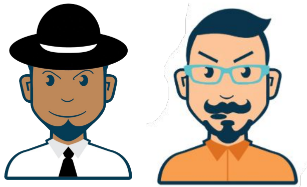
Save *My* Web Things For Me

and Others

but Not All

Mat Kelly, PhD

Assistant Professor, Information Science
Drexel University
College of Computing & Informatics
mkelly@drexel.edu
https://matkelly.com



Some figures adapted from work of Agata Krych CC-BY-SA 4.0

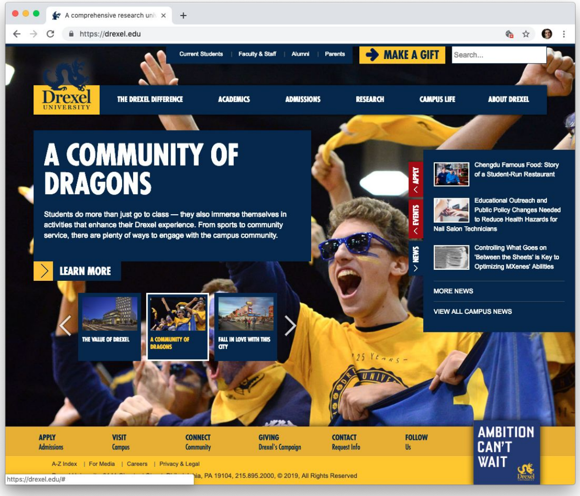
Master’s of Information Webinar Series
February 4, 2020

Drexel UNIVERSITY

Drexel CCI’S DCM Major

- Specialization in Digital Content Management for Masters in Information Science degree
- Courses
  - 3 foundation
  - 5 core
  - 6 electives
  - Capstone (2 quarters)
- Sample Courses in Major:
  - INFO633 Information Visualization
  - INFO654 Enterprise Content Management (being taught by Dr. Kelly, Spring 2020)
  - INFO 676 Applied Ontology

The Web



A COMMUNITY OF DRAGONS

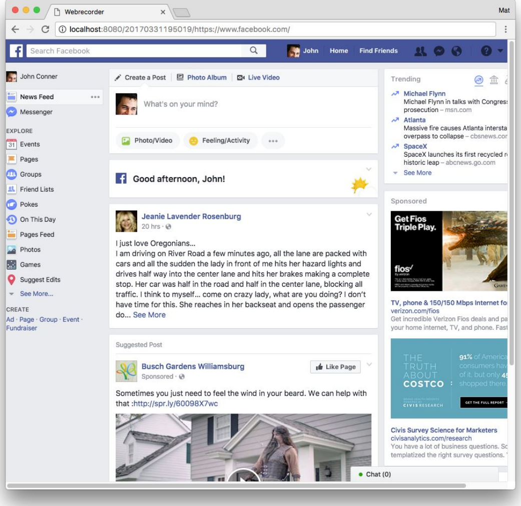
Students do more than just go to class — they also immerse themselves in activities that enhance their Drexel experience. From sports to community service, there are plenty of ways to engage with the campus community.

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

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

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

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

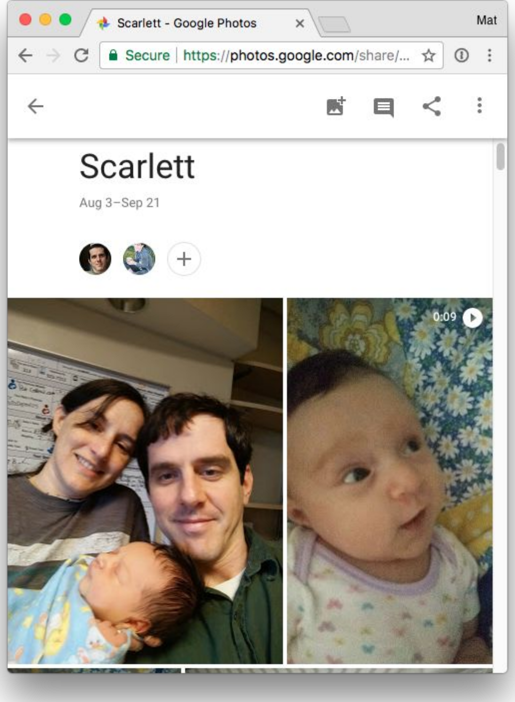
SpaceX
SpaceX launches its first recycled r
historic leap – abcnews.go.com

See More

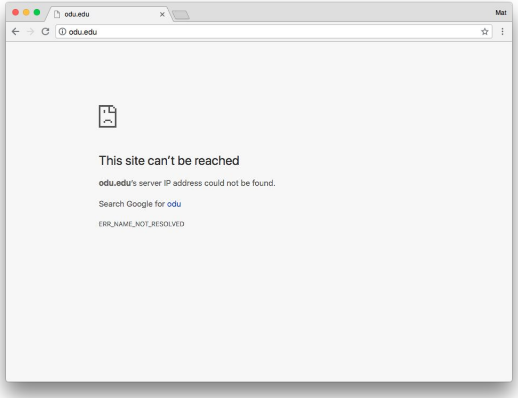
TV, phone & 150/150 Mbps Internet fo
verizon.com/fios
Get incredible Verizon Fios deals and pa
your home internet, TV, and phone. Fast

91% of America consumers hav of it, but only 45 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

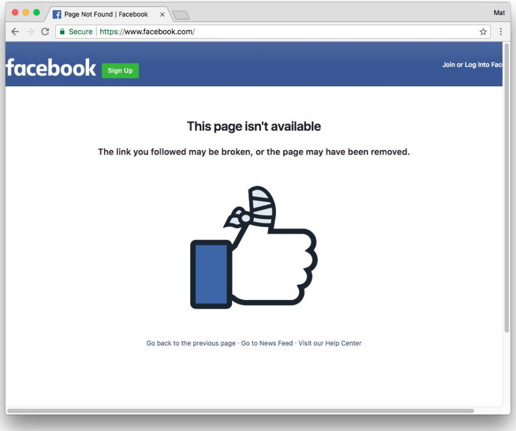


This site can't be reached

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

Search Google for odu

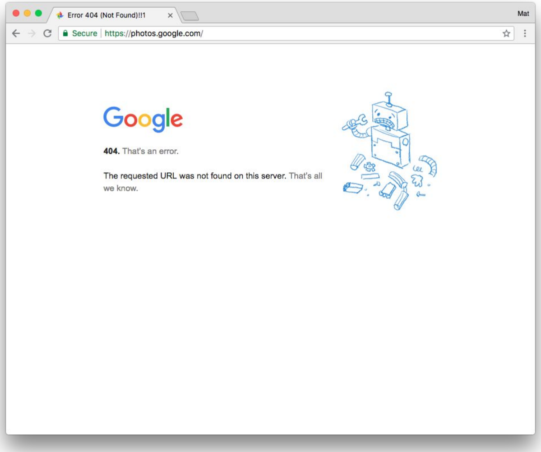
ERR_NAME_NOT_RESOLVED



This page isn't available

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

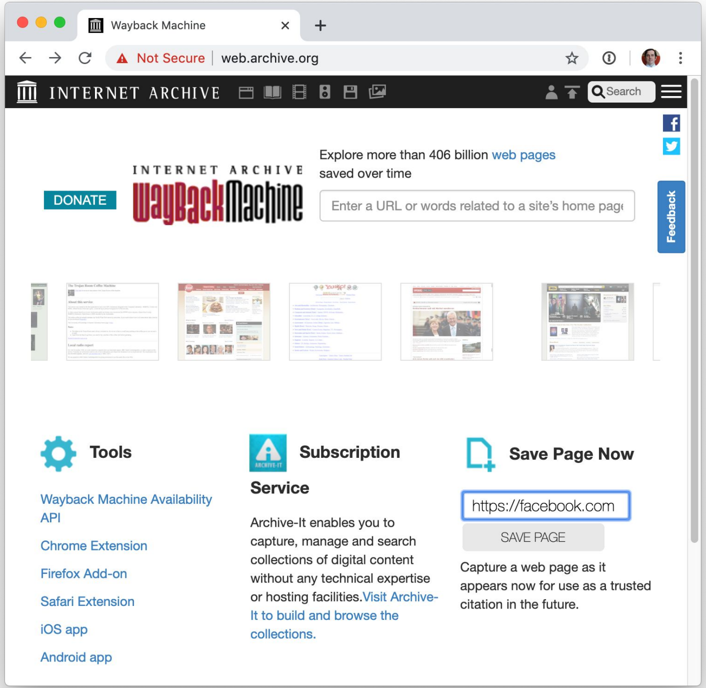
Go back to the previous page · Go to News Feed · Visit our Help Center



404. That's an error.

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

The Web...can be preserved



Explore more than 406 billion 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

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.

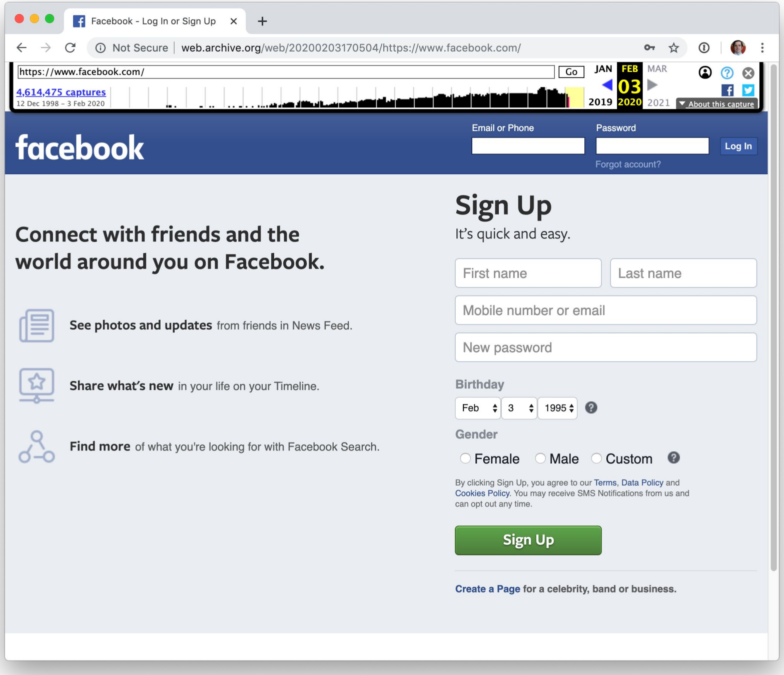
Save Page Now

https://facebook.com

SAVE PAGE

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

Internet Archive's interface to "Save Page Now"



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 quick and easy.

First name | Last name

Mobile number or email

New password

Birthday

Feb 3 1995

Gender

○ Female ○ Male ○ Custom

By clicking Sign Up, you agree to our Terms, Data Policy and Cookies Policy. You may receive SMS Notifications from us and can opt out any time.

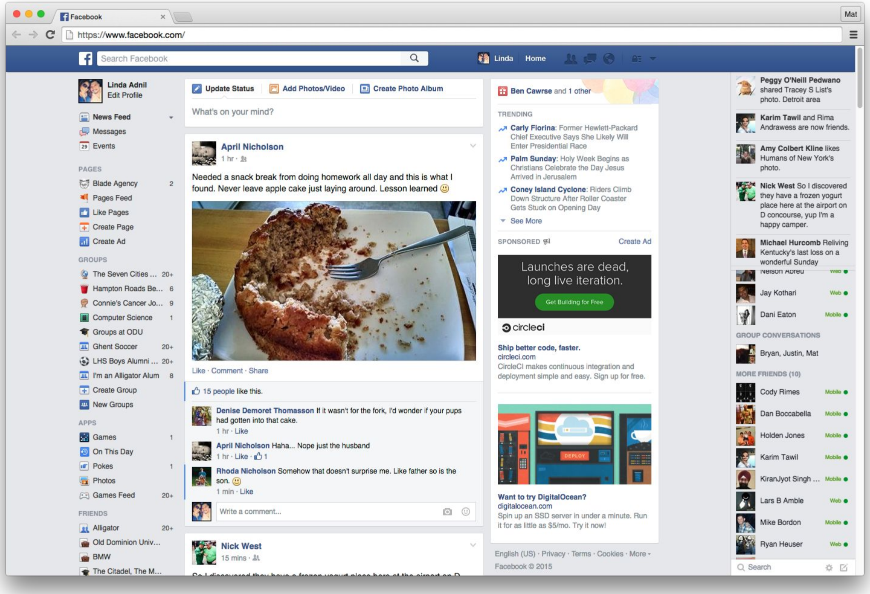
Sign Up

Create a Page for a celebrity, band or business.

IA's result of preserving facebook.com

* except when it's not

The Web...can be preserved*



April Nicholson
1 hr ·

Needed a snack break from doing homework all day and this is what I found. Never leave apple cake just laying around. Lesson learned 🙂

15 people like this.

Denise Demoret Thomasson If it wasn't for the fork, I'd wonder if your pups had gotten into that cake.
1 hr · Like

April Nicholson Haha... Nope just the husband
1 hr · Like · 1

Rhoda Nicholson Somehow that doesn't surprise me. Like father so is the son. 🙂
1 min · Like

Nick West
15 mins ·

Ben Cawrse and 1 other

Carly Florina: Former Hewlett-Packard Chief Executive Says She Likely Will Enter Presidential Race

Palm Sunday: Holy Week Begins as Christians Celebrate the Day Jesus Arrived in Jerusalem

Coney Island Cyclone: Riders Climb Down Structure After Roller Coaster Gets Stuck on Opening Day

Launches are dead, long live iteration.

Ship better code, faster.
circleci.com
CircleCI makes continuous integration and deployment simple and easy. Sign up for free.

Want to try DigitalOcean?
digitalocean.com
Spin up an SSD server in under a minute. Run it for as little as $5/mo. Try it now!

Peggy O'Neill Pedwano shared Tracey S List's photo. Detroit area

Karim Tawil and Rima Andrawess are now friends.

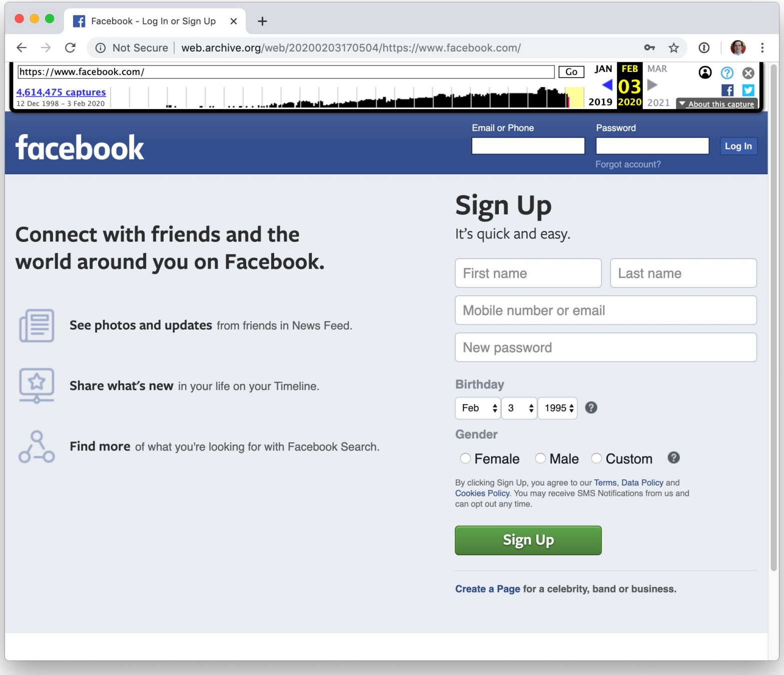
Amy Colbert Kline likes Humans of New York's photo.

Nick West So I discovered they have a frozen yogurt place here at the airport on D concourse, yup I'm a happy camper.

Michael Hurcomb Reliving Kentucky's last loss on a wonderful Sunday

What you see on the live web

≠



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

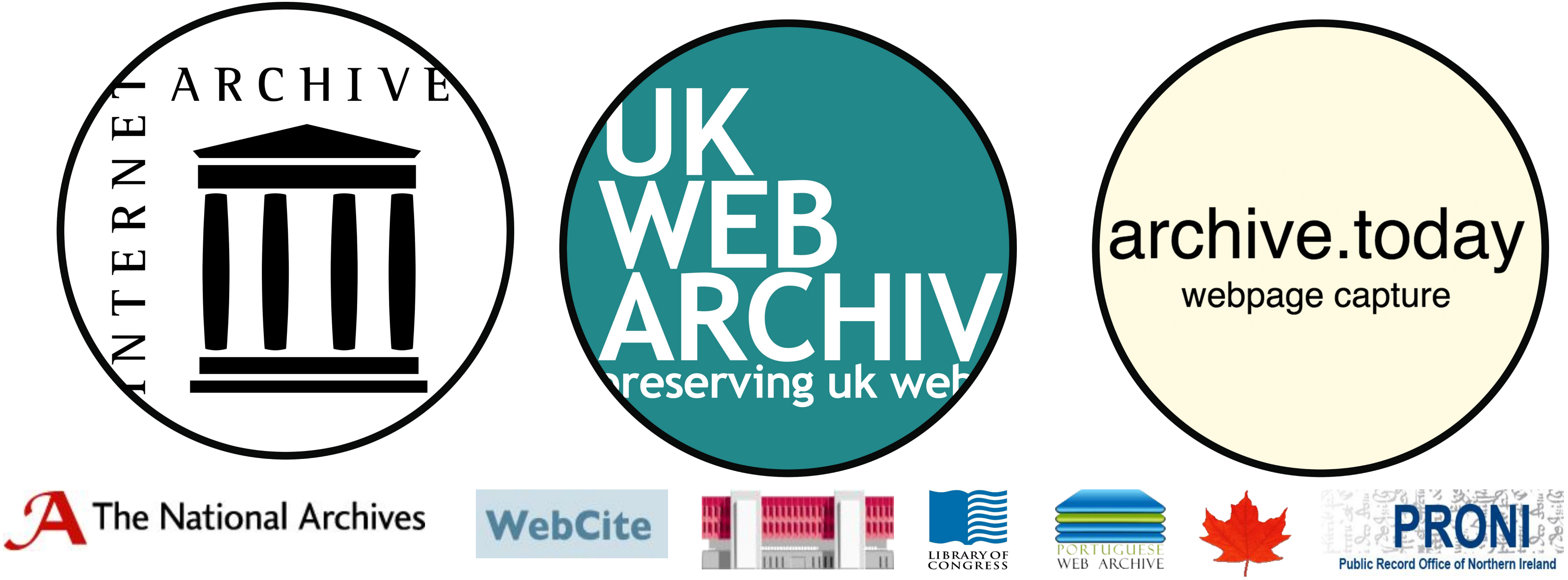
It's quick and easy.

By clicking Sign Up, you agree to our Terms, Data Policy and Cookies Policy. You may receive SMS Notifications from us and can opt out any time.

Create a Page for a celebrity, band or business.

IA's result of preserving facebook.com

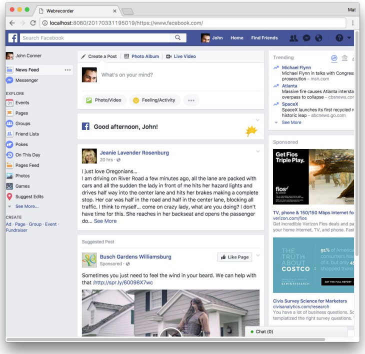
Archiving Institutions' Efforts


INTERNET ARCHIVE
UK WEB ARCHIVE preserving uk web
archive.today webpage capture
The National Archives
WebCite
LIBRARY OF CONGRESS
PORTUGUESE WEB ARCHIVE
PRONI Public Record Office of Northern Ireland

S. Alam et al., Web Archive Profiling Through CDX Summarization, International Journal on Digital Libraries, 17(3), pp. 223--238, 2016.

Institutions' Efforts...likely won't have your things

Private social media pages



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

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

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

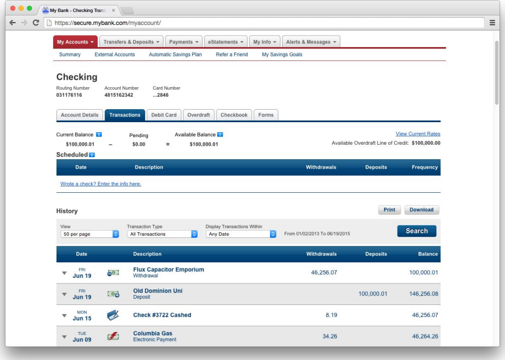
SpaceX
SpaceX launches its first recycled r historic leap – abcnews.go.com

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

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

Bank account ledger



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

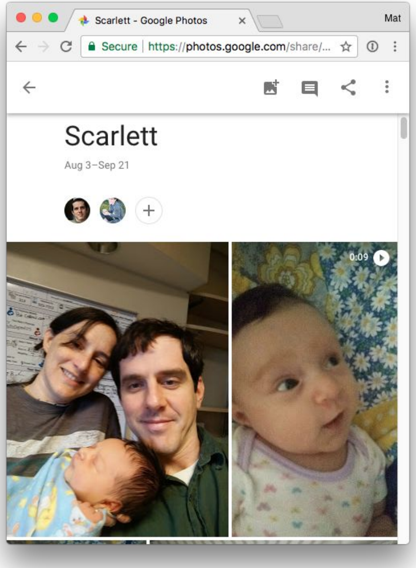
Date	Description	Withdrawals	Deposits	Frequency
Wrote a check? Enter the info here.				

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

Online photo storage



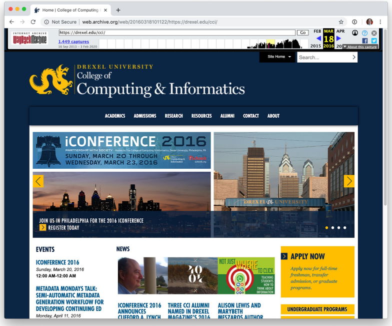
- Niche Sites
- Corporate *intranets*
- etc.

...and it might be best they didn't.

The Divide between the Live & Archived Web

For what they ***do*** capture:

- What (URI?) do these captures represent?
- When were they captured?
- Can the entities in the archive give us more context?



Home | College of Computing & ...

web.archive.org/web/20160318101122/https://drexel.edu/cci/

https://drexel.edu/cci/

1,449 captures
16 Sep 2013 – 3 Feb 2020

FEB MAR APR
18
2015 2016 20

DREXEL UNIVERSITY College of Computing & Informatics

iCONFERENCE 2016
PARTNERSHIP WITH SOCIETY
SUNDAY, MARCH 20 THROUGH WEDNESDAY, MARCH 23, 2016

JOIN US IN PHILADELPHIA FOR THE 2016 ICONFERENCE
REGISTER TODAY

EVENTS

ICONFERENCE 2016
Sunday, March 20, 2016
12:00 AM-12:00 AM

METADATA MONDAYS TALK: SEMI-AUTOMATIC METADATA GENERATION WORKFLOW FOR DEVELOPING CONTINUING ED
Monday, April 11, 2016

NEWS

ICONFERENCE 2016 ANNOUNCES CLIFFORD A. LYNCH

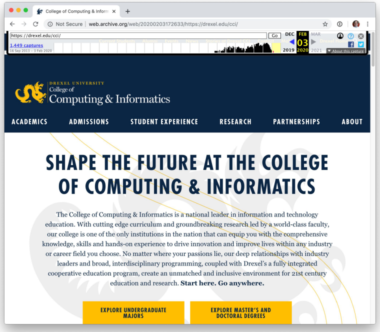
THREE CCI ALUMNI NAMED IN DREXEL MAGAZINE'S 2016

ALISON LEWIS AND MARYBETH MESZAROS AUTHOR

APPLY NOW

Apply now for full-time freshman, transfer admission, or graduate programs.

UNDERGRADUATE PROGRAMS



College of Computing & Inform...

web.archive.org/web/20200203172633/https://drexel.edu/cci/

https://drexel.edu/cci/

1,449 captures
16 Sep 2013 – 3 Feb 2020

DEC FEB MAR
03
2019 2020 2021

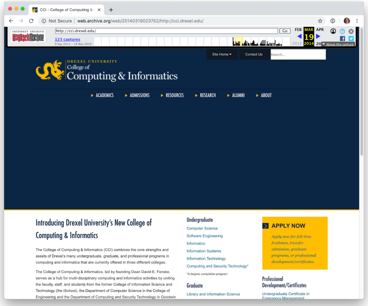
DREXEL UNIVERSITY College of Computing & Informatics

SHAPE THE FUTURE AT THE COLLEGE OF COMPUTING & INFORMATICS

The College of Computing & Informatics is a national leader in information and technology education. With cutting edge curriculum and groundbreaking research led by a world-class faculty, our college is one of the only institutions in the nation that can equip you with the comprehensive knowledge, skills and hands-on experience to drive innovation and improve lives within any industry or career field you choose. No matter where your passions lie, our deep relationships with industry leaders and broad, interdisciplinary programming, coupled with Drexel's a fully integrated cooperative education program, create an unmatched and inclusive environment for 21st century education and research. **Start here. Go anywhere.**

EXPLORE UNDERGRADUATE MAJORS

EXPLORE MASTER'S AND DOCTORAL DEGREES



CCI - College of Computing & ...

web.archive.org/web/20140319023752/http://cci.drexel.edu/

http://cci.drexel.edu/

123 captures
9 Sep 2013 – 19 Nov 2019

FEB MAR APR
19
2013 2014 20

DREXEL UNIVERSITY College of Computing & Informatics

Introducing Drexel University's New College of Computing & Informatics

The College of Computing & Informatics (CCI) combines the core strengths and assets of Drexel's many undergraduate, graduate, and professional programs in computing and informatics that are currently offered in three different colleges.

The College of Computing & Informatics, led by founding Dean David E. Fenske, serves as a hub for multi-disciplinary computing and informatics activities by uniting the faculty, staff, and students from the former College of Information Science and Technology (the iSchool), the Department of Computer Science in the College of Engineering and the Department of Computing and Security Technology in Goodwin

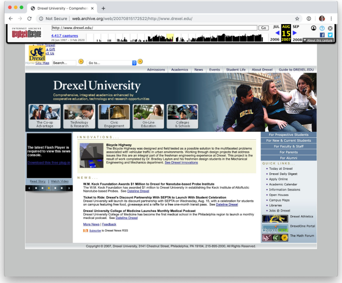
Undergraduate
Computer Science
Software Engineering
Informatics
Information Systems
Information Technology
Computing and Security Technology*
*A degree completion program

Graduate
Library and Information Science

APPLY NOW

Apply now for full-time freshman, transfer admission, graduate programs, or professional development/certificates.

Professional Development/Certificates
Undergraduate Certificate in Emergency Management



Drexel University - Comprehen...

web.archive.org/web/20070815172522/http://www.drexel.edu/

http://www.drexel.edu/

4,417 captures
26 Jun 1997 – 3 Feb 2020

JUL AUG SEP
15
2006 2007 2008

Drexel University

Comprehensive, integrated academics enhanced by cooperative education, technology and research opportunities

The latest Flash Player is required to view this news console.

Download this free plug-in

INNOVATIONS...

Bicycle Highway
The Bicycle Highway was designed and field tested as a possible solution to the multifaceted problems associated with vehicular traffic in urban environments. Working through design projects that address issues like this are an integral part of the freshman engineering experience at Drexel. This project is the result of work completed by Dr. Bradley Layton and his freshmen design students in the Mechanical Engineering and Mechanics department. See Drexel Innovations

NEWS...

W.M. Keck Foundation Awards $1 Million to Drexel for Nanotube-based Probe Institute
The W.M. Keck Foundation has awarded $1 million to Drexel University in establishing the Keck Institute of Attofluidic Nanotube-based Probes. See Dateline Drexel

Ticket to Ride: Drexel's Discount Partnership With SEPTA to Launch With Student Celebration
Drexel University will launch its discount partnership with SEPTA on Wednesday, Aug. 15, with a celebration for students on campus featuring free food, giveaways and a raffle for a free one-month transit pass. See Dateline Drexel

Drexel University College of Medicine Launches Monthly Medical Podcast
Drexel University College of Medicine has become the first medical school in the Philadelphia region to launch a monthly medical podcast. See Dateline Drexel

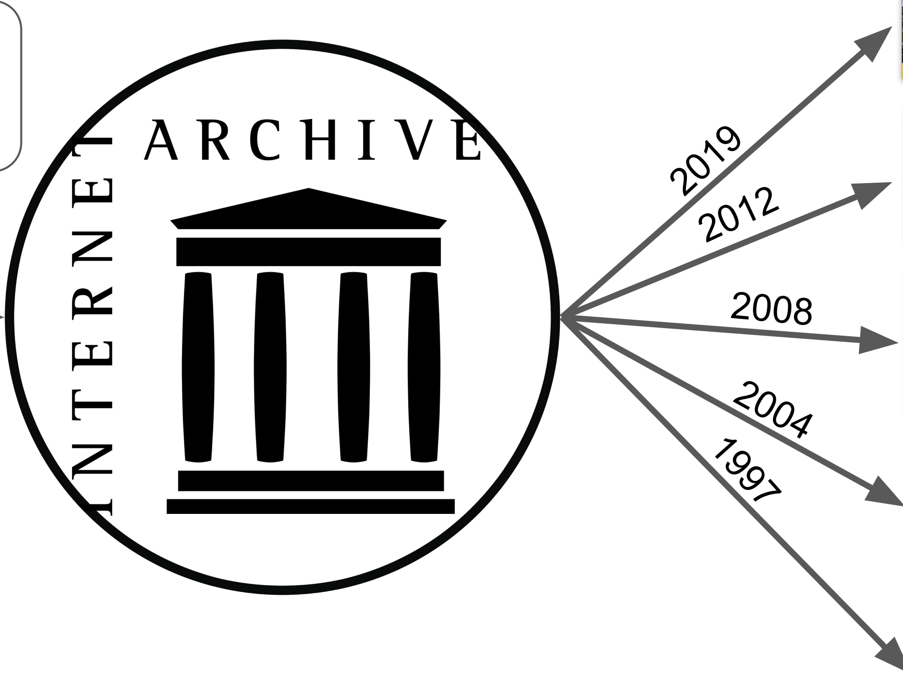
More News | Feedback

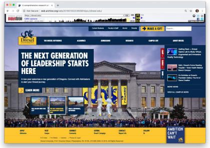
Subscribe to Drexel News RSS

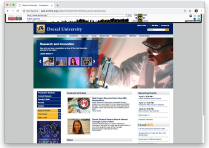
Copyright © 2007, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, 215-895-2000, All Rights Reserved.

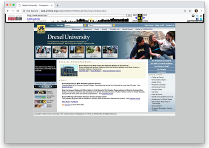
Needed a Standard Way for Archives to Express this Association

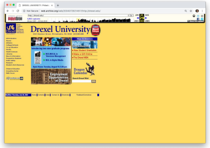
What did drexel.edu look like in the past?

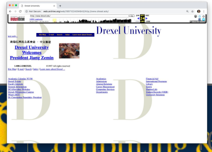



THE NEXT GENERATION OF LEADERSHIP STARTS HERE


Drexel University


Drexel University


Drexel University


Drexel University

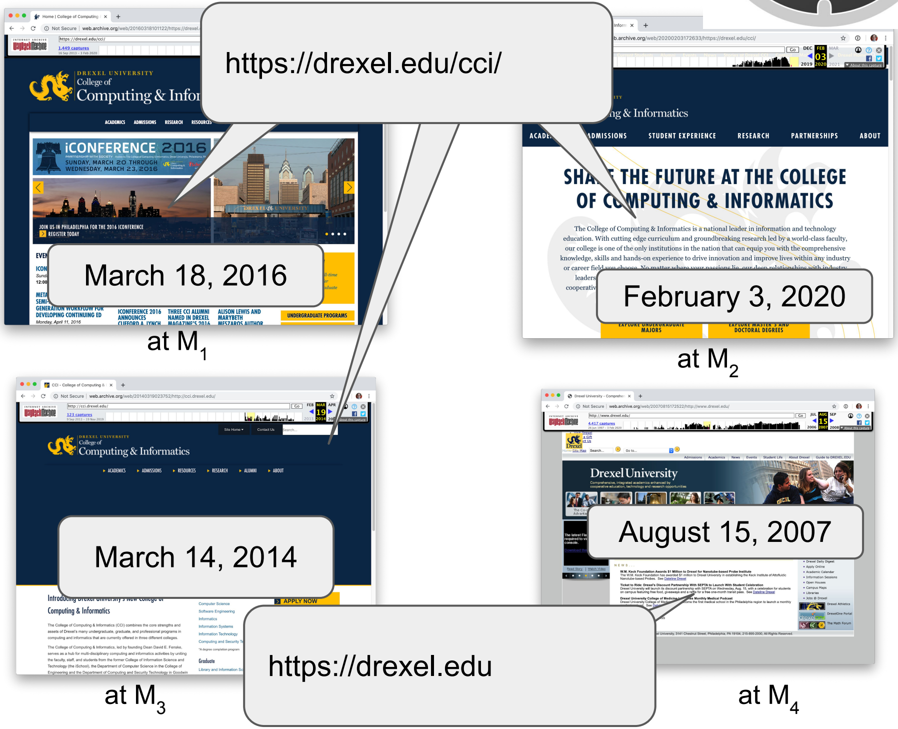
Memento Bridges the Divide

TimeMap for https://drexel.edu/cci/

original: https://drexel.edu/cci/
Memento1: March 14, 2014 @ M_3
Memento2: March 18, 2016 @ M_1
Memento3: February 3, 2020 @ M_2

TimeMap for https://drexel.edu

original: https://drexel.edu/
Memento1: August 15, 2007 @ M_4
...



https://drexel.edu/cci/

March 18, 2016

at M_1

February 3, 2020

at M_2

March 14, 2014

at M_3

August 15, 2007

https://drexel.edu

at M_4

Web Archiving Components$^{\text{(two of them)}}$

Preservation

- Saving Live Web Content Offline in a Standard, Interoperable Format (WARC)
  - Includes HTML, CSS, JavaScript, Flash, MP4, etc. and resources within each

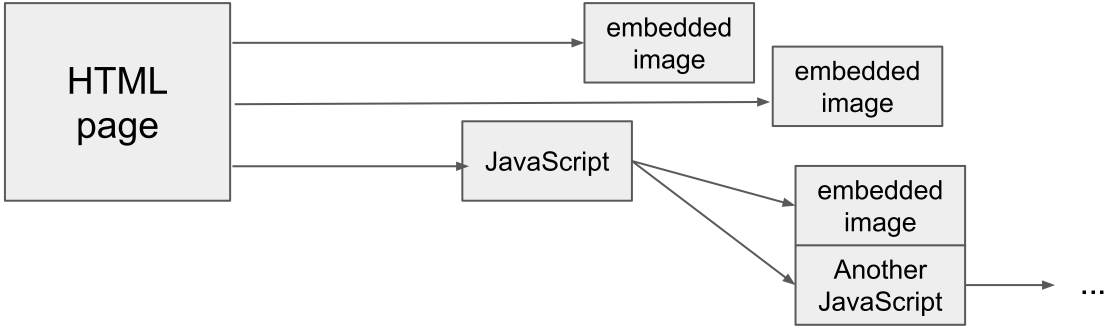
Access

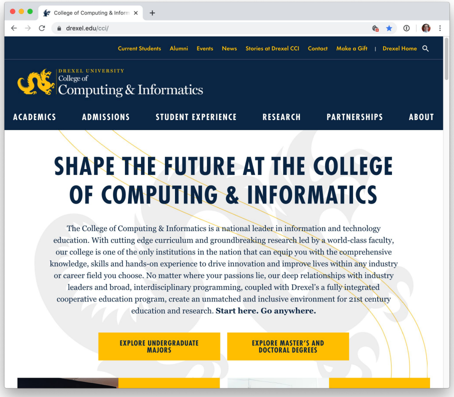
- Associating a request for a URI in time to a WARC
- Re-assembling Pages (no easy task)

Web Archiving Components

Preservation

- Capture an HTML page at a URI
- Capture all resource representations embedded within a page
- Associate the what (URI) and when (time) of the captured resource representations





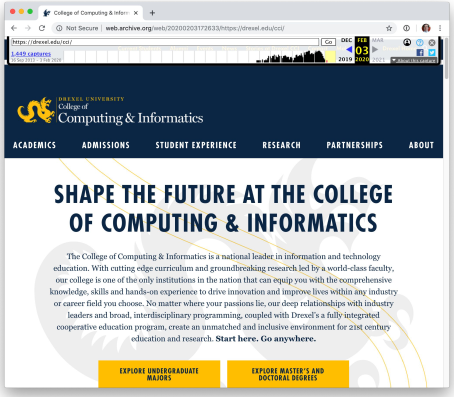
SHAPE THE FUTURE AT THE COLLEGE OF COMPUTING & INFORMATICS

The College of Computing & Informatics is a national leader in information and technology education. With cutting edge curriculum and groundbreaking research led by a world-class faculty, our college is one of the only institutions in the nation that can equip you with the comprehensive knowledge, skills and hands-on experience to drive innovation and improve lives within any industry or career field you choose. No matter where your passions lie, our deep relationships with industry leaders and broad, interdisciplinary programming, coupled with Drexel's a fully integrated cooperative education program, create an unmatched and inclusive environment for 21st century education and research. **Start here. Go anywhere.**

EXPLORE UNDERGRADUATE MAJORS
EXPLORE MASTER'S AND DOCTORAL DEGREES

CCI homepage ~now

CCI homepage, “recently” archived



SHAPE THE FUTURE AT THE COLLEGE OF COMPUTING & INFORMATICS

The College of Computing & Informatics is a national leader in information and technology education. With cutting edge curriculum and groundbreaking research led by a world-class faculty, our college is one of the only institutions in the nation that can equip you with the comprehensive knowledge, skills and hands-on experience to drive innovation and improve lives within any industry or career field you choose. No matter where your passions lie, our deep relationships with industry leaders and broad, interdisciplinary programming, coupled with Drexel's a fully integrated cooperative education program, create an unmatched and inclusive environment for 21st century education and research. **Start here. Go anywhere.**

EXPLORE UNDERGRADUATE MAJORS
EXPLORE MASTER'S AND DOCTORAL DEGREES

Web Archiving Components

Preservation

- Concatenated “records”
- WARC header provides metadata and provenance info
- HTTP header blocks preserved as Part of the transaction

WARC response header

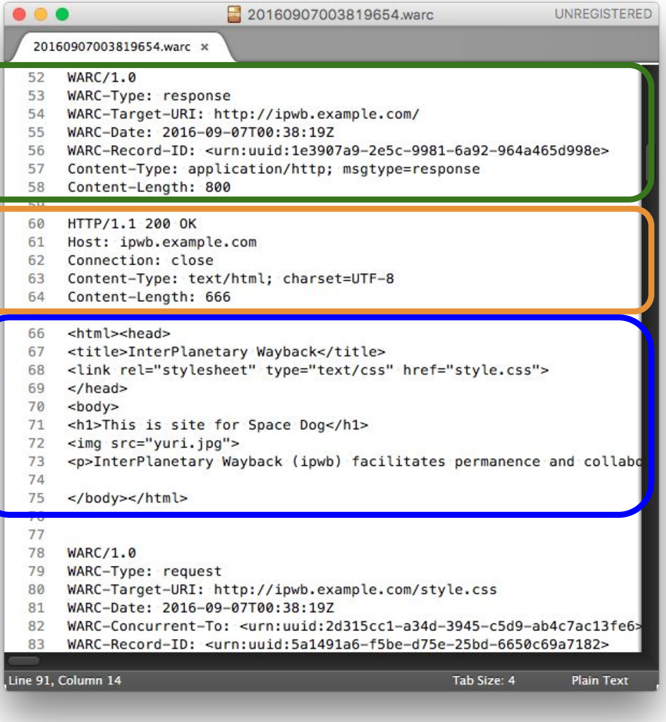
HTTP header block

HTTP payload block

ICS › 35 › 35.240 › 35.240.30

ISO 28500:2017

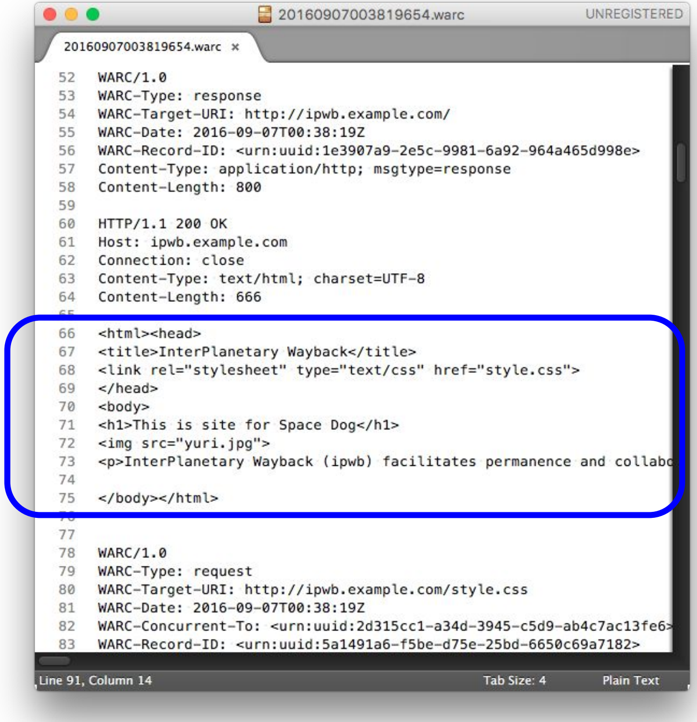
WARC file format



```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://ipwb.example.com/
WARC-Date: 2016-09-07T00:38:19Z
WARC-Record-ID: <urn:uuid:1e3907a9-2e5c-9981-6a92-964a465d998e>
Content-Type: application/http; msgtype=response
Content-Length: 800

HTTP/1.1 200 OK
Host: ipwb.example.com
Connection: close
Content-Type: text/html; charset=UTF-8
Content-Length: 666

<html><head>
<title>InterPlanetary Wayback</title>
<link rel="stylesheet" type="text/css" href="style.css">
</head>
<body>
<h1>This is site for Space Dog</h1>
<img src="yuri.jpg">
<p>InterPlanetary Wayback (ipwb) facilitates permanence and collabo

</body></html>


WARC/1.0
WARC-Type: request
WARC-Target-URI: http://ipwb.example.com/style.css
WARC-Date: 2016-09-07T00:38:19Z
WARC-Concurrent-To: <urn:uuid:2d315cc1-a34d-3945-c5d9-ab4c7ac13fe6>
WARC-Record-ID: <urn:uuid:5a1491a6-f5be-d75e-25bd-6650c69a7182>
```

Web Archiving Components

Access

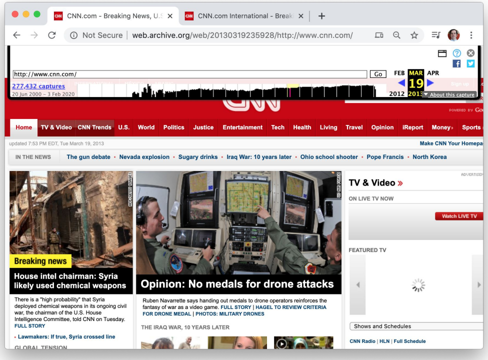


```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://ipwb.example.com/
WARC-Date: 2016-09-07T00:38:19Z
WARC-Record-ID: <urn:uuid:1e3907a9-2e5c-9981-6a92-964a465d998e>
Content-Type: application/http; msgtype=response
Content-Length: 800

HTTP/1.1 200 OK
Host: ipwb.example.com
Connection: close
Content-Type: text/html; charset=UTF-8
Content-Length: 666

<html><head>
<title>InterPlanetary Wayback</title>
<link rel="stylesheet" type="text/css" href="style.css">
</head>
<body>
<h1>This is site for Space Dog</h1>
<img src="yuri.jpg">
<p>InterPlanetary Wayback (ipwb) facilitates permanence and collabo

</body></html>


WARC/1.0
WARC-Type: request
WARC-Target-URI: http://ipwb.example.com/style.css
WARC-Date: 2016-09-07T00:38:19Z
WARC-Concurrent-To: <urn:uuid:2d315cc1-a34d-3945-c5d9-ab4c7ac13fe6>
WARC-Record-ID: <urn:uuid:5a1491a6-f5be-d75e-25bd-6650c69a7182>
```

- WARCs are indexed for URI→ record lookup
- HTML content is served with URIs rewritten to “point” back into the archive
  - Inclusive of embedded resources and “external links” pointing to other archived web pages
- Further lookups result with subsequent rewriting until no further HTTP requests are made

Web Archiving Components

Access



http://www.cnn.com/

277,432 captures

20 Jun 2000 – 3 Feb 2020

updated 7:53 PM EDT, Tue March 19, 2013

House intel chairman: Syria likely used chemical weapons

There is a "high probability" that Syria deployed chemical weapons in its ongoing civil war, the chairman of the U.S. House Intelligence Committee, told CNN on Tuesday. FULL STORY

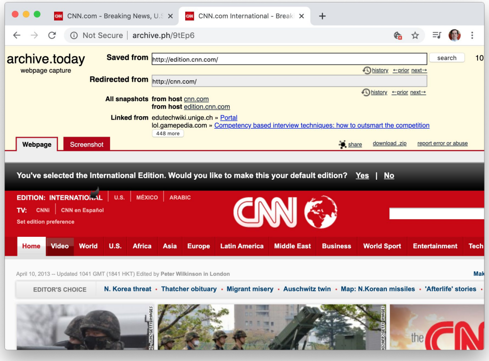
- Lawmakers: If true, Syria crossed line

Opinion: No medals for drone attacks

Ruben Navarrette says handing out medals to drone operators reinforces the fantasy of war as a video game. FULL STORY | HAGEL TO REVIEW CRITERIA FOR DRONE MEDAL | PHOTOS: MILITARY DRONES

Experience the past representation online, e.g.,

- http://web.archive.org/web/20130319235928/http://www.cnn.com/
  - Cnn.com at Tue, 19 Mar 2013 23:59:28 GMT
- http://archive.ph/9tEp6
  - cnn.com at Wed, 10 Apr 2013 11:33:55 GMT



Saved from http://edition.cnn.com/

Redirected from http://cnn.com/

All snapshots from host cnn.com
from host edition.cnn.com

Linked from edutechwiki.unige.ch » Portal
lol.gamepedia.com » Competency based interview techniques: how to outsmart the competition

You've selected the International Edition. Would you like to make this your default edition?

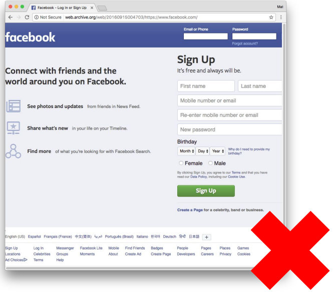
April 10, 2013 -- Updated 1041 GMT (1841 HKT) Edited by Peter Wilkinson in London

URIs are opaque*, semantics should not be inferred

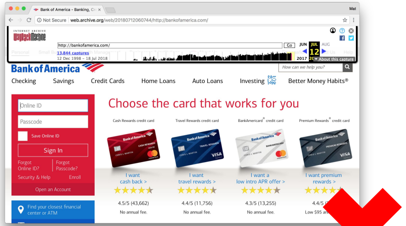
e.g., **20130319** does not necessarily mean the capture represents the page on **March 19, 2013**

* https://www.w3.org/DesignIssues/Axioms.html#opaque

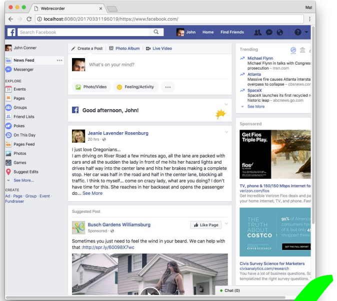
Backtracking - Save My Things


Facebook - Log In or Sign Up
web.archive.org/web/20160916004703/https://www.facebook.com/
facebook
Connect with friends and the world around you on Facebook.
Sign Up
It's free and always will be.
Sign Up

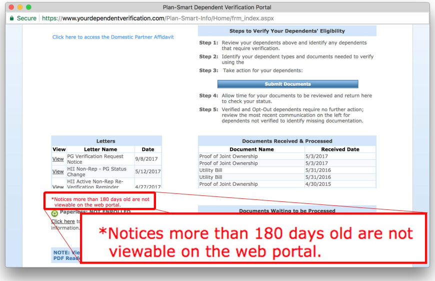
✗


Bank of America - Banking, Credit Cards, Home Loans
web.archive.org/web/20180712060744/http://bankofamerica.com/
Bank of America
Checking
Savings
Credit Cards
Home Loans
Auto Loans
Investing
Better Money Habits®
Choose the card that works for you
Sign In
Open an Account

✗

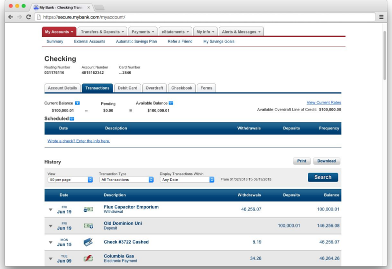

Webrecorder
localhost:8080/20170331195019/https://www.facebook.com/
Good afternoon, John!
Jeanie Lavender Rosenburg
Busch Gardens Williamsburg

✓

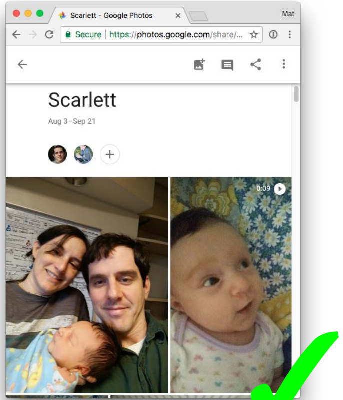

Plan-Smart Dependent Verification Portal
https://www.yourdependentverification.com/Plan-Smart-Info/Home/frm_index.aspx
Steps to Verify Your Dependents' Eligibility
Submit Documents
Letters
Documents Received & Processed
Documents Waiting to be Processed

*Notices more than 180 days old are not viewable on the web portal.

✓



✓

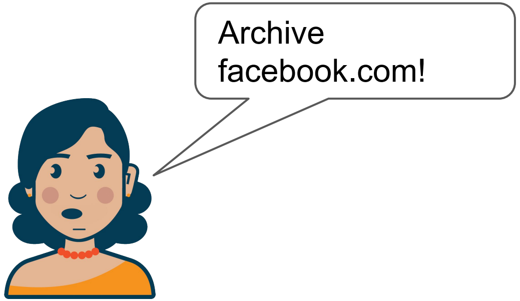


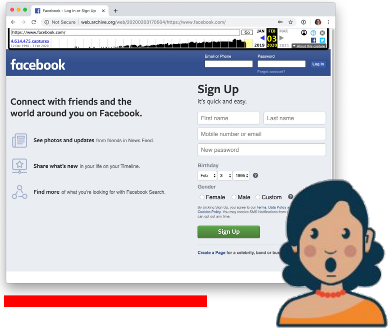
✓

Preserving Content That *Was* Behind Authentication

- By **reference** (URI) vs. **By-Value** (representation)*
- Proxy Approaches
  - MITM: warcprox, Webrecorder (link these, screenshots)
- Browser-based approach (by-value)
  - **WARCreate**

WARCreate
Google Chrome Extension




Facebook - Log In or Sign Up
Not Secure | web.archive.org/web/20200203170504/https://www.facebook.com/
https://www.facebook.com/
4,614,475 captures
12 Dec 1998 - 3 Feb 2020
JAN FEB 03 MAR 2019 2020 2021
About this capture

facebook
Email or Phone
Password
Log In
Forgot account?

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.

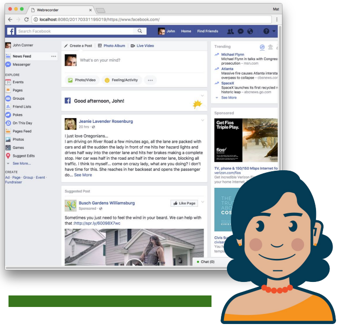
Find more of what you're looking for with Facebook Search.

Sign Up
It's quick and easy.
First name | Last name
Mobile number or email
New password
Birthday: Feb 3 1995
Gender: ○ Female ○ Male ○ Custom

By clicking Sign Up, you agree to our Terms, Data Policy a Cookies Policy. You may receive SMS Notifications from can opt out any time.

Sign Up

Create a Page for a celebrity, band or bus


Webrecorder
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.com
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 Fi your home internet

* Previously funded by grant, *"Archive What I See Now"*

NATIONAL ENDOWMENT FOR THE HUMANITIES

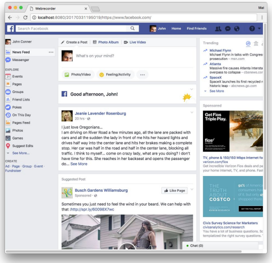
#HK-50181-14

Archival Replay

- Authentication mechanism is not exhibited once a page is preserved
  - For example, entering credentials for facebook.com does not send the request to facebook.com, which might not exist in the future.
- By value stores the resource representation
- This content might contain personally identifiable or sensitive information
  - sharing the captures to others needs strategic consideration.
- URI-M alone for access could make the memento publicly accessible (bad)

Sharing Captures

Check out my capture at
http://archive.alice.com/web/202002031205/https://facebook.com



Webrecorder

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

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

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

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

That probably should not be publicly available...

Memento TimeMaps

- Coalescence URI-Ms by the original URI (URI-R) they represented on the live web
  - **https://cnn.com**
  - **http://www.cnn.com**
  - **https://cnn.com/index.php**
  - **http://www2.cnn.com:80/homepage.html**

Various identifiers (original URI, URI-Rs) of the CNN homepage over time

- ...might all represent the same site in the past
- “Canonicalization” helps to normalize these variants for retrieval

Cross-archive Association

Live web “original URI” (URI-R)

```
<https://cnn.com>; rel="original",
<https://memgator.cs.odu.edu/timemap/link/https://cnn.com>; rel="self"; type="application/link-format",
<https://arquivo.pt/wayback/20000620180259mp_/http://cnn.com/>; rel="first memento"; datetime="Tue, 20 Jun 2000 18:02:59 GMT",
<http://web.archive.org/web/20000621011731/http://cnn.com:80/>; rel="memento"; datetime="Wed, 21 Jun 2000 01:17:31 GMT",
<http://wayback.vefsafn.is/wayback/20000621140928/http://cnn.com/>; rel="memento"; datetime="Wed, 21 Jun 2000 14:09:28 GMT",
<http://web.archive.org/web/20000621140928/http://cnn.com:80/>; rel="memento"; datetime="Wed, 21 Jun 2000 14:09:28 GMT",
<https://arquivo.pt/wayback/20000706192838mp_/http://cnn.com/>; rel="memento"; datetime="Thu, 06 Jul 2000 19:28:38 GMT",
<http://wayback.vefsafn.is/wayback/20000706192838/http://cnn.com/>; rel="memento"; datetime="Thu, 06 Jul 2000 19:28:38 GMT",
<http://web.archive.org/web/20000706192838/http://cnn.com:80/>; rel="memento"; datetime="Thu, 06 Jul 2000 19:28:38 GMT",
<http://wayback.archive-it.org/all/20150601215659/http://cnn.com/>; rel="memento"; datetime="Mon, 01 Jun 2015 21:56:59 GMT",
...
```

INTERNET ARCHIVE WayBackMachine

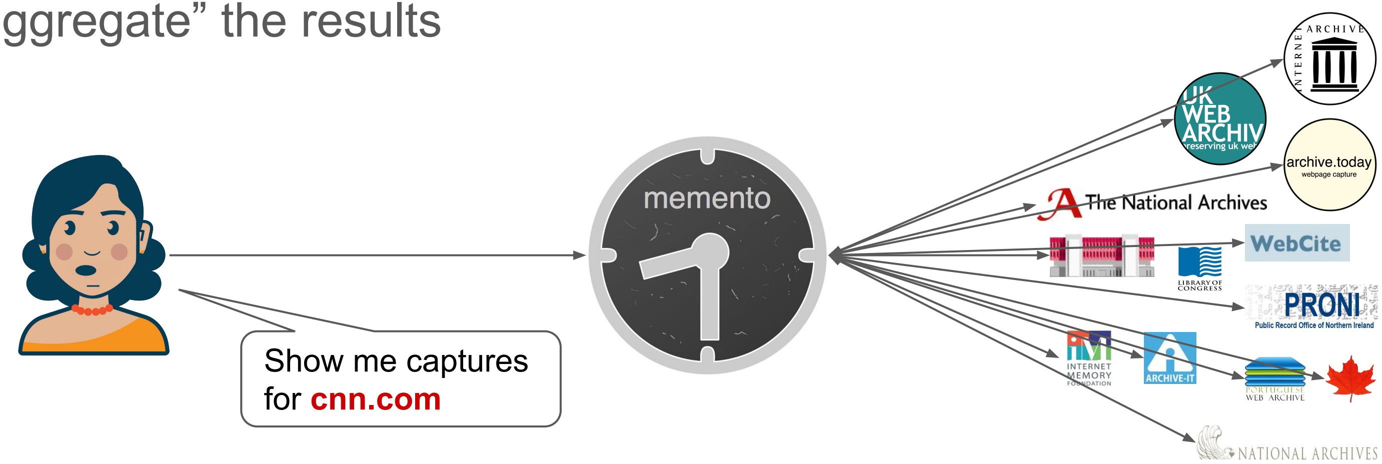
ARQUIVO.PT

ARCHIVE-IT

Vefsafn.is

Memento Aggregation

- A web archive is unlikely to provide links to other archives' captures (mementos)
- Aggregation services send a request for a **URI-R** (live web URI) to multiple archives and "aggregate" the results



Show me captures for **cnn.com**

Who Runs the Aggregators?

- Los Alamos National Laboratory hosts their Memento aggregator
  - e.g., http://timetravel.mementoweb.org/timemap/link/http://cnn.com
- The set of archives it uses is unchangeable by the querying user
- When new archives come about, to be included:
  - The archive's **MUST** be Memento compliant
  - The archive's endpoints **MUST** be manually added by someone at LANL
- The captures on *your machine* will never be included in these results.

Example TimeMap from a Memento Aggregator

```
<http://drexel.edu>; rel="original",
<http://some.aggregator.com/timemap/link/http://drexel.edu>; rel="self"; type="application/link-format",
<http://web.archive.org/web/20060514123511/http://drexel.edu:80/>; rel="first memento"; datetime="Sun, 14 May 2006
12:35:11 GMT",
<http://web.archive.org/web/20060516213852/http://www.drexel.edu/>; rel="memento"; datetime="Tue, 16 May 2006 21:38:52
GMT",
...
<http://web.archive.org/web/20180128152125/http://drexel.edu>; rel="memento"; datetime="Sun, 28 Jan 2018 15:21:25 GMT",
<http://web.archive.org/web/20180319141920/http://drexel.edu/>; rel="last memento"; datetime="Mon, 19 Mar 2018 14:19:20
GMT",
<http://some.aggregator.com/timemap/link/http://drexel.edu>; rel="timemap"; type="application/link-format",
<http://some.aggregator.com/timemap/json/http://drexel.edu>; rel="timemap"; type="application/json",
<http://some.aggregator.com/timemap/cdxj/http://drexel.edu>; rel="timemap"; type="application/cdxj+ors",
<http://some.aggregator.com/timegate/http://drexel.edu>; rel="timegate"
```

Original URI (URI-R)

Other TimeMaps (URI-Ts)

TimeGate (URI-G)

Relative Relations

MemGator

- *Open Source* Memento aggregator software
- Individuals can easily deploy an aggregator *onto their system*.
- Allows users to define which archives are queried
  - provides a comprehensive set of defaults

S. Alam and M. Nelson, MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go, Proceedings of JCDL 2016, pp. 243-244, 2016.

https://github.com/oduwsdl/memgator

A Framework for Aggregating Private and Public Web Archives

- Aggregation with private web archives
- Client-side archive specification

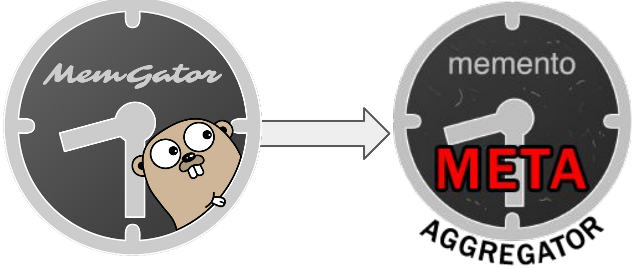
- Authentication layer to systematically interface with private Web archives

- Archival negotiation in dimensions beyond time

more information:

M. Kelly et al., "A Framework for Aggregating Private and Public Web Archives," In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2018, pp. 273-282.

Memento Meta Aggregator

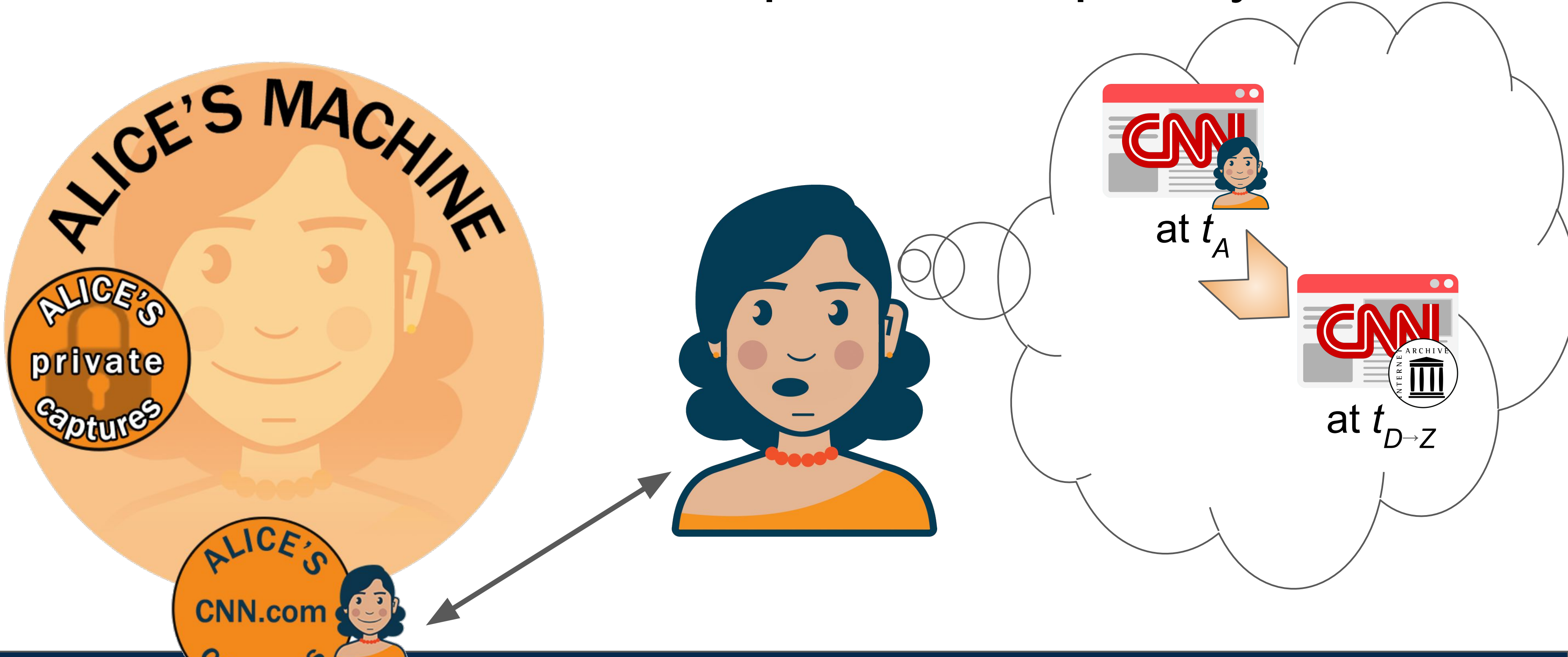


- Enhanced MemGator to enable more powerful client-side querying
    - Includes client specifying set of archives to aggregate
- Empowers querying client to:
    - optionally specify the set of archives aggregated, beyond what might be done by default
    - Specify archival precedence
    - Interface with mementos that require access beyond simply dereferencing URI-M
    - ...and more.

Personal Archive Aggregation: Alice Saves the Web

ALICE'S MACHINE
ALICE'S
private
captures
ALICE'S
CNN.com
captures

Alice Wants to See Her Captures Temporally Inline



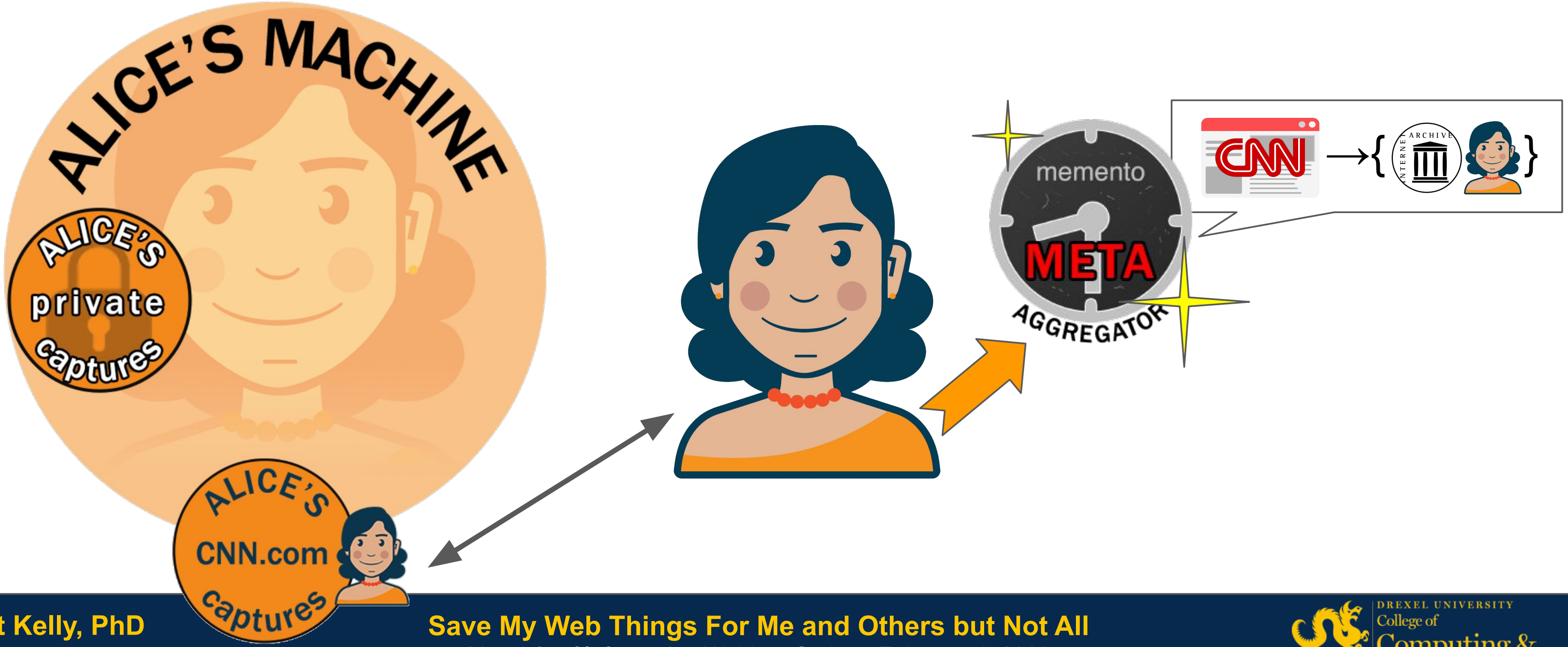
Personal Archive Aggregation

Mementity Dynamics - Alice & Her Archives

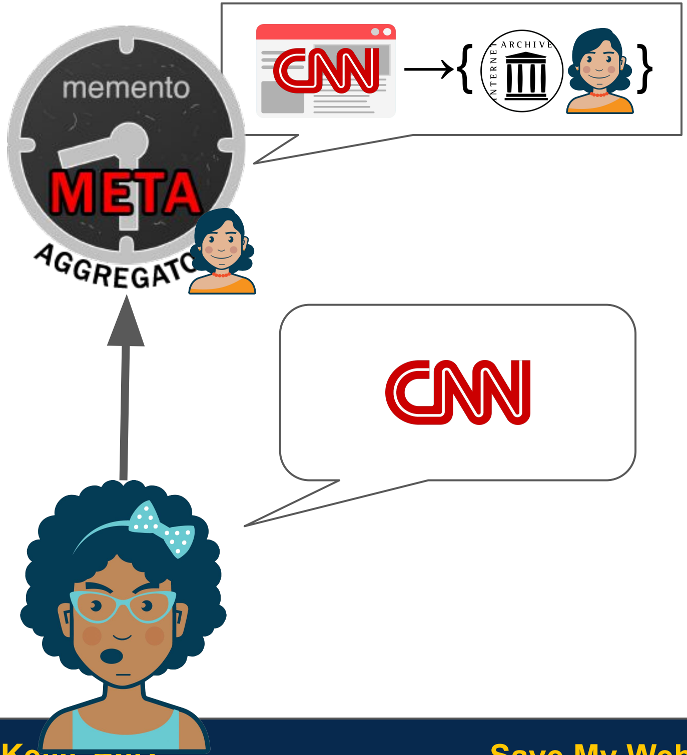
ALICE'S MACHINE
ALICE'S
private
captures
ALICE'S
CNN.com
captures
CNN
INTERNET ARCHIVE

Personal Archive Aggregation

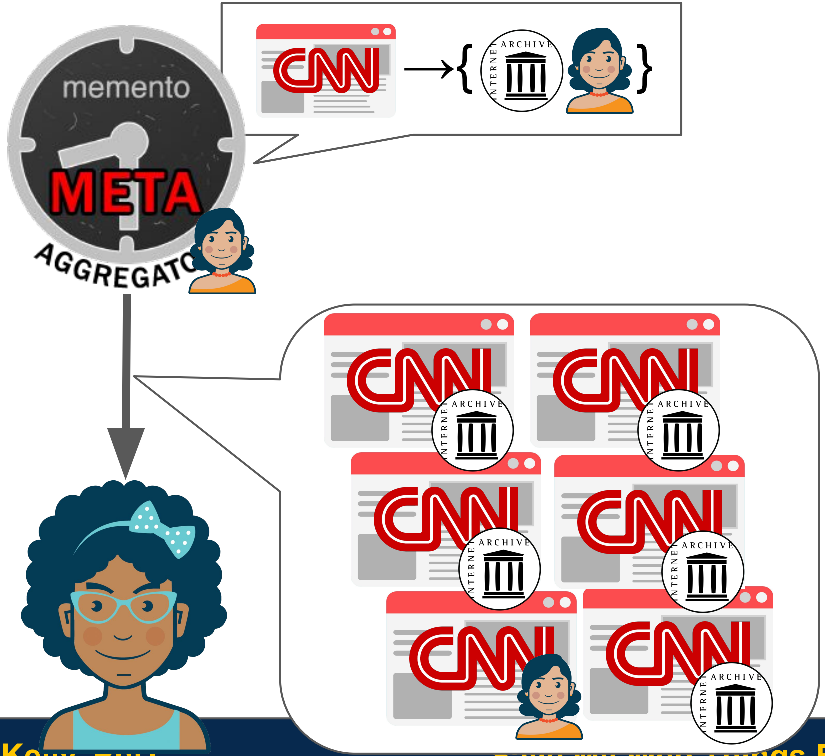
Alice Deploys MMA



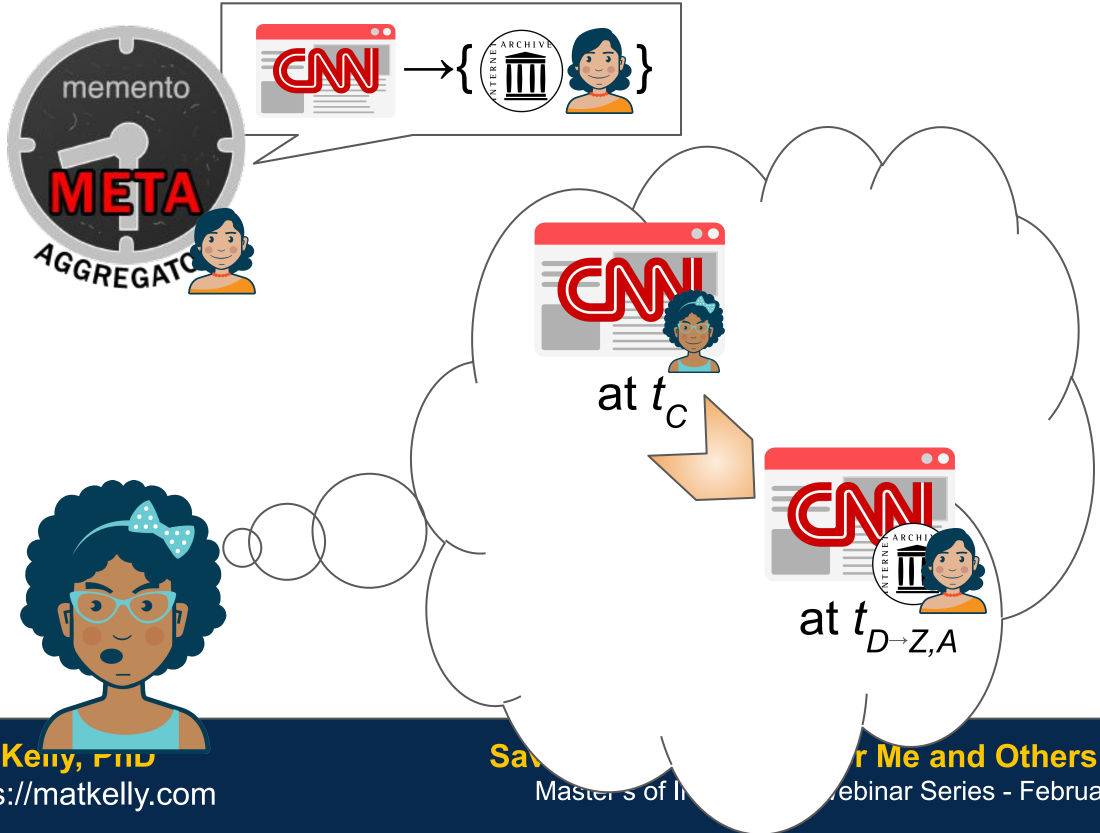
Carol Asks MMA for CNN



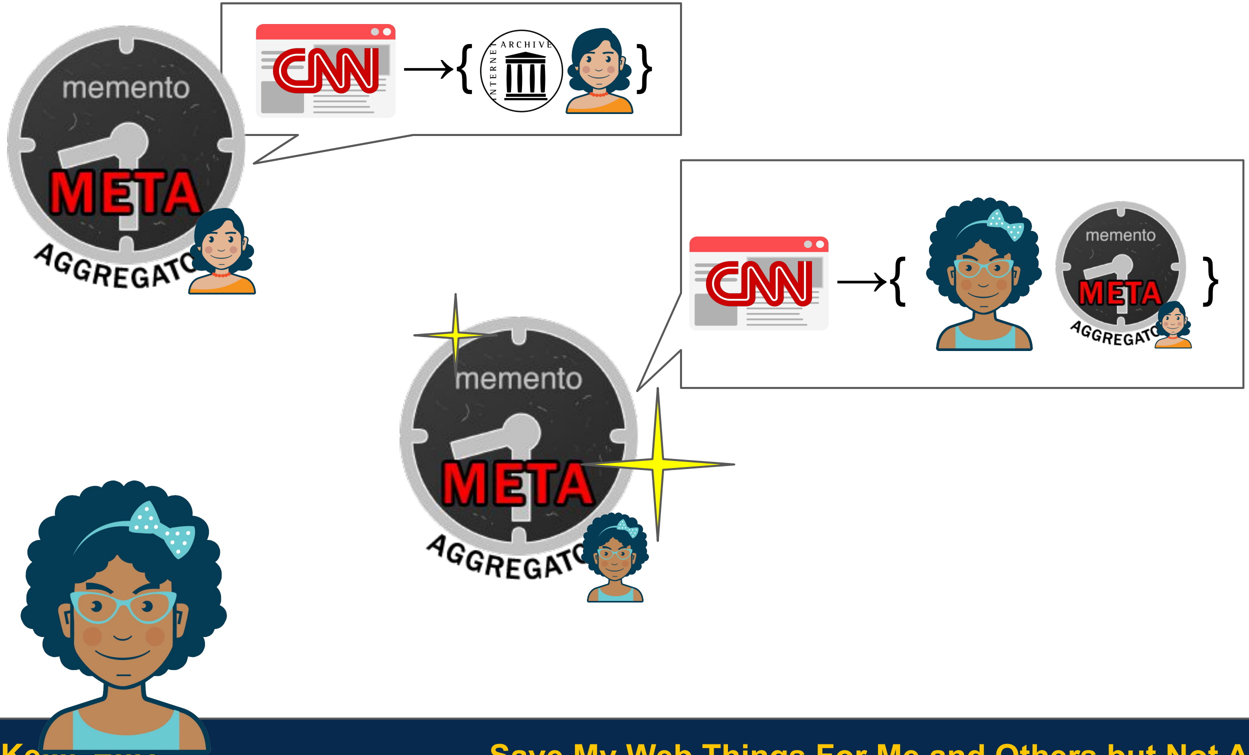
MMA Returns CNN Memento



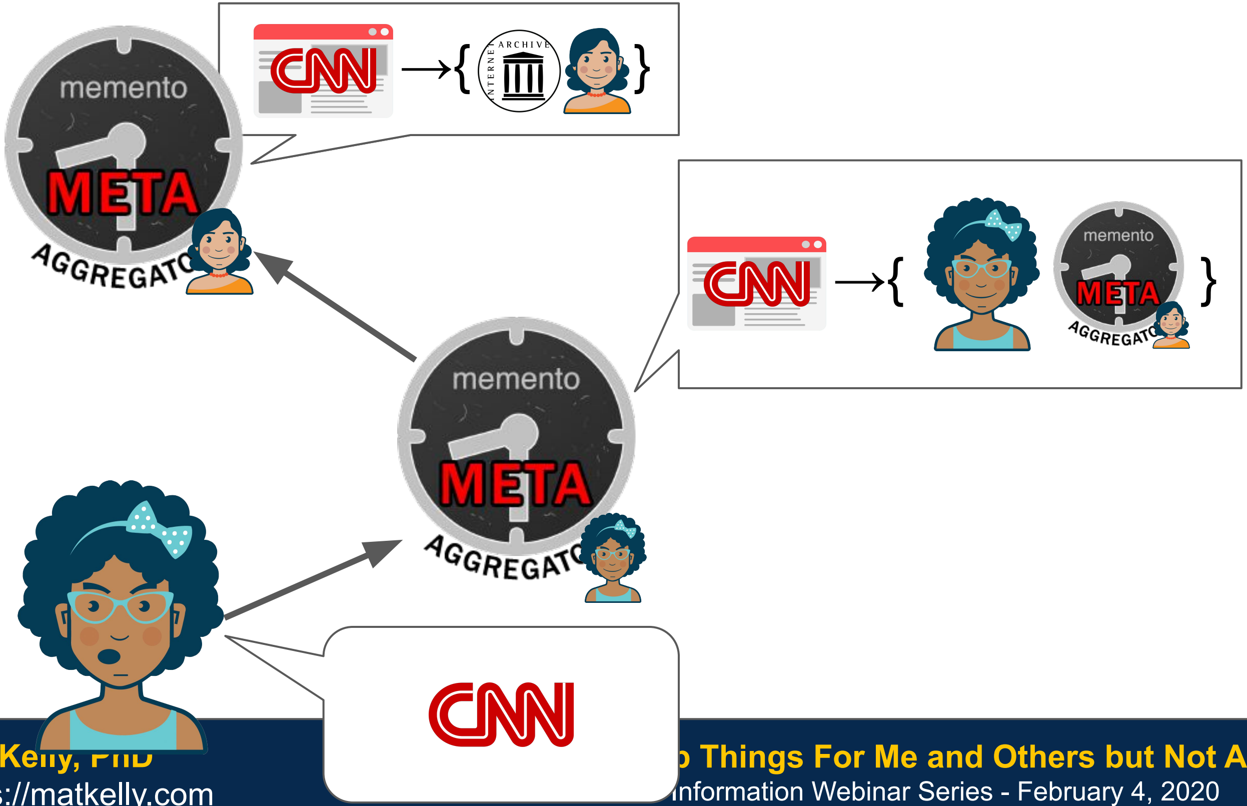
Carol Wants to Aggregate Her Own Captures



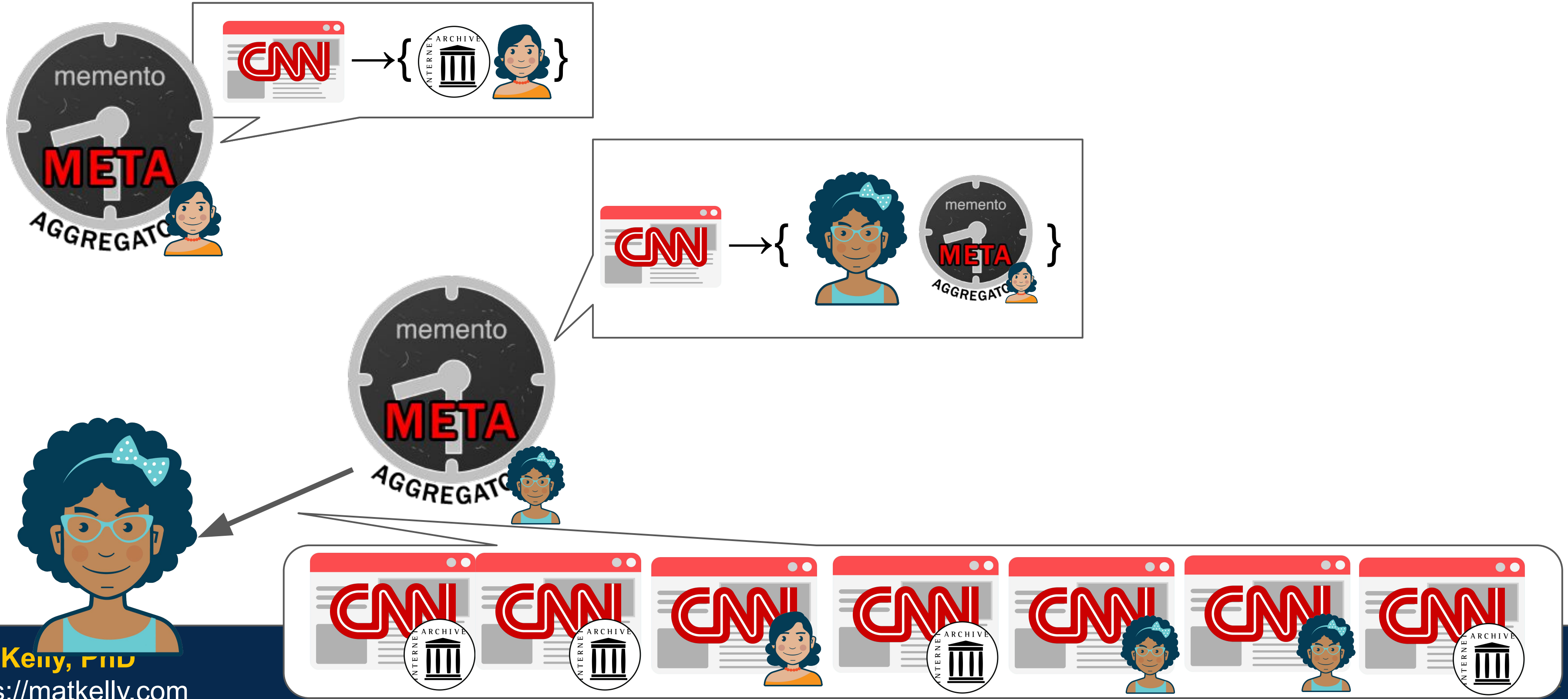
Carol Spawn Her Own MMA to Access Alice’s Web Archive and Alice’s MMA



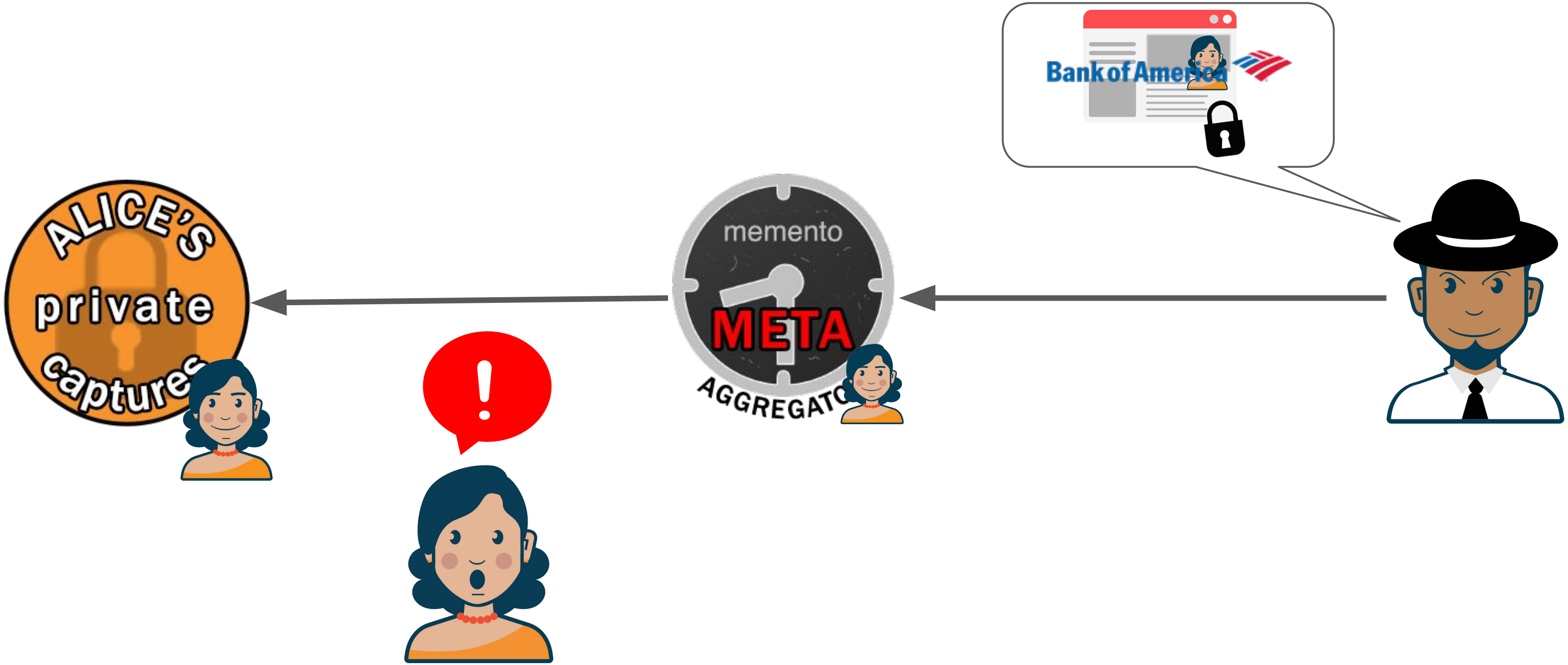
Carol Asks Her MMA For CNN



Carol's MMA Returns CNN Mementos from the Sources She Specified



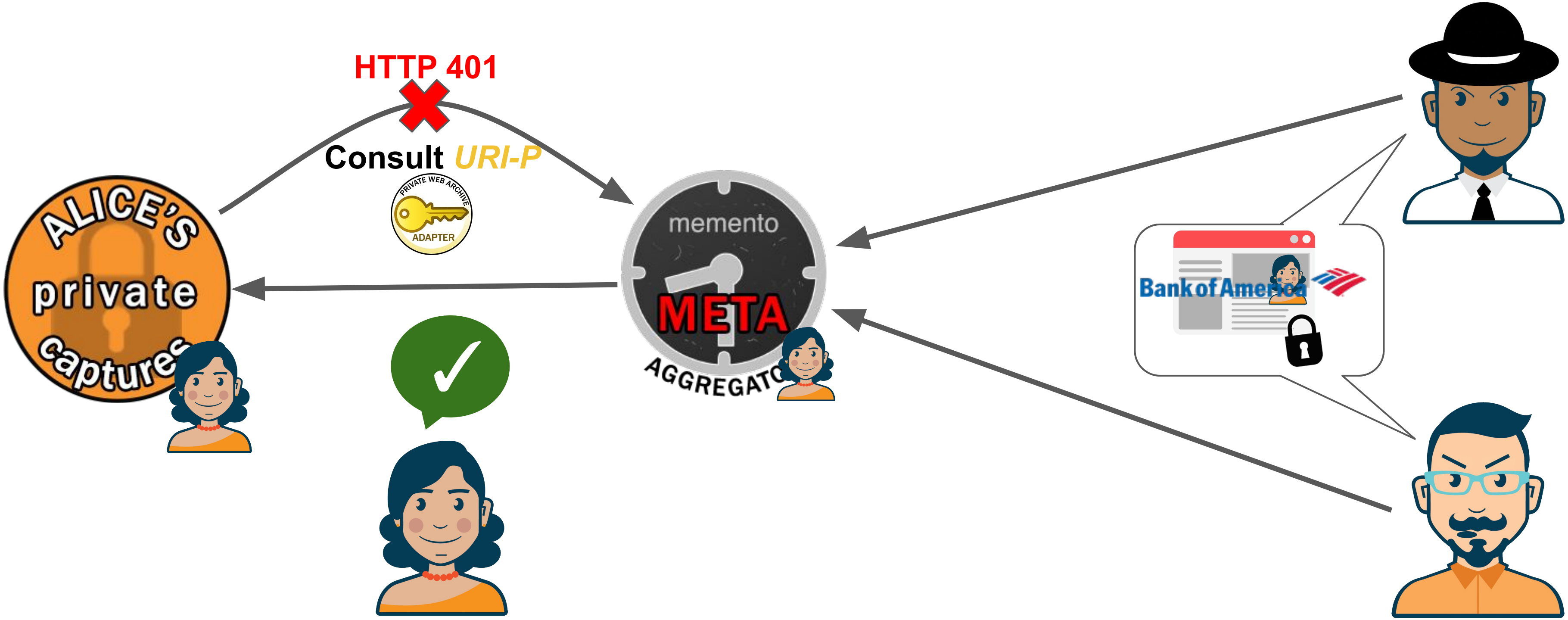
Hooray, Aggregation!



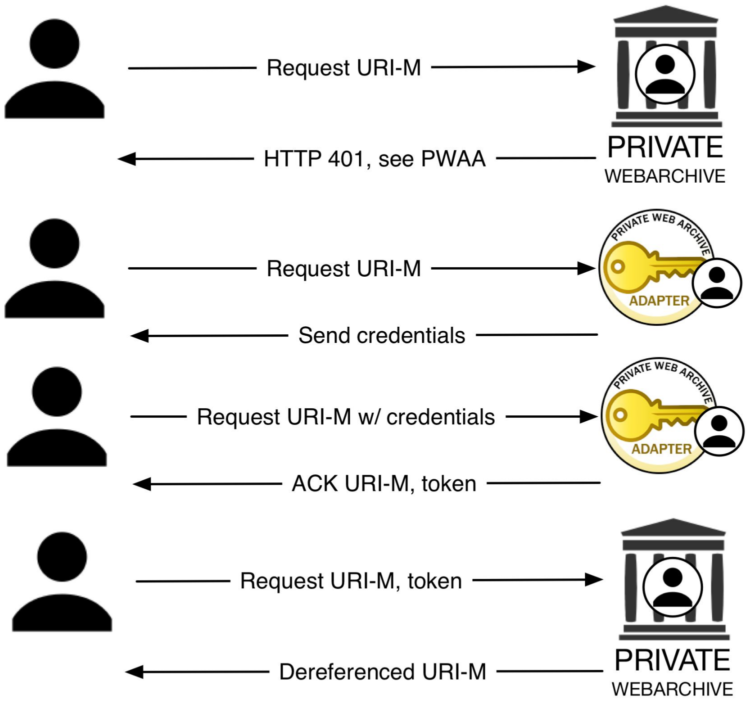
Regulating Access to Private Captures

- Some form of access control should be used to regulate private web archives to prevent information leakage and exposure of sensitive information.
- Private Web Archive Adapter
  - Enables systematic (OAuth2) based access control to private web archives
- Delegates authentication/authorization to a separate “entity”
  - archives remain functionally cohesive
  - decouples archives from authentication role to facilitate interoperability

Dereferencing Beyond Simple URI Access



Adaptation of OAuth2 to Web Archives

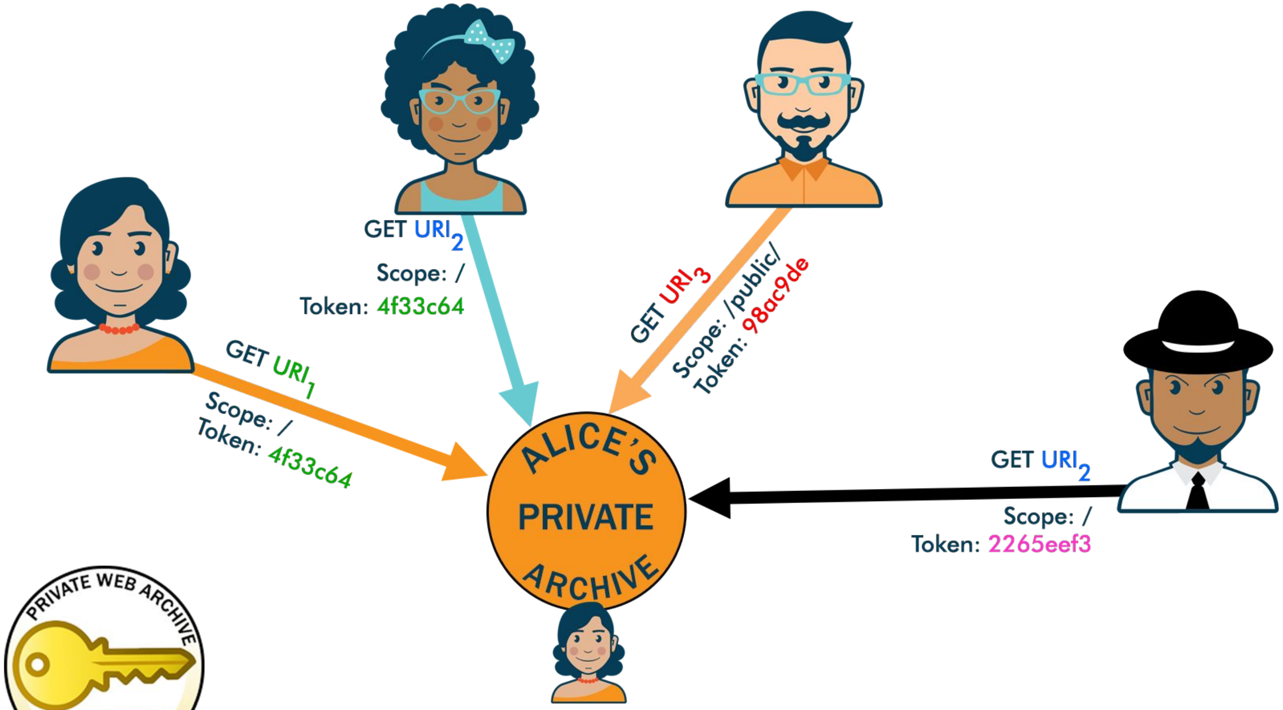


- Auth Layer for to encourage Private Web archive aggregation
- Typical OAuth 2.0 flow
- Auth role cohesive to PWAA
- Persistent access through tokenization

Using Auth Tokens for Private Web Archive Access

Token grants can be:

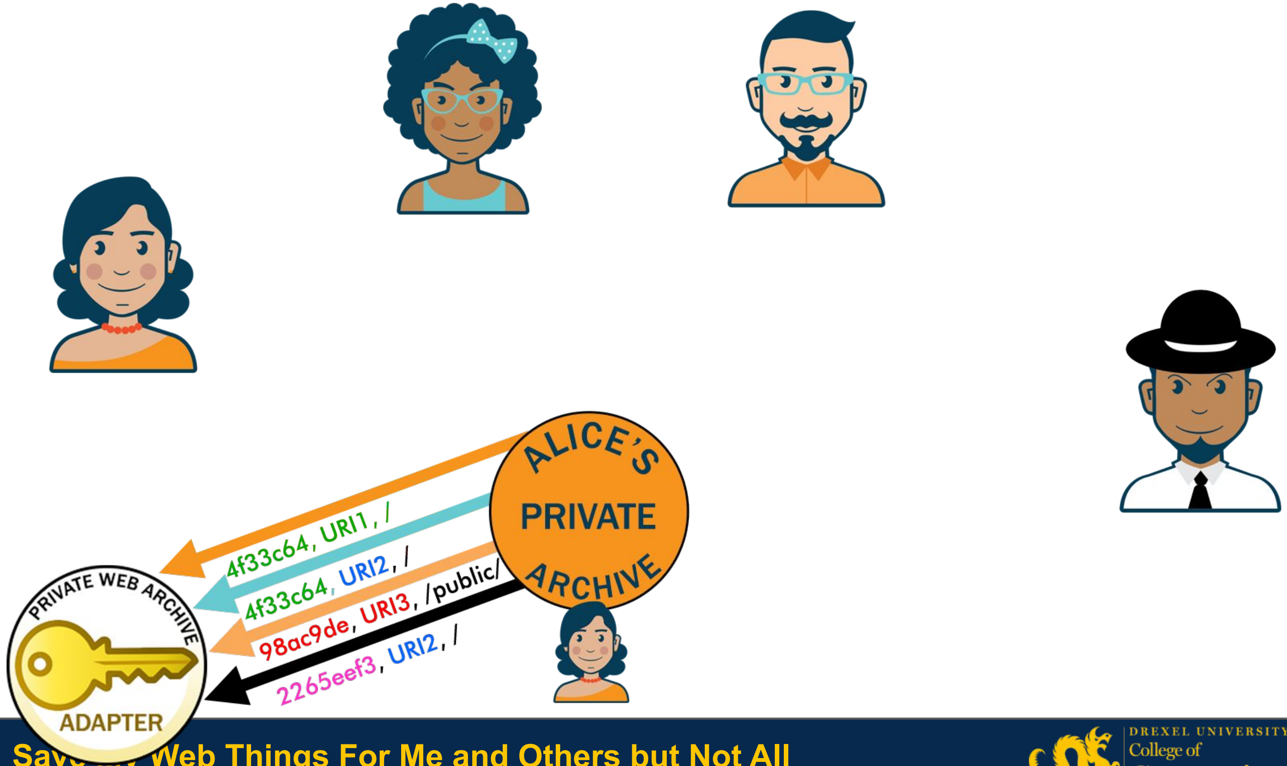
- Shared
- Disavowed
- Limit scope



Using Auth Tokens for Private Web Archive Access

Token grants can be:

- Shared
- Disavowed
- Limit scope



Richer Querying of Web Archives Makes Them More Useful

Memento introduces negotiation in **time**, also useful would be negotiation:

- private/public captures, only, first, if-and-only-if conditional
- Memento quality above a threshold*
  - Crawlers sometimes miss necessary resource representations!
- Only mementos that are not redirects†
  - Over 80% of the google.com URI-Ms at Internet Archive are HTTP 3XX redirects!

* J. Brunelle et al., “Not All Mementos Are Created Equal: Measuring the Impact of Missing Resources,” International Journal on Digital Libraries (IJDL), 16(3), pp. 283-301, 2015.

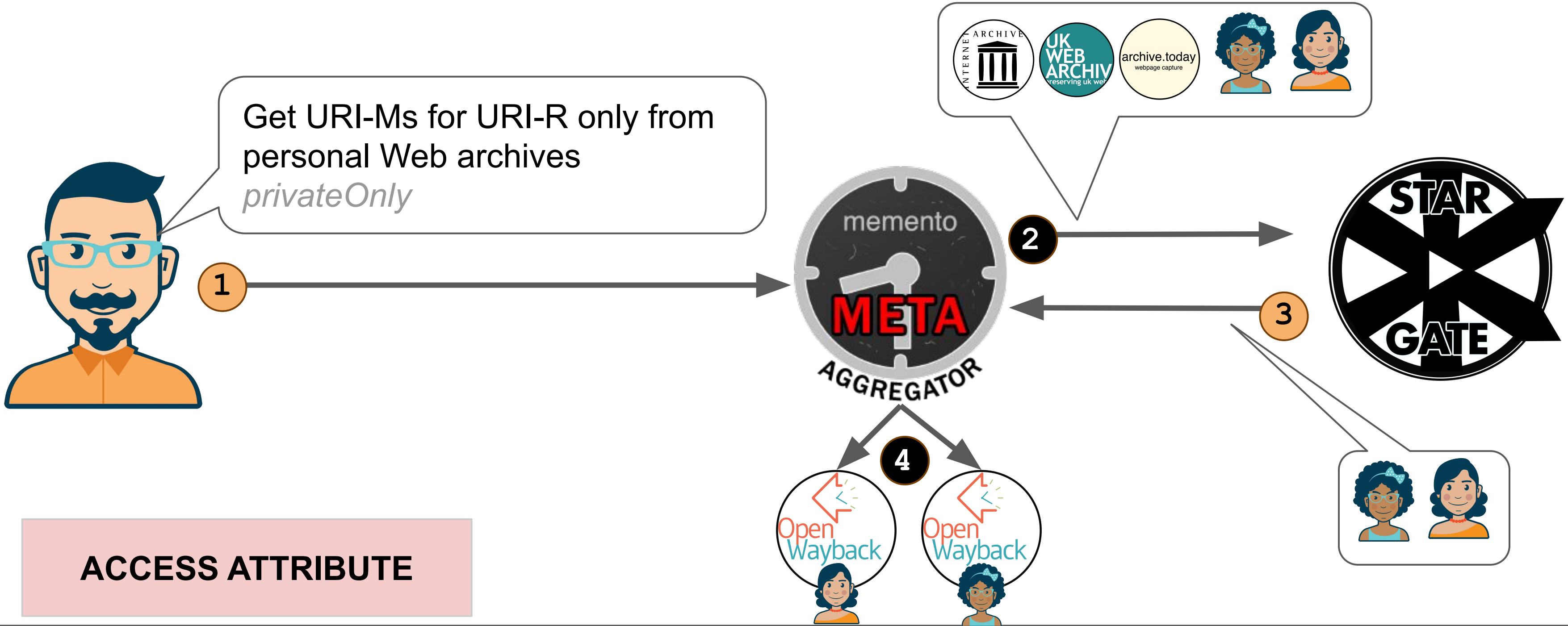
† M. Kelly et al., “Impact of URI Canonicalization on Memento Count,” Technical Report arXiv:1703.03302, 2017.

There exists a need in querying web archives to...

- Filter results on the server side (cf. making users sort through verbose results)
- Specify the sources used in aggregation (e.g., include *my* captures)
- Allow aggregators to optionally comply
- Allow clients/users of web archives to express these PREFERENCES

- The **StarGate** “concepts” facilitates
  content negotiation with web archives in dimensions beyond time

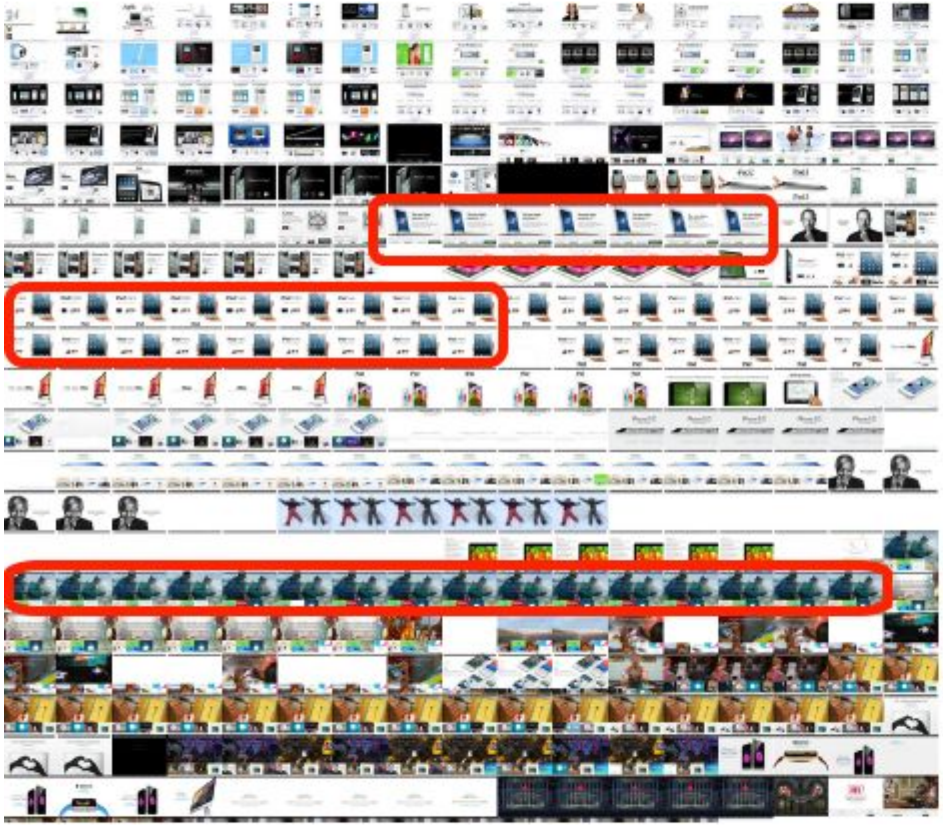
Despite Your Sources, Use A Subset



ACCESS ATTRIBUTE

Many Nearly Identical

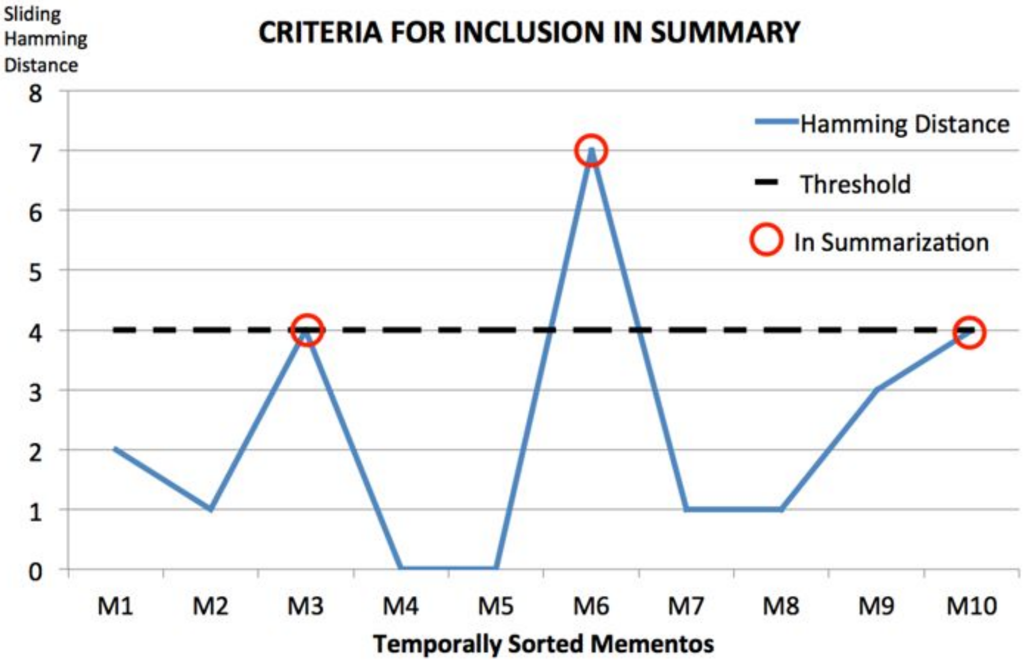
- Temporal adjacent captures might be similar or identical
- Finding a summary of representations in time can be expensive
- Generating screenshots of all would be wasteful



Identifying Similarity by Calculating Hamming Distance

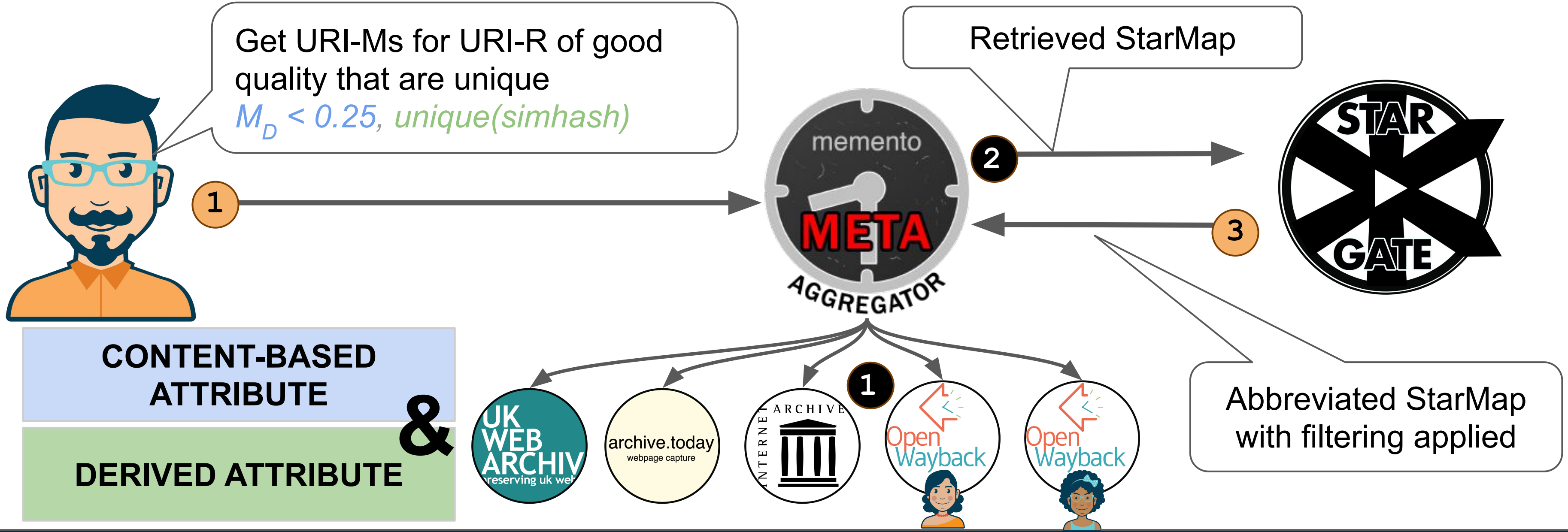
		HAMMING DISTANCE
HTML of apple.com March 3, 2008	c39f0abc...b9	N/A pivot
HTML of apple.com March 5, 2008	c39d0abc...c9	2
HTML of apple.com April 12, 2008	c39d0abc...b9	1
HTML of apple.com October 4, 2008	c770ad1b...b9	7

Use dissimilarity from pivot as basis for inclusion



StarGate

- Allows for content negotiation with archives in dimensions beyond time
  - e.g., Bob asks for only captures from the set of archives that are of a sufficient quality of capture.



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

Retrieved StarMap

Abbreviated StarMap with filtering applied

CONTENT-BASED ATTRIBUTE

&

DERIVED ATTRIBUTE

Summation

SAVE MY WEB THINGS FOR ME

Preserve content behind authentication

...AND OTHERS

Use an Aggregator for others’ access

...BUT NOT ALL

Use a PWAA for Access Control

...and further relevant topics beyond the title!

Open-Ended Research Efforts

- Efficient querying in numerous dimensions to mitigate the combinatorial explosion of permutations
- Make software easy, intuitive, native, self-hosted
  - Users that preserve might not be technical, this should not be a barrier
- Facilitate resilience -- allow users to have captures **safely** distributed
  - See InterPlanetary Wayback https://github.com/oduwsdl/ipwb
- Introduce more sophisticated, stronger ways to regulate access

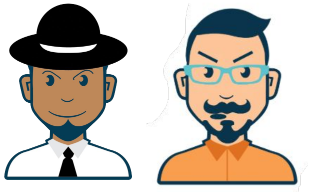
Save *My* Web Things For Me

and Others

but Not All

Mat Kelly, PhD

Assistant Professor, Information Science
Drexel University
College of Computing & Informatics
mkelly@drexel.edu
https://matkelly.com



Some figures adapted from work of Agata Krych CC-BY-SA 4.0

Master’s of Information Webinar Series
February 4, 2020

Drexel UNIVERSITY