advice from Howie Roseman 35 minutes ago

Will the Truist Championship help land the Philadelphia area more PGA Tour events? 40 minutes ago

Philly school district leaders expected to unveil $4.6 billion 2025-26 budget, but forecast chaos ahead 52 minutes ago

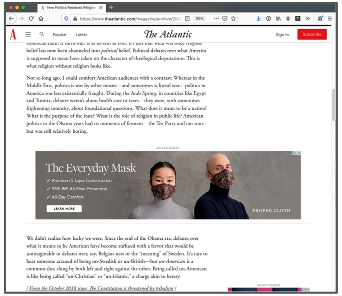
Commerce Director Alba Martinez is



Outstanding Problems Under Investigation

A few of many

- Web Advertisements are intentionally lost but are important
    - AI-driven personas to surface deep web content; reusing browser profile to inform crawls
- Private/Personal web archive resilience through replication with security
- Resilience of remote resources through sophisticated linked data structures
- User-informed preference for information retrieval of the past in dimensions beyond time
    - e.g., topical, quality, ad hoc supplied lambda functions
- *Better* defining fuzziness of public/private and personal/institutional captures to better align with non-boolean reality


How Politics Replaced Religion
https://www.theatlantic.com/magazine/archive/2021
90%
Popular
Latest
The Atlantic
Sign In
Subscribe

American faith, it turns out, is as fervent as ever; it's just that what was once *religious* belief has now been channeled into *political* belief. Political debates over what America is supposed to mean have taken on the character of theological disputations. This is what religion without religion looks like.

Not so long ago, I could comfort American audiences with a contrast: Whereas in the Middle East, politics is war by other means—and sometimes is literal war—politics in America was less existentially fraught. During the Arab Spring, in countries like Egypt and Tunisia, debates weren't about health care or taxes—they were, with sometimes frightening intensity, about foundational questions: What does it mean to be a nation? What is the purpose of the state? What is the role of religion in public life? American politics in the Obama years had its moments of ferment—the Tea Party and tan suits—but was still relatively boring.

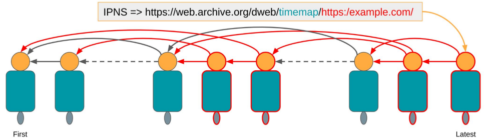
ADVERTISEMENT

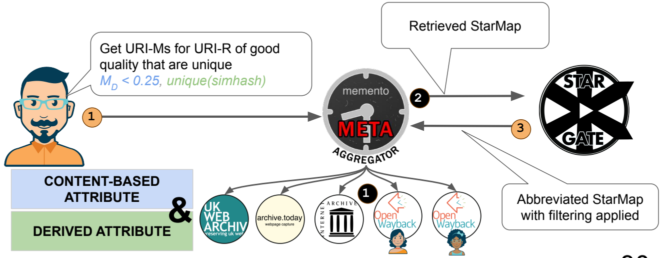
The Everyday Mask
✓ Premium 5-Layer Construction
✓ 95% BFE Air Filter Protection
✓ All-Day Comfort
LEARN MORE
PROPER CLOTH

We didn't realize how lucky we were. Since the end of the Obama era, debates over what it means to be American have become suffused with a fervor that would be unimaginable in debates over, say, Belgian-ness or the "meaning" of Sweden. It's rare to hear someone accused of being un-Swedish or un-British—but *un-American* is a common slur, slung by both left and right against the other. Being called un-American is like being called "un-Christian" or "un-Islamic," a charge akin to heresy.

[From the October 2018 issue: The Constitution is threatened by tribalism]

ADVERTISEMENT


IPNS => https://web.archive.org/dweb/timemap/https:/example.com/
First
Latest


Get URI-Ms for URI-R of good quality that are unique
M_D < 0.25, unique(simhash)
Retrieved StarMap
1
2
3
memento
META
AGGREGATOR
STAR GATE
CONTENT-BASED ATTRIBUTE
&
DERIVED ATTRIBUTE
UK WEB ARCHIVE
archive.today
INTERNET ARCHIVE
OpenWayback
OpenWayback
1
Abbreviated StarMap with filtering applied