

# Web Archiving

Mat Kelly, PhD

Assistant Professor, Information Science  
Drexel University, College of Computing & Informatics (CCI)  
Philadelphia, PA

[mkelly@drexel.edu](mailto:mkelly@drexel.edu)  
<https://matkelly.com>  
[@machawk1](https://twitter.com/machawk1)

LIS Education And Data Science Integrated Network Group (LEADING) Bootcamp  
Week 2: June 14, 2021

# What We Will Cover

- Archival Crawling
- Access and Replay
- Formats and metadata
- Services
  - Internet Archive
  - Archive-It
  - WebRecorder
- Tools

# The Web

A screenshot of a web browser showing the LEADING: LIS Education And Data Science Integrated Network Group website. The URL is [ccr.drexel.edu/mrc/leading/](http://ccr.drexel.edu/mrc/leading/). The page features the Drexel University logo and the Metadata Research Center logo. The main navigation menu includes HOME, ABOUT, RESEARCH, PUBLICATIONS & SCHOLARLY ACTIVITY, PEOPLE, and NEWS & EVENTS. Below the menu, the text "LEADING: LIS Education And Data Science Integrated Network Group" is displayed. Two prominent blue links are shown: "LEADING Home | People |" and "LEADING Fellows | Application". A note below states, "\*\*Application deadline for 2021 is closed\*\*". The LEADING logo, consisting of three red curved lines above the word "LEADING", is also present. The footer contains text about the project's funding and partners.

LEADING: LIS Education And Data Science  
Integrated Network Group

**LEADING Home | People |**  
**LEADING Fellows | Application**

\*\*Application deadline for 2021 is closed\*\*

The LIS Education and Data Science Integrated Network Group (LEADING) is a Laura Bush 21st Century Librarian (LB21) National Digital Infrastructures and Initiatives project, supported by the Institute of Museum and Library Services (IMLS). The LEADING

# Our Web

The screenshot shows the LEADING: LIS Education And Data Science Integrated Network Group website. At the top, there's a header with the Drexel University logo, the Metadata Research Center name, and a sub-header 'College of Computing & Information'. Below the header, a navigation bar includes links for HOME, ABOUT, RESEARCH, PUBLICATIONS & SCHOLARLY ACTIVITY, PEOPLE, and NEWS & EVENTS. The main content area features a yellow banner with the text 'LEADING: LIS Education And Data Science Integrated Network Group'. Below the banner, there are two large blue buttons: 'LEADING Home | People' and 'LEADING Fellows | Application'. A note at the bottom left says '\*\*Application deadline for 2021 is closed\*\*'. To the right, there's a logo for 'LEADING' with three red curved lines above the word 'LEADING'. Below the logo is a small text: 'LIS Education And Data Science Integrated Network Group'. At the very bottom, there's a note: 'The LIS Education and Data Science Integrated Network Group (LEADING), is a Laura Bush 21st Century Librarian (LB21) National Digital Infrastructures and Initiatives project, supported by the Institute of Museum and Library Services (MLS). The LEADING'.

The screenshot shows the Twitter profile for the account @all\_metadata. The profile picture is a purple circle with white text '<MRC>'. The bio reads: 'All things #metadata: The Metadata Research Center aims to advance research in metadata, semantics, and ontologies @DrexelCCI.' It also mentions 'Drexel University' and the URL 'ccid.drexel.edu/mrc'. The account has 259 tweets, 231 followers, and 574 following. The 'Tweets' tab is active. A recent tweet from May 21, 2021, states: 'Mid-May, 25 LEADING Fellows started their online data science curriculum. In June, they will attend a virtual boot camp, which will involve more interactivity.' Below the tweet is a link: 'Visit the LEADING website to learn about our fellows: ccid.drexel.edu/mrc/leading/fellows'. The 'What's happening' section shows a live tweet from E3 2021: 'E3 2021: Xbox, Bethesda and Square Enix expected to feature on Day 2'.

# My Web

The screenshot shows the homepage of the LEADING website. At the top, there's a yellow header bar with the Drexel University logo and the Metadata Research Center branding. Below this, a navigation bar includes links for HOME, ABOUT, RESEARCH, PUBLICATIONS & SCHOLARLY ACTIVITY, PEOPLE, and NEWS & EVENTS. The main content area features a large yellow banner with the text "LEADING: LIS Education And Data Science Integrated Network Group". Below the banner, there are two prominent buttons: "LEADING Home | People" and "LEADING Fellows | Application". A note below the buttons states, "Application deadline for 2021 is closed\*\*". To the right, there's a logo for "LEADING" with three red curved lines above the word "LEADING". Below the logo, it says "LIS Education And Data Science Integrated Network Group". At the bottom left, there's a paragraph about the project being supported by the Institute of Museum and Library Services (IMLS). The bottom right contains a small image of a baby.

The screenshot shows the Twitter profile page for the MRC (@all\_metadata). The profile picture is a purple circle with white text. The bio reads: "All things #metadata: The Metadata Research Center aims to advance research in metadata, semantics, and ontologies @DrexelCCl." It also mentions "Drexel University" and the URL "cc.drexel.edu/mrc". The account has 259 tweets, 231 followers, and 574 following. There are sections for "Tweets", "Tweets & replies", "Media", and "Likes". A tweet from May 21 is highlighted: "Mid-may, 25 LEADING Fellows started their online data science curriculum. In June, they will attend a virtual boot camp, which will involve more interactivity." Below the tweets, there's a "What's happening" section with a link to "E3 2021: Xbox, Bethesda and Square Enix expected to feature on Day 2".

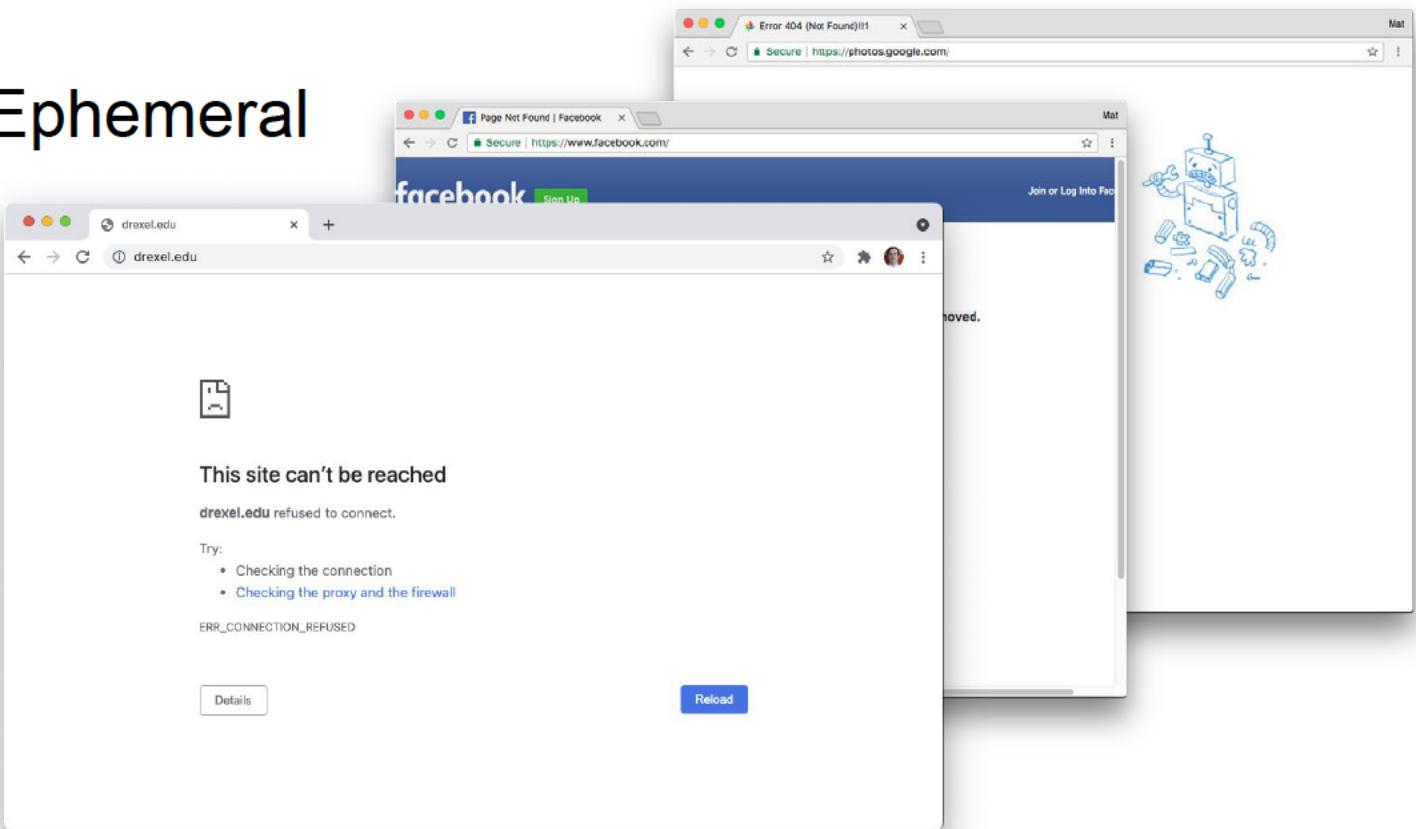
The screenshot shows a Google Photos album titled "Scarlett" from August 3 to September 21. The album contains several photos of a family, including a man, a woman, and a baby. One photo shows the man holding the baby. The interface includes a search bar, a "You might like" section with other users, and a "What's happening" feed at the bottom.

# The Web is Ephemeral

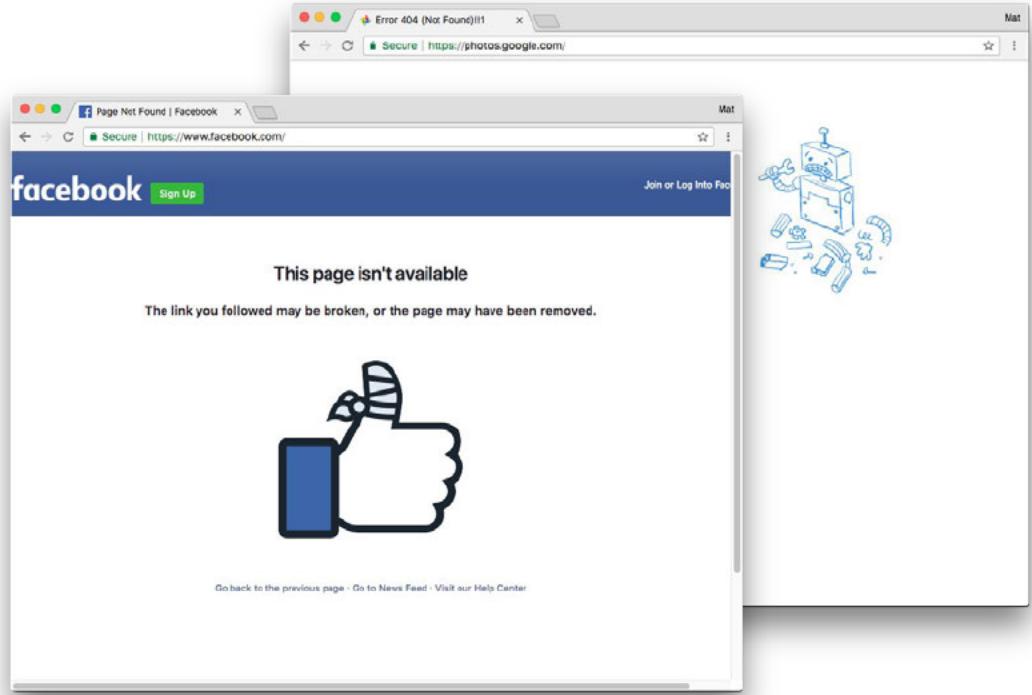
A screenshot of a web browser window displaying the Drexel University admissions page. The page features a yellow header bar with a 'Latest COVID-19 Updates' button. Below this is a navigation bar with links for Current Students, Faculty & Staff, Alumni, Parents, MAKE A GIFT, and a search bar. The main content area has a large banner with the text 'THIS IS YOUR MOMENT' over a background image of a campus building and trees. To the right, there's a sidebar with news items: 'Urban Ethnographer Elijah Anderson to Address Graduates at Drexel's University-wide Commencement', 'Drexel University Announces 2021 Commencement Speakers and Honorary Degree Recipients', and 'Dragons Unite to Bring Vaccines, Resources to the Surrounding West Philadelphia Community'. At the bottom, there are links for Apply, Visit, Connect, Giving, Contact, and Follow, along with a 'View All Campus News' link.

A screenshot of a web browser window showing a 404 Not Found error for the Facebook homepage. The URL in the address bar is https://www.facebook.com/. The page displays the standard 404 error message: 'Page Not Found | Facebook' and 'Secure | https://www.facebook.com/'. To the right of the browser window, there is a cartoon illustration of a blue robot or computer character sitting on a chair, looking sad and holding its head, with a speech bubble saying 'I'm sorry, I can't help you.' The overall theme of the slide is that the web is ephemeral and can change rapidly.

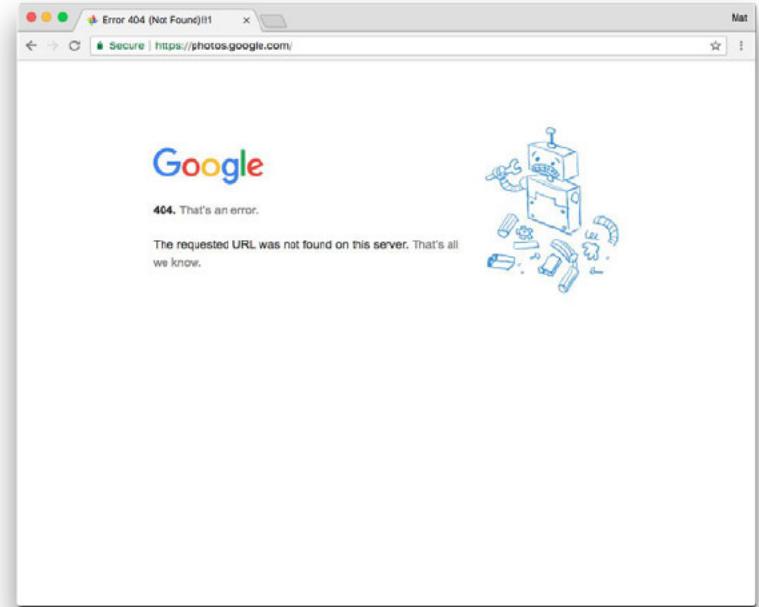
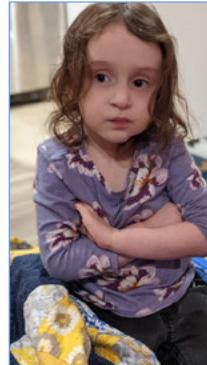
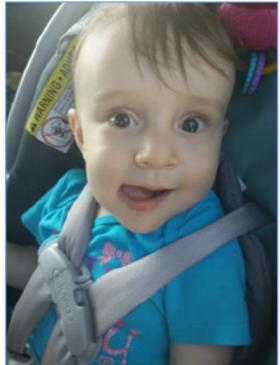
# The Web is Ephemeral



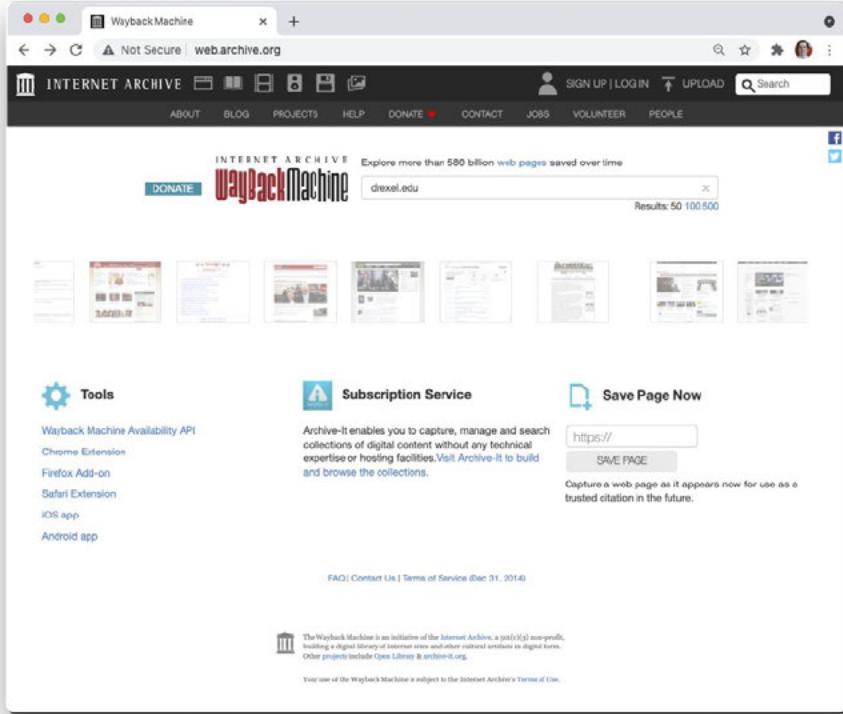
# The Web is Ephemeral



# The Web is Ephemeral

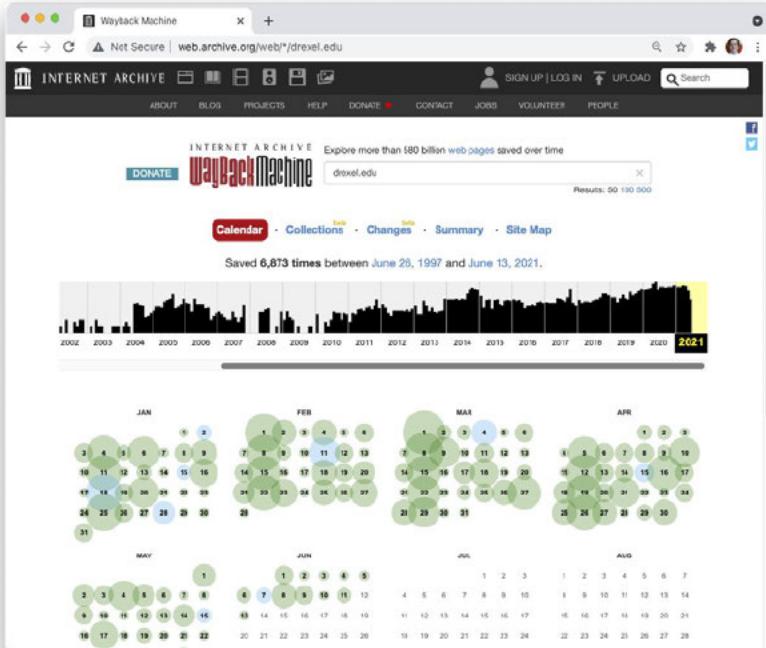


# Web Archives to the Rescue: Typical Access



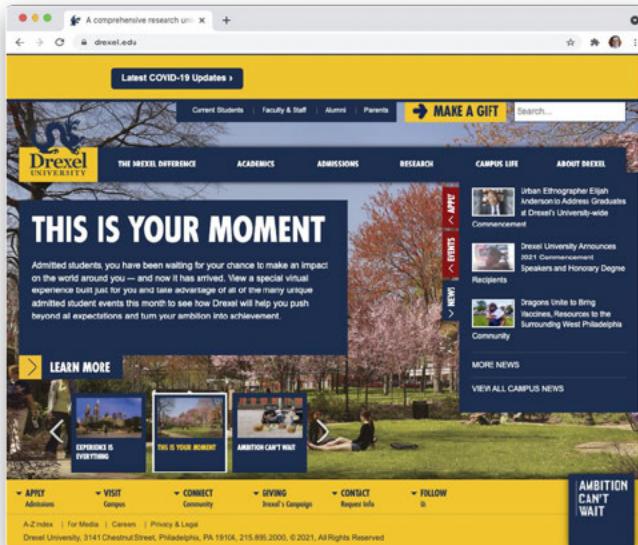
1. Go to `archive.org` in your browser
2. Enter the URL you want to see in the past in the form field
3. Submit your query

# Web Archives to the Rescue: Typical Access



4. Locate the archived Web page on the calendar or histogram view
5. Select the year/selection for the day
6. Repeat until you find the closest date and time

# Web Archiving: View The Web of the Past



Now

Web Archiving  
Week 2: June 14, 2021

LEADING Bootcamp  
Mat Kelly, PhD

# Web Archiving: View The Web of the Past

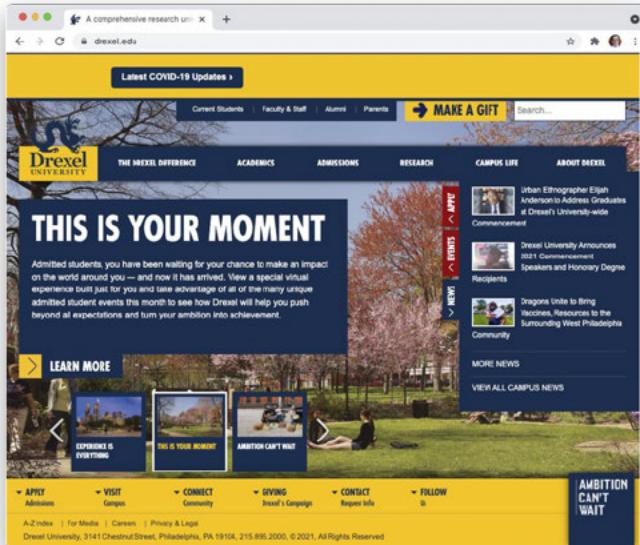
A screenshot of the Drexel University website homepage. The header features a yellow bar with links for "Latest COVID-19 Updates", "Current Students", "Faculty & Staff", "Alumni", "Parents", "MAKE A GIFT", and a search bar. Below the header, the main content area has a large banner with the text "THIS IS YOUR MOMENT". It includes a subtext about admitted students and their opportunities. On the right side of the banner, there are three small images with captions: "Urban Ethnographer Elijah Anderson to Address Graduates at Drexel University-wide Commencement", "Drexel University Announces 2021 Commencement Speakers and Honorary Degree Recipients", and "Dragons Unite to Bring Resources, Resources to the Surrounding West Philadelphia Community". Below the banner are sections for "MORE NEWS" and "VIEW ALL CAMPUS NEWS". At the bottom, there are navigation links for "APPLY", "VISIT", "CONNECT", "SAVING", "CONTACT", and "FOLLOW". The footer contains links for "A-Z Index", "For Media", "Careers", "Privacy & Legal", and copyright information.

Now

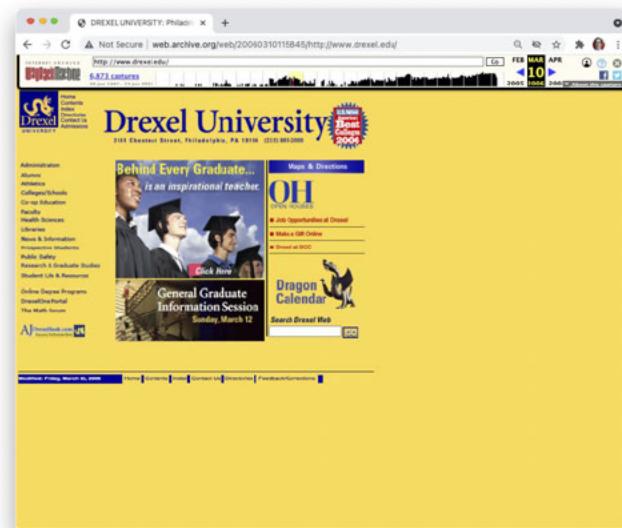
A screenshot of the Drexel University website homepage from June 14, 2011. The layout is different, with a blue header and a larger central content area. The main headline reads "Congratulations 2011 Grads!". Below it, there's a subtext about the online commencement experience. To the right, there are several images of graduates. The sidebar on the left lists "Prospective Students", "Current Students", "Faculty & Staff", "Parents", and "Alumni". The sidebar on the right lists "Upcoming Events" (Jun 16 7:00 PM - A Headlin' U Meet Joe Singletor) and "Academic Calendars". At the bottom, there's a banner for the "WACE 17th Annual World Association of Cooperative Education Conference". A red arrow points from the "Now" screenshot to this one.

June 14, 2011

# Web Archiving: View The Web of the Past

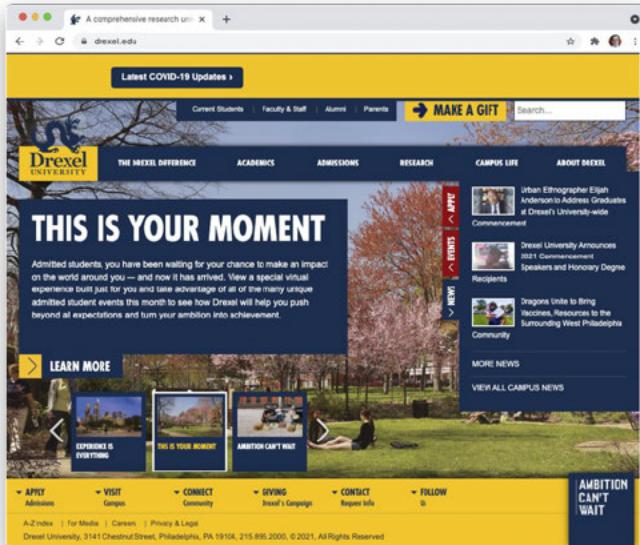


Now

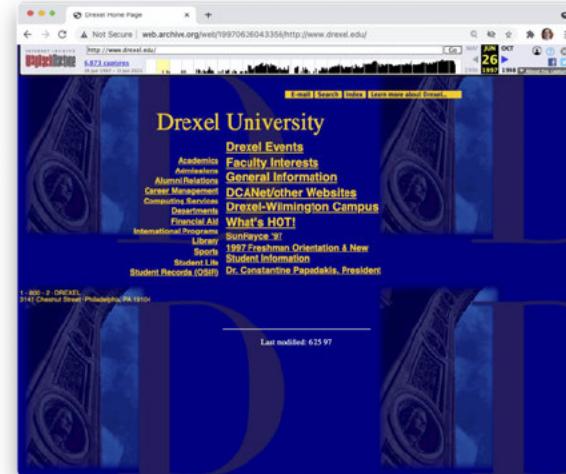


March 10, 2006

# Web Archiving: View The Web of the Past

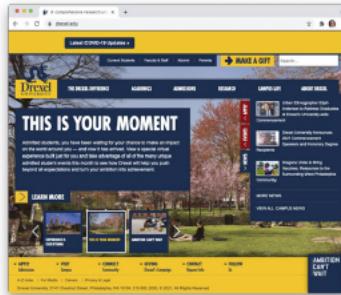


Now

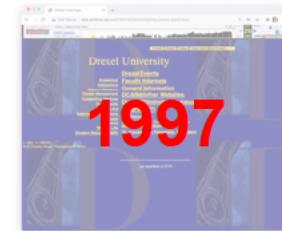


March 26, 1997

# Web Archiving



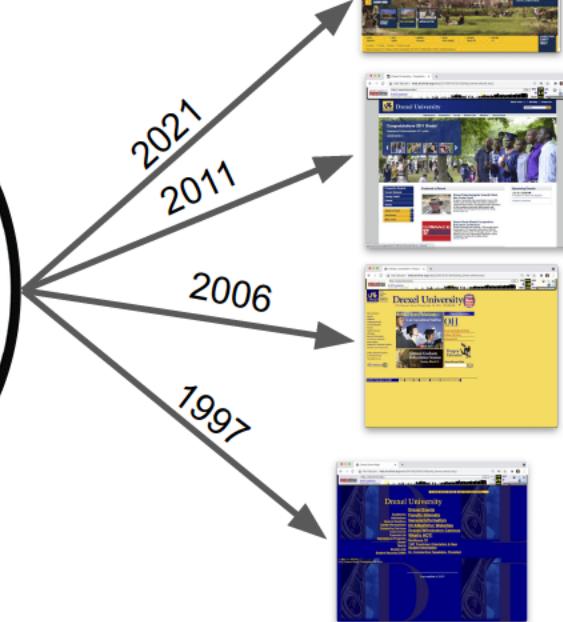
Associate a live Web URI  
with their **archived representations**



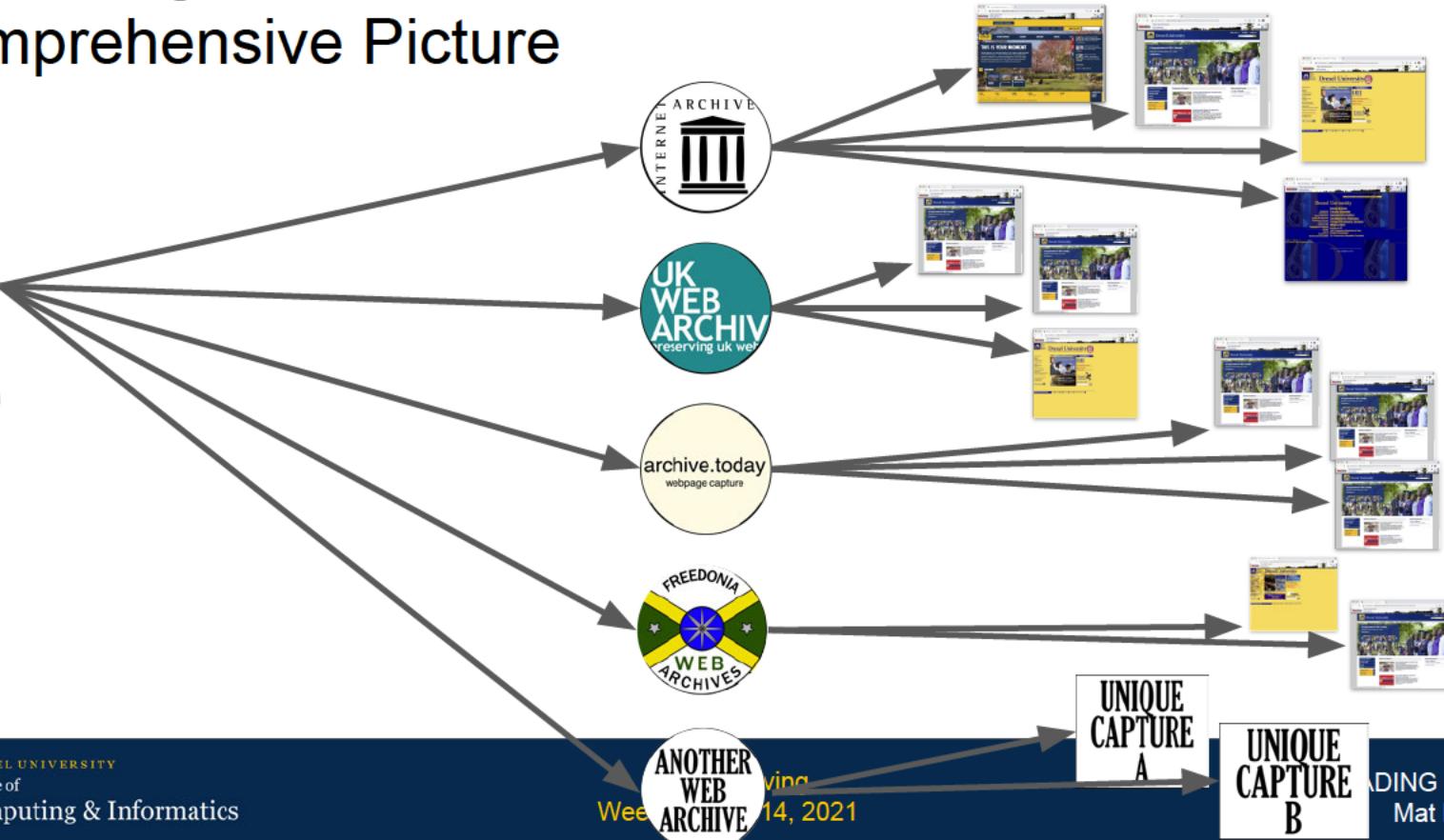
# Web Archives Provide Access to the Web that was



What did `drexel.edu` look like in the past?



# Consulting More Archives Produces a More Comprehensive Picture



# Even Then, Not Everything is Preserved

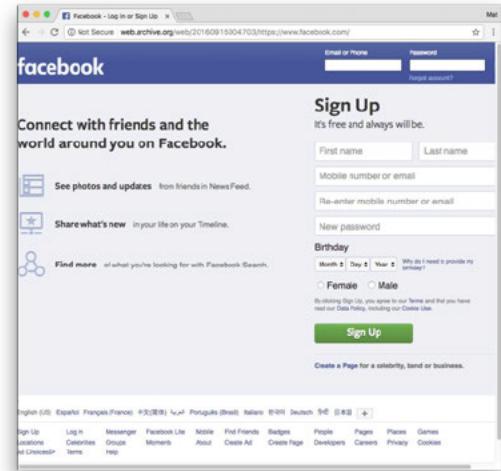
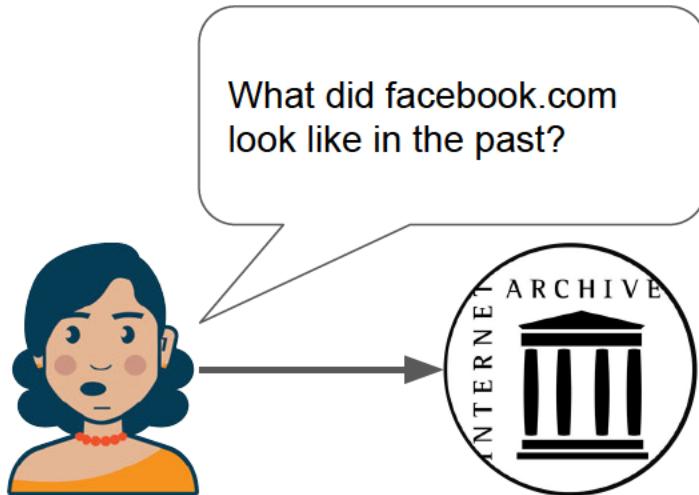
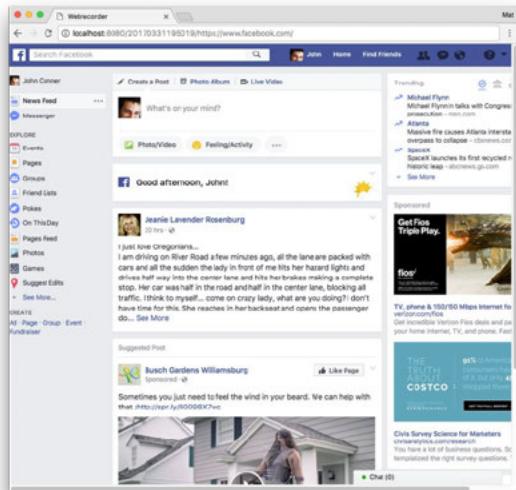


What did obscuresite.com look like in the past?

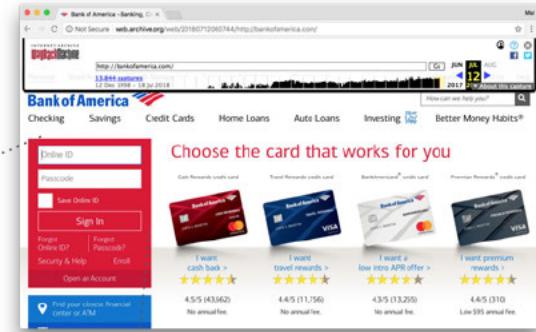
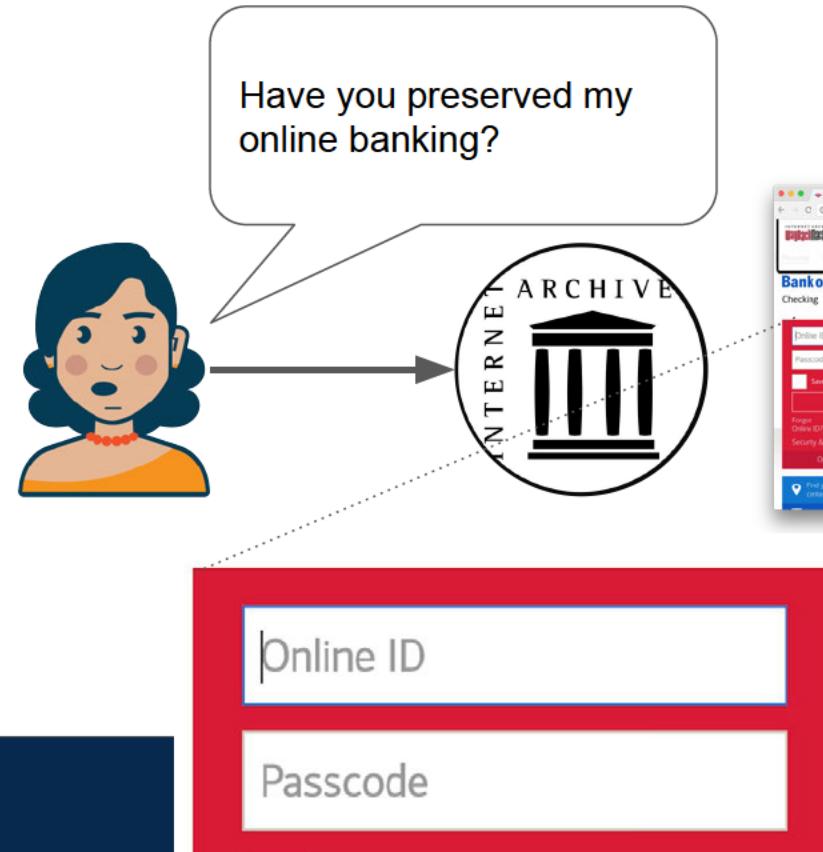
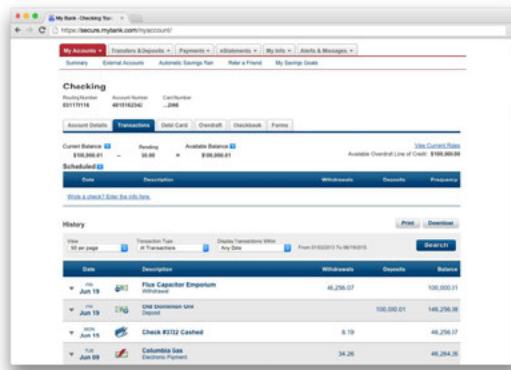


0 captures for  
obscuresite.com

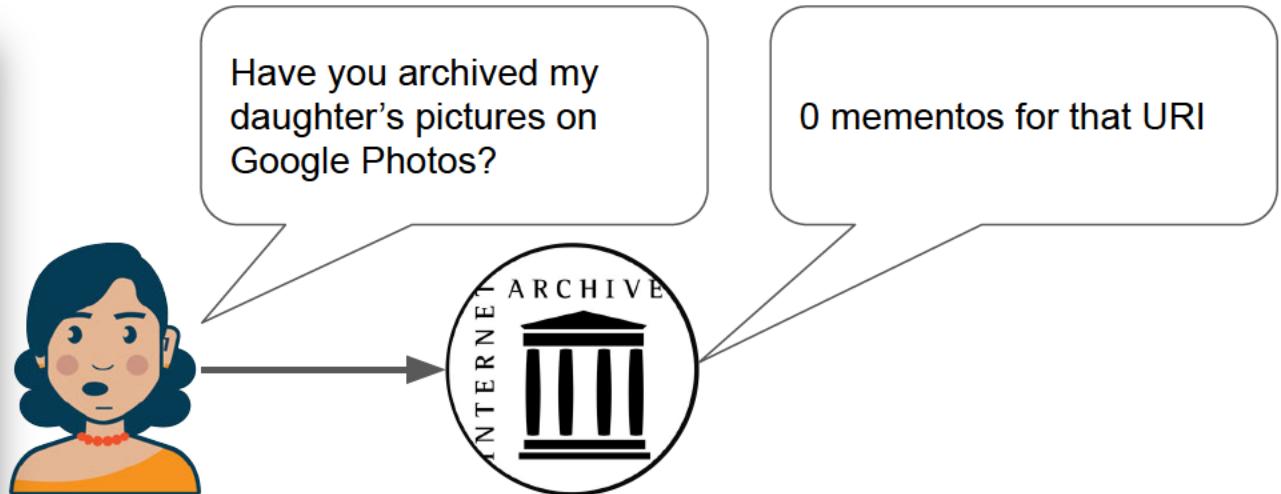
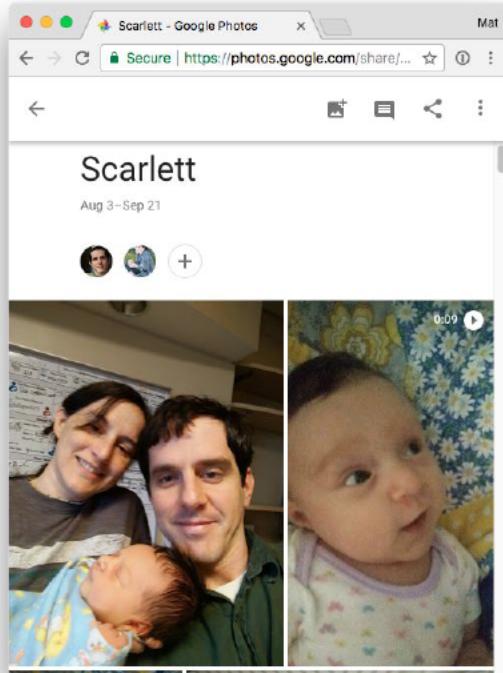
# User Sees on Live Web May Not Be What is Captured



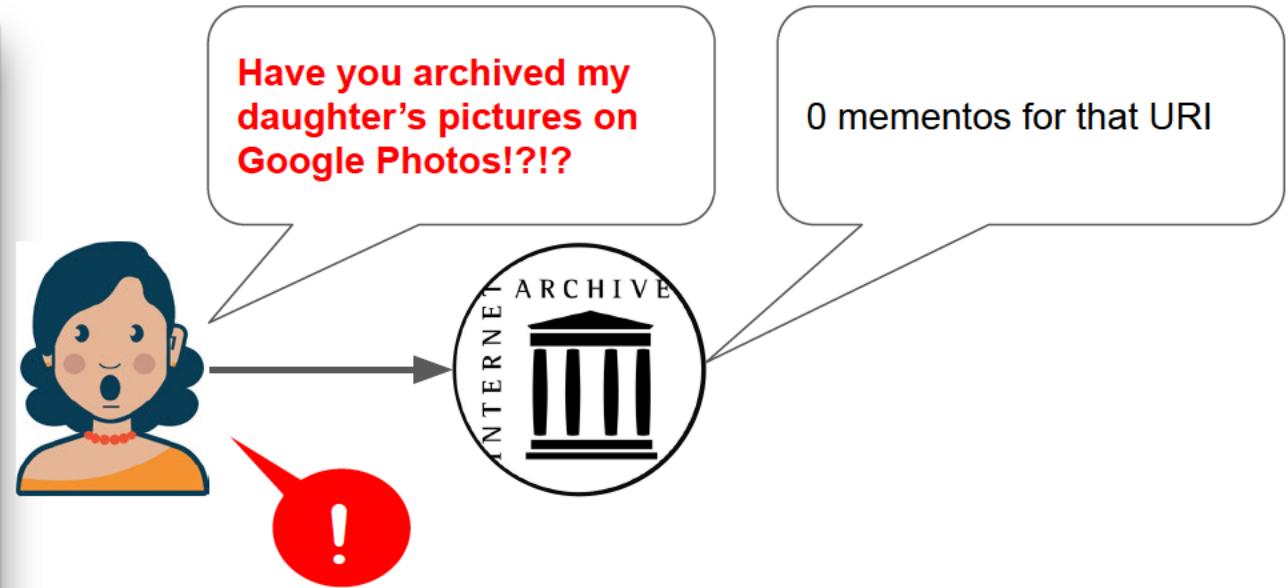
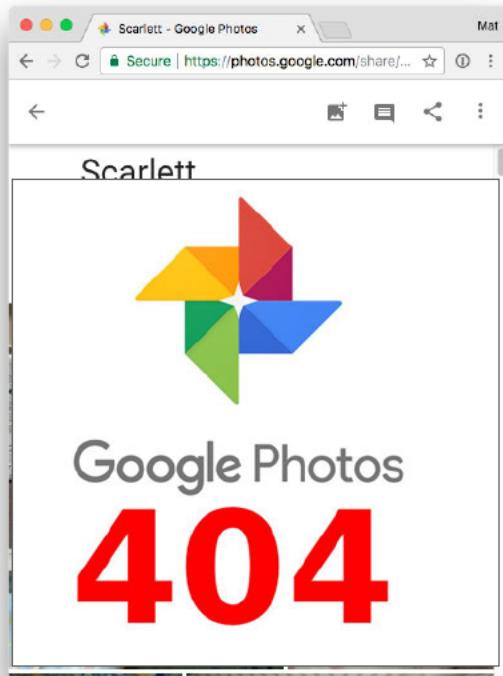
# ...And Oftentimes That is For the Best



# Other Times, We May Want Our Content Archived

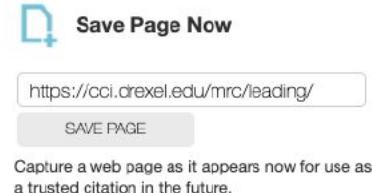


# ...Especially When It Has Disappeared



# Approaches Toward Archiving the Web

- Crawling
- Page-at-a-time Web Archiving
- Proxy-style Web archiving



# Crawlers, in General

- Originally intended for search engines
  - Googlebot
  - Alexa
- Follow links, add to queue until exhausted (never)
- Custom scripts (Python, Perl, etc.) for ad hoc applications
  - Intent must be known prior to crawling
  - Can be temporally and spatially expensive

# Archival Crawler

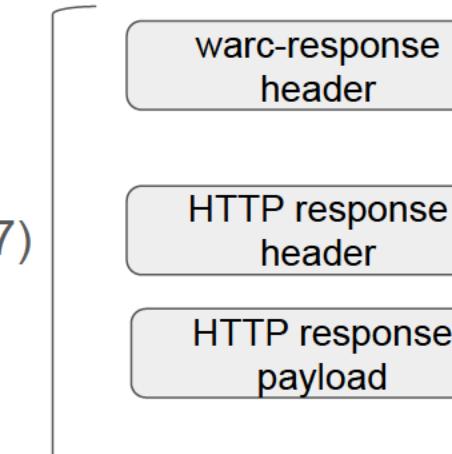
- Record HTTP transactions
- Associate metadata with crawls
- Links extracted from web page are added to queue (frontier) and subsequently processed
- Store in a portable, standard format
  - e.g., WARC

**HERITRIX**

# Crawler Output: Web Archive (WARC) file

- As compared to...
  - raw collection of files
  - screenshots
- ISO-backed (28500:2017)
- Checksums/Fixity
- Metadata
- Extensible

WARC response record



```
20160905022013693.warc • UNREGISTERED  
20160905022013693.warc  
54 WARC/1.0  
55 WARC-Type: response  
56 WARC-Target-URI: http://ipwb.example.com/  
57 WARC-Date: 2016-09-05T02:20:13Z  
58 WARC-Record-ID: <urn:uuid:06d837b9-5747-3f9e-a7b1-5431274b8aaa>  
59 Content-Type: application/http; msgtype=response  
60 Content-Length: 806  
61  
62 HTTP/1.1 200 OK  
63 Host: ipwb.example.com  
64 Connection: close  
65 Content-Type: text/html; charset=UTF-8  
66 Content-Length: 684  
67  
68 <html><head>  
69 <title>InterPlanetary Wayback</title>  
70 <link rel="stylesheet" type="text/css" href="style.css">  
71 </head>  
72 <body>  
73   
74 <p>InterPlanetary Wayback (ipwb) facilitates permanence and coll  
75   
76   
77  
78 </body></html>  
79  
80  
81 WARC/1.0  
82 WARC-Type: response  
83 WARC-Target-URI: http://ipwb.example.com/style.css  
84 WARC-Date: 2016-09-05T02:20:13Z  
85 WARC-Record-ID: <urn:uuid:b9f7761e-e6b4-d4c7-317b-49894413e6a5>  
86 Content-Type: application/http; msgtype=response
```

Line 72, Column 7      Tab Size: 4      Plain Text

# Access

- Files readable but “replay” required
  - Reassembles resource representations captured
  - Allow one to experience as if a web page
- Account for “rewriting” or “redirecting” links back to the archive
  - Does this compromise the integrity of the record?
  - Some advanced (reconstructive) approaches mitigate this

# Circling back: Crawling the Dynamic Web

- Crawlers should be fast, often written as scripts
- Typically are not functionally complete to modern web techs
  - Lag behind in some newer features of the web
  - This is problematic! Resources are missed
- Other approaches: leverage a “headless” browser
  - A lot slower but crawls are higher fidelity -- more accurate record

## More info:

Justin F. Brunelle, Mat Kelly, Michele C. Weigle and Michael L. Nelson, “The Impact of JavaScript on Archivability,” International Journal on Digital Libraries (IJDL), 17(2), pp. 95-117. January 2016.

# Access beyond Replay

- Captures may be distributed between archives
  - Different accounts of the same web page over time can be aggregated
- Aggregation assists in revolving temporal voids
  - Higher temporal granularity is proportional to a more complete record
  - Sources should be trusted/vetted, else the record is questionable
- An archive likely does not have the precise time requested
  - Datetime negotiation assists in resolving the closest



# Memento (RFC7089)

- Specification for **temporal negotiation** on the web
- Provides *syntax and semantics* for representing archival captures
- Concepts like TimeMaps and TimeGates allow for representation of captures' identifiers from a single or multiple sources
- Most web archives implement Memento
  - including **WayBack Machine**
- Provides an alternate means of access beyond simply replaying the archived web page.

<https://datatracker.ietf.org/doc/html/rfc7089>

# TimeMap

Live web URL

Original URI (URI-R)

Note "last" here, providing temporal context

Relative Relations

Same representation in other formats

Other TimeMaps (URI-Ts)

Access point for temporal negotiation

TimeGate (URI-G)

```
<https://drexel.edu>; rel="original",
<https://aggregator.matkelly.com/timemap/link/https://drexel.edu>;  
rel="self"; type="application/link-format",
<http://web.archive.bibalex.org:80/web/19970626040823/http://www.drexel.  
edu/>; rel="first memento"; datetime="Thu, 26 Jun 1997 04:08:23 GMT",
<https://web.archive.org/web/19970626040823/http://www.drexel.edu:80/>;  
rel="memento"; datetime="Thu, 26 Jun 1997 04:08:23 GMT",
<https://web.archive.org/web/20060615011116/http://www.drexel.edu:80/>;  
rel="memento"; datetime="Thu, 15 Jun 2006 01:11:16 GMT",
<https://wayback.archive-it.org/all/20210611040826/https://drexel.edu/>;  
rel="memento"; datetime="Fri, 11 Jun 2021 04:08:26 GMT",
<https://web.archive.org/web/20210614165405/http://drexel.edu/>;  
rel="last memento"; datetime="Mon, 14 Jun 2021 16:54:05 GMT",
<https://aggregator.matkelly.com/timemap/link/https://drexel.edu>;  
rel="timemap"; type="application/link-format",
<https://aggregator.matkelly.com/timemap/json/https://drexel.edu>;  
rel="timemap"; type="application/json",
<https://aggregator.matkelly.com/timemap/cdxj/https://drexel.edu>;  
rel="timemap"; type="application/cdxj+ors",
<https://aggregator.matkelly.com/timegate/https://drexel.edu>;  
rel="timegate"
```

Link (RFC 7089) TimeMap

# Memento is Extensible -- e.g., other TimeMap formats

```
<http://matkelly.com>; rel="original",
<http://localhost:1208/timemap/link/http://matkelly.com>;
rel="self"; type="application/link-format",
<http://web.archive.org/web/20060514123511/http://www.matkelly.co
m:80/>; rel="first memento"; datetime="Sun, 14 May 2006 12:35:11
GMT",
<http://web.archive.org/web/20060516213852/http://www.matkelly.co
m/>; rel="memento"; datetime="Tue, 16 May 2006 21:38:52 GMT",
...
<http://web.archive.org/web/20180128152125/http://matkelly.com>;
rel="memento"; datetime="Sun, 28 Jan 2018 15:21:25 GMT",
<http://web.archive.org/web/20180319141920/http://matkelly.com/>;
rel="last memento"; datetime="Mon, 19 Mar 2018 14:19:20 GMT",
<http://localhost:1208/timemap/link/http://matkelly.com>;
rel="timemap"; type="application/link-format",
<http://localhost:1208/timemap/json/http://matkelly.com>;
rel="timemap"; type="application/json",
<http://localhost:1208/timemap/cdxj/http://matkelly.com>;
rel="timemap"; type="application/cdxj+ors",
<http://localhost:1208/timegate/http://matkelly.com>;
rel="timegate"
```

## Link (RFC 7089) TimeMap

Original URI (URI-R)

Other TimeMaps (URI-Ts)

TimeGate (URI-G)

Relative Relations

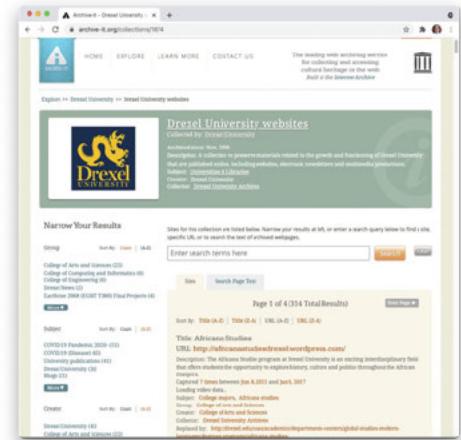
```
!context ["http://tools.ietf.org/html/rfc7089"]
!id {"uri": "http://localhost:1208/timemap/cdxj/http://matkelly.com"}
!keys ["memento_datetime_YYYYMMDDhhmmss"]
!meta {"original_uri": "http://matkelly.com"}
!meta {"timegate_uri": "http://localhost:1208/timegate/http://matkelly.com"}
!meta {"timemap_uri": {"link_format":
"http://localhost:1208/timemap/link/http://matkelly.com", "json_format":
"http://localhost:1208/timemap/json/http://matkelly.com", "cdxj_format":
"http://localhost:1208/timemap/cdxj/http://matkelly.com"}}
20060514123511 {"uri":
"http://web.archive.org/web/20060514123511/http://www.matkelly.com:80/",
"rel": "first memento", "datetime": "Sun, 14 May 2006 12:35:11 GMT"}
20060516213852 {"uri":
"http://web.archive.org/web/20060516213852/http://www.matkelly.com/",
"rel": "memento", "datetime": "Tue, 16 May 2006 21:38:52 GMT"}
...
20180128152125 {"uri":
"http://web.archive.org/web/20180128152125/http://matkelly.com", "rel":
"memento", "datetime": "Sun, 28 Jan 2018 15:21:25 GMT"}
20180319141920 {"uri":
"http://web.archive.org/web/20180319141920/http://matkelly.com/", "rel":
"last memento", "datetime": "Mon, 19 Mar 2018 14:19:20 GMT"}
```

## CDXJ TimeMap

# Archival Institutions and Web Services



- Internet Archive (IA): <https://web.archive.org>
    - Mostly automated crawling but now allows one-off user-supplied URI-specification
  - Archive-It: <https://archive-it.org>
    - Subsidiary of IA
    - \$-based subscription service
    - Introduces curated “collections”
    - Allows metadata association
  - Webrecorder: <https://webrecorder.io>
    - Newer technologies, higher-fidelity captures
    - Collection-based organization
    - Free and Open Source, actively developed
    - Web service (Conifer) and suite of tools



# Web Archiving Tools

- Capture
  - Heritrix (institutional grade crawler)
  - Proxy-style capture
    - Webrecorder's pywb
    - Warcprox
  - Browser-based capture: WARCreate
- Replay
  - OpenWayback
  - Webrecorder's pywb



A screenshot of the Heritrix 3.2.0 web interface. The top navigation bar shows the title 'Heritrix Engine 3.2.0' and a 'Not Secure' warning. Below the title, the word 'HERITRIX' is displayed in large, bold, black letters. The main content area is divided into sections: 'Engine' (Memory: 29146 KiB used; 125952 KiB current heap; 232960 KiB max heap, with a 'run garbage collector' button), 'Job Directories' (0 detected, with a 'rescan' button), and 'Add Job Directory' (a form field with placeholder text: 'Create new job directory with recommended starting configuration').

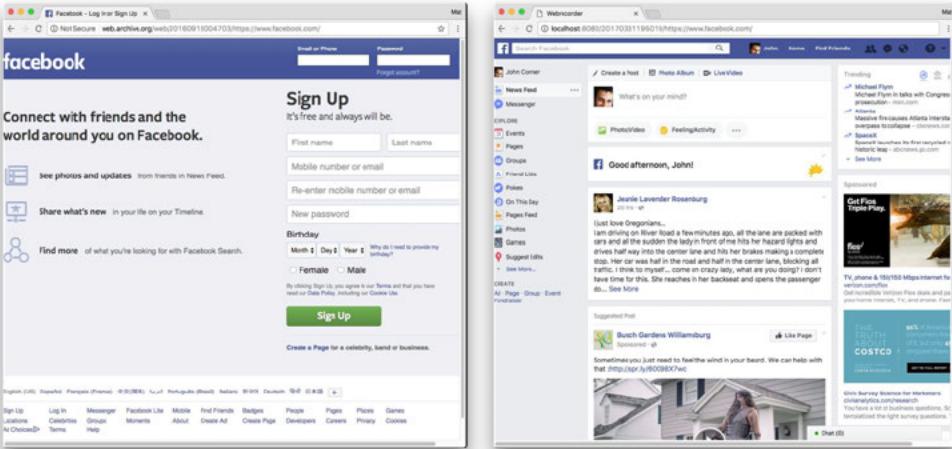
**Heritrix web interface**  
For control control - not definition

# Using Web Archives beyond simply Re-experiencing Web pages

- Historical research
- Technologies over time
- Recent discourse
- Fact-checking

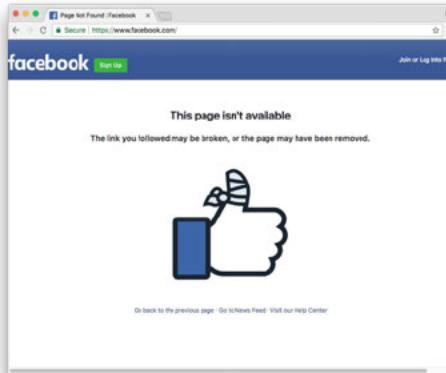
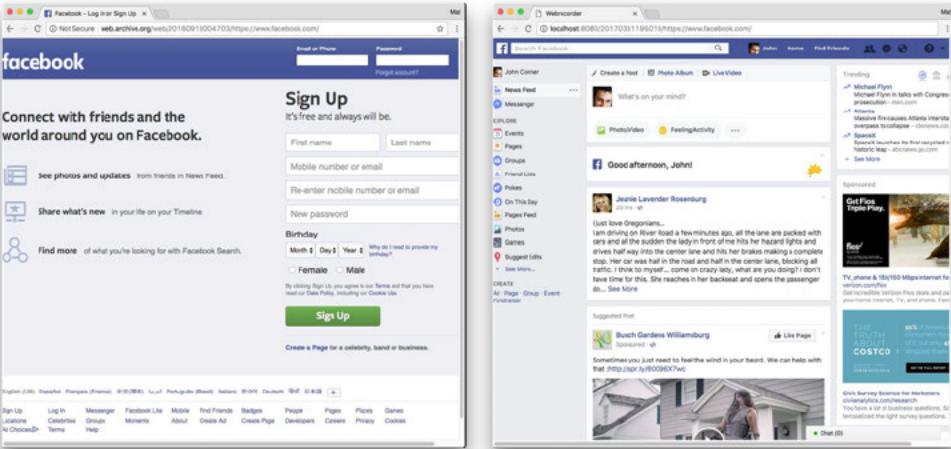
# Personal Web Archiving?

- Archives provide a representation of past web
  - Correct representation?
  - Verifiable?
- Personalization, authenticated representations allow for a custom experience on the web
  - Archivable?
  - Does it matter?



# Personal Web Archiving?

- Archives provide a representation of past web
  - Correct representation?
  - Verifiable?
- Personalization, authenticated representations allow for a custom experience on the web
  - Archivable?
  - Does it matter?



# Beyond Replay

- What if the computer with my WARCs dies?
  - InterPlanetary Wayback: peer-to-peer distributed integration of WARCs and IPFS
- While on the live web, how do I know how well this page has been preserved?
  - Mink: browser extension, view archival prevalence while browsing live web
- Archival tools are hard to manage and require CLI interaction, ugh!
  - WAIL: packages Heritrix, OpenWayback, other archiving tools into a native, desktop-based app



<https://github.com/oduwsdl/ipwb>



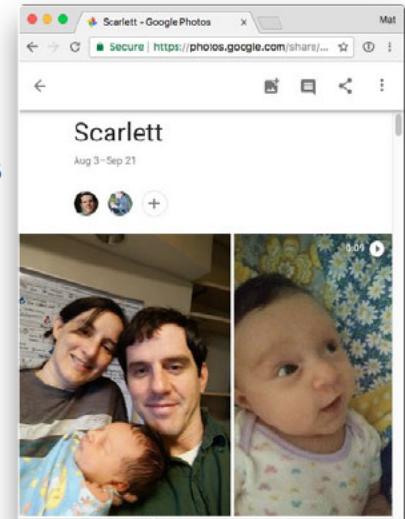
<https://github.com/machawk1/mink>



<https://github.com/machawk1/wail>

# Open Questions

- Capturing content behind authentication
  - Large part of the problem
  - What many think is important on the web requires credentials to access
  - PII
- Trusting institutions' captures as true
- Distributed web archives / aggregation
- Vetting captures over time
  - An approach: distributed archival fixity



# Archiving the Web is not a Solved Problem

- Web pages are culturally significant, should be preserved
- Contemporary archival tooling is always catching up to browsers' functionality
- Some web content should only reside in personal/private captures
- There is lots more to do regarding tooling, formats, verification, validating, etc.

