

Web Archiving

Mat Kelly, PhD

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

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

LIS Education And Data Science Integrated Network Group (LEADING) Bootcamp
Week 2: June 12, 2023

What We Will Cover

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

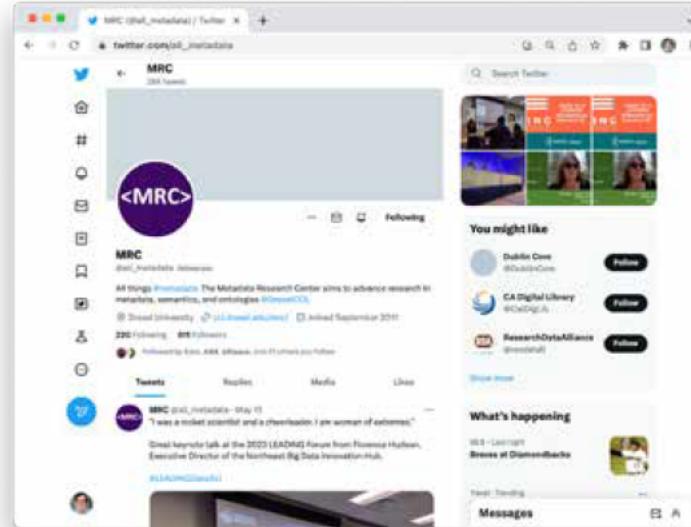
The Web



Our Web



The screenshot shows the homepage of the Metadata Research Center (MRC) at Drexel University. The header features the MRC logo and navigation links for HOME, ABOUT, RESEARCH, PUBLICATIONS & SCHOLARLY ACTIVITY, PEOPLE, and NEWS & EVENTS. Below the header, a yellow banner displays the text "LEADING: LIS Education And Data Science Integrated Network Group". Two blue buttons are present: "LEADING Home | People" and "LEADING Fellows | Application". A note indicates that the application deadline for 2021 is closed. At the bottom, there is information about the LEADING project, including its status as a Laura Bush 21st Century Librarian/SLIS National Digital Infrastructure and Initiatives project, supported by the Institute of Museum and Library Services (IMLS). The LEADING logo is also shown.



The screenshot shows the Twitter profile page for the MRC (@mrc_metadata). The profile picture is a purple circle containing the text "<MRC>". The bio reads: "All things [Promises](#). The Metadata Research Center aims to advance research in metadata, semantics, and ontologies ([M2O](#)).". It lists 220 followers and 416 following. A tweet from May 13, 2023, by @mrc_metadata states: "I was a panelist at a cheerleader. I am women of extremes." Below the tweet, it says "Great keynote talk at the 2022 LEADING Focus from Florence Huygen, Executive Director of the Northeast Big Data Innovation Hub." The interface includes sections for Tweets, Replies, Media, Likes, and a sidebar for "You might like" and "What's happening".

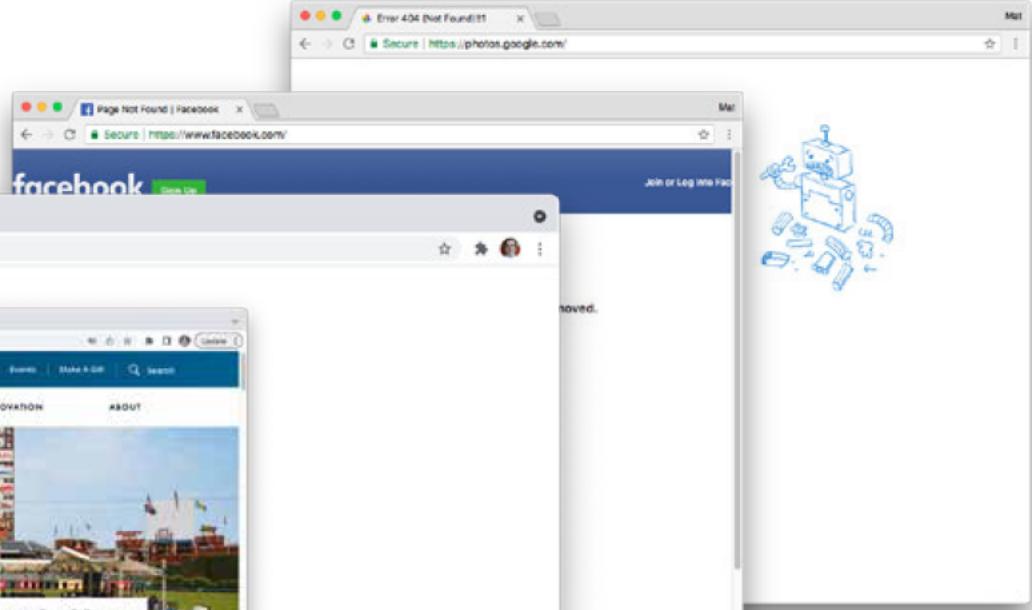
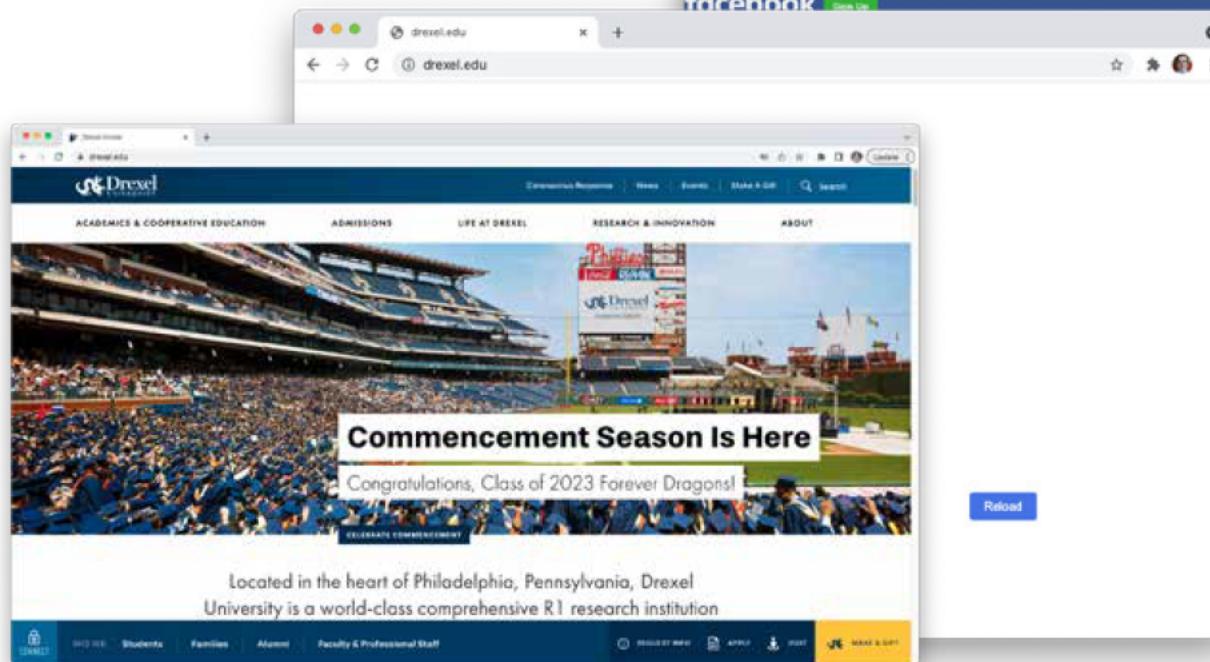
My Web

The screenshot shows the MRC website. At the top, there's a yellow header bar with the MRC logo and the text "LEADING: LIS Education And Data Science Integrated Network Group". Below this, there are two main buttons: "LEADING Home | People" and "LEADING Fellows | Application". A note below the buttons states: "Application deadline for 2021 is closed**". The bottom section contains the LEADING logo and some descriptive text about the project.

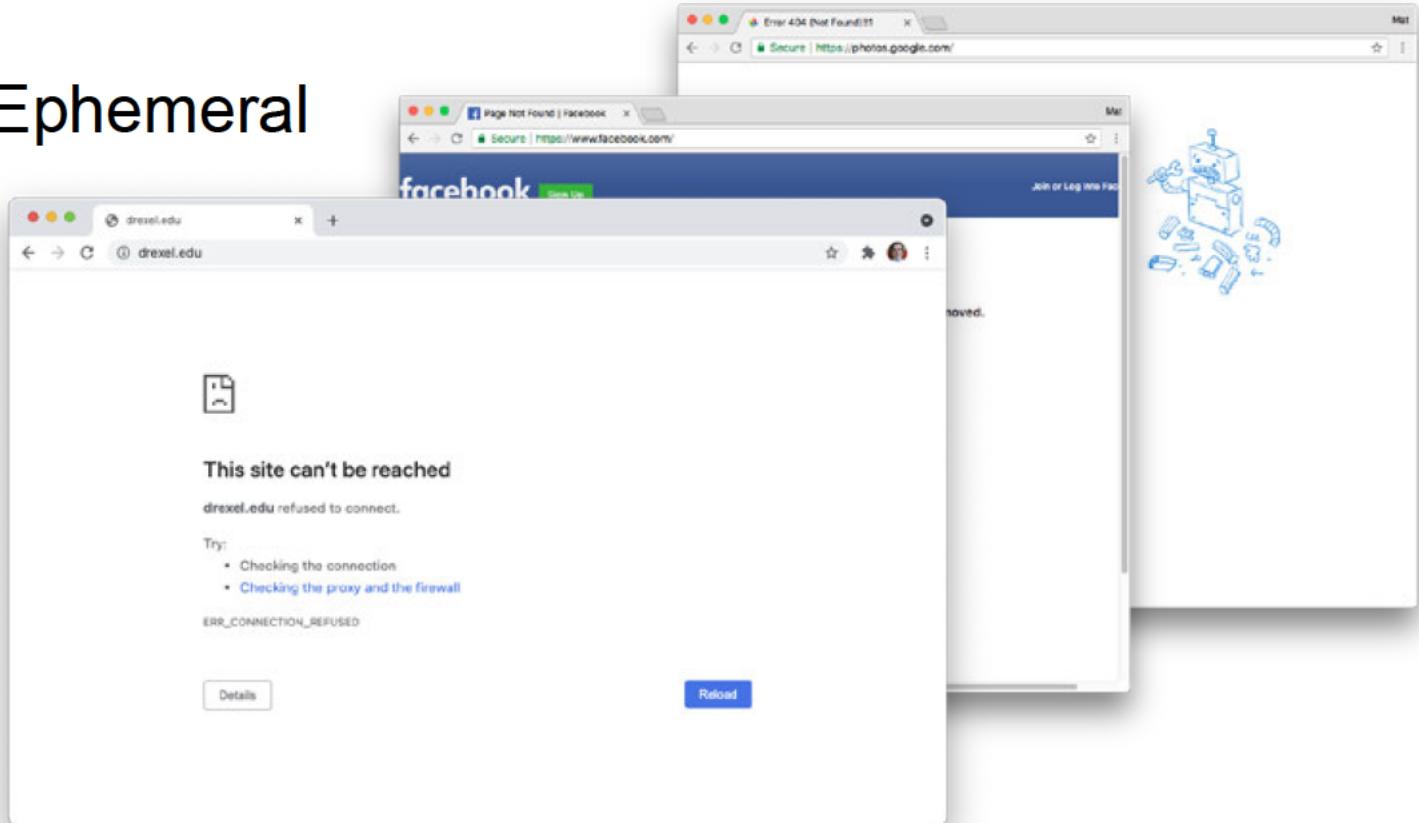
The screenshot shows the Twitter profile of the MRC (@mrc_metadata). The profile picture is a purple circle with the letters "MRC". The bio reads: "All things [#informatics](#). The Metadata Research Center aims to advance research in metadata, semantics, and ontologies [@mrc_metadata](#)". The account has 220 followers and 416 following. A tweet from May 13, 2023, is visible: "MRC grad, metadata - May 13 'I was a matak scientist and a cheerleader. I am women of extremes' Great keynote talk at the 2023 LEADING focus from Florence Huppen, Executive Director of the Northeast Big Data Innovation Hub. [@fjhuppen](#)"

The screenshot shows a Google Photos album titled "Scarlett" from August 9 to September 21. It features several photos of a family, including a man, a woman, and a baby. The woman is smiling in one photo, and the baby is looking towards the camera in another.

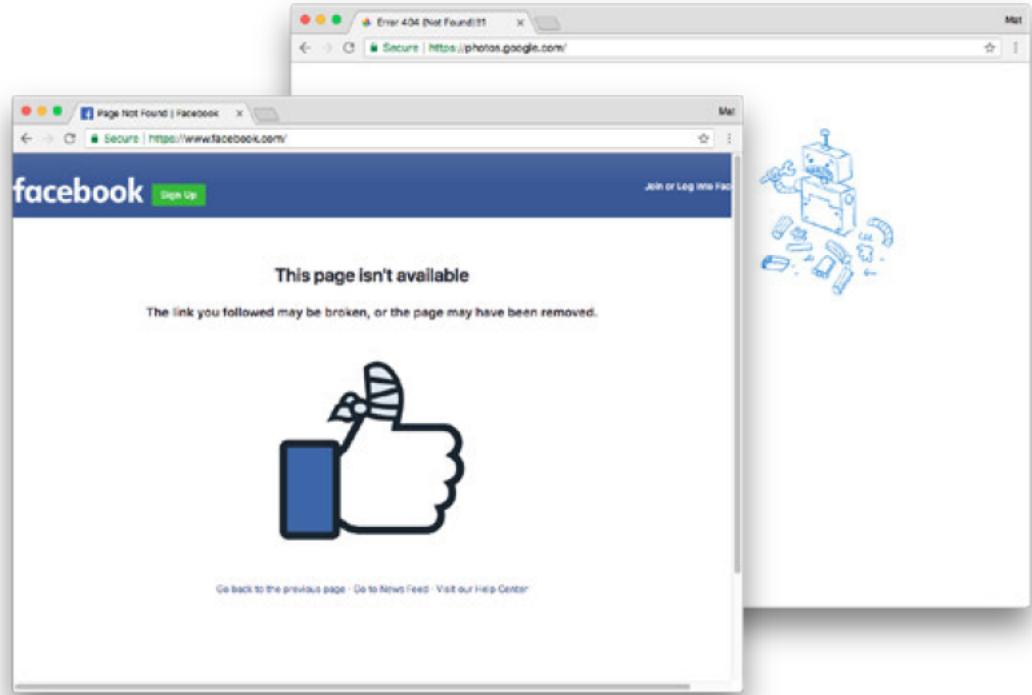
The Web is Ephemeral



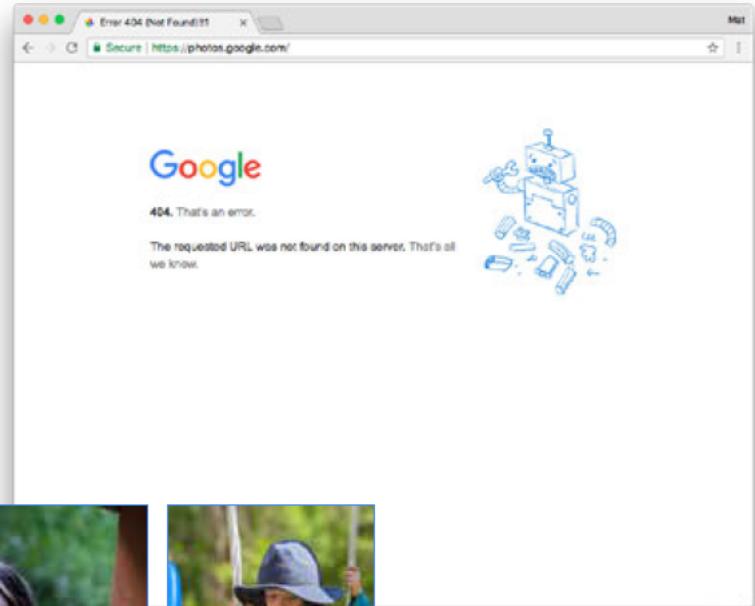
The Web is Ephemeral



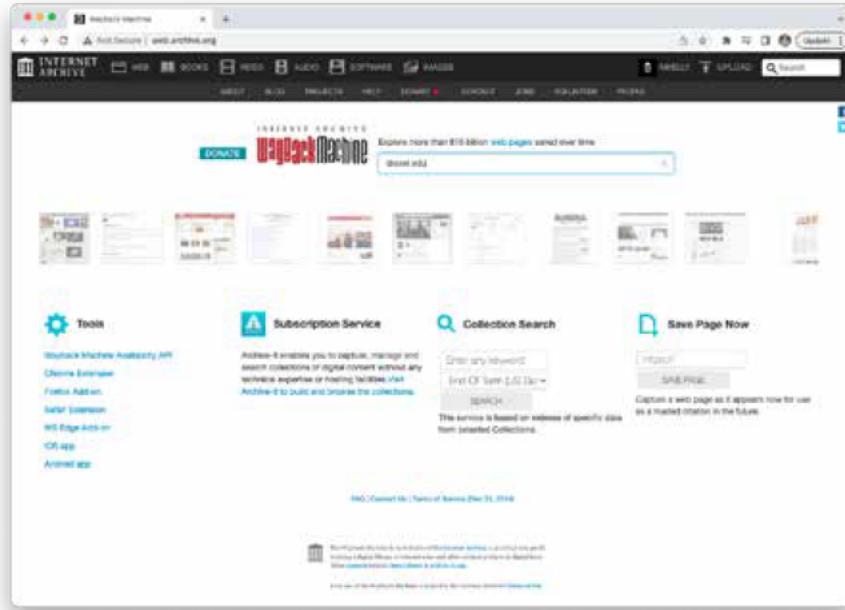
The Web is Ephemeral



The Web is Ephemeral

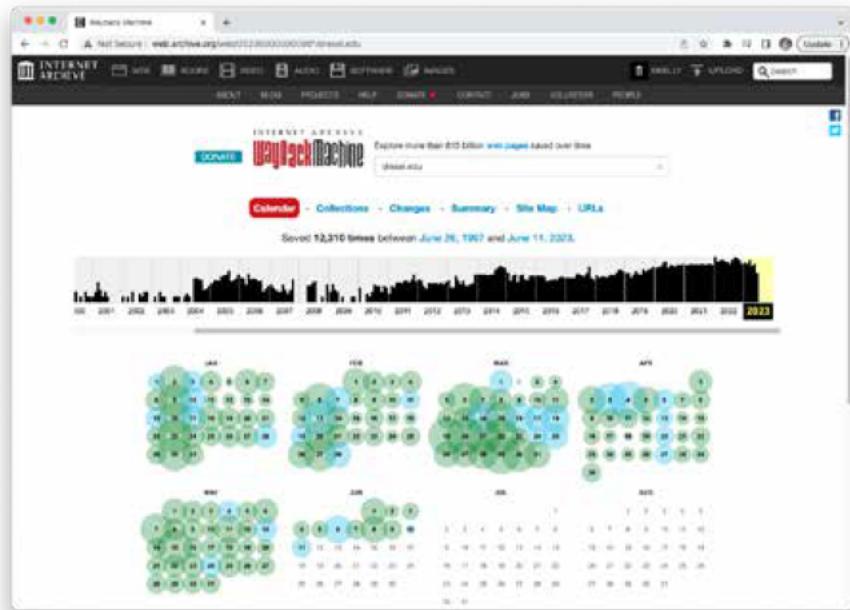


Web Archives to the Rescue: Typical Access



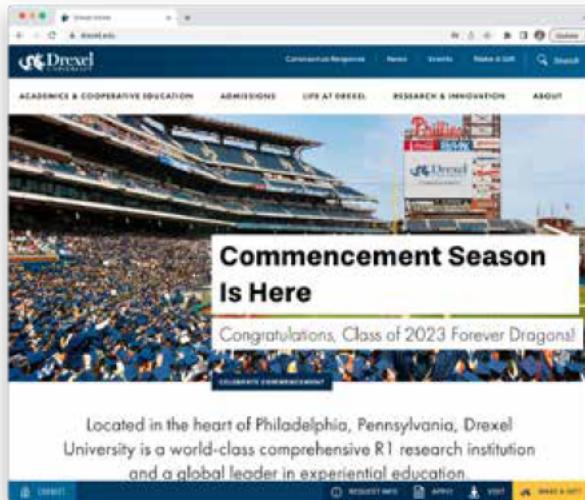
1. Go to `archive.org` in your browser
2. Enter the URL you want to see in the past in the form field
3. Submit your query

Web Archives to the Rescue: Typical Access



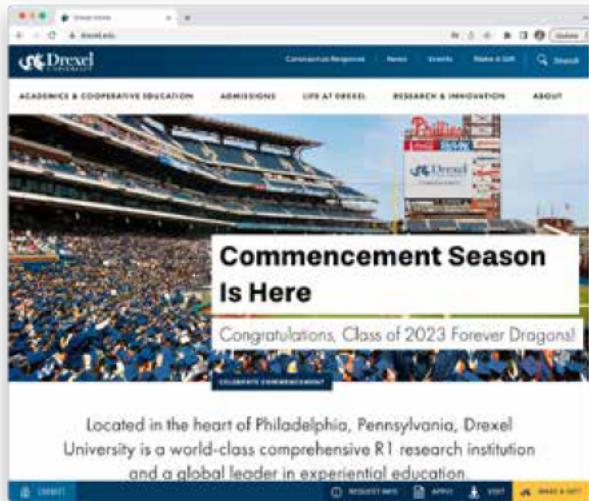
4. Locate the archived Web page on the calendar or histogram view
5. Select the year/selection for the day
6. Repeat until you find the closest date and time

Web Archiving: View The Web of the Past

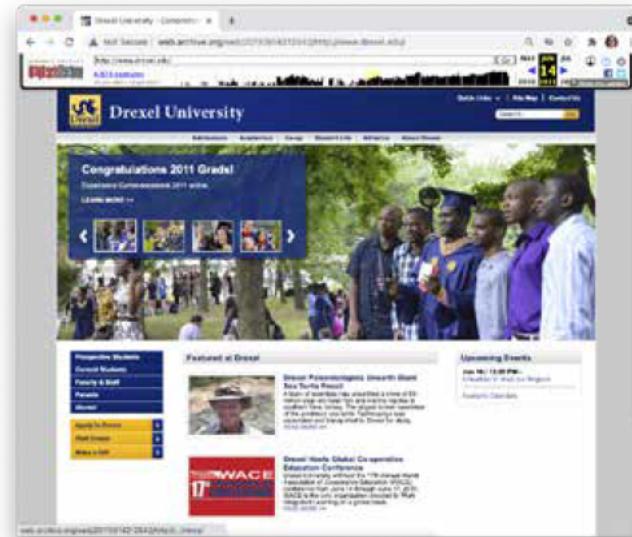


Now

Web Archiving: View The Web of the Past

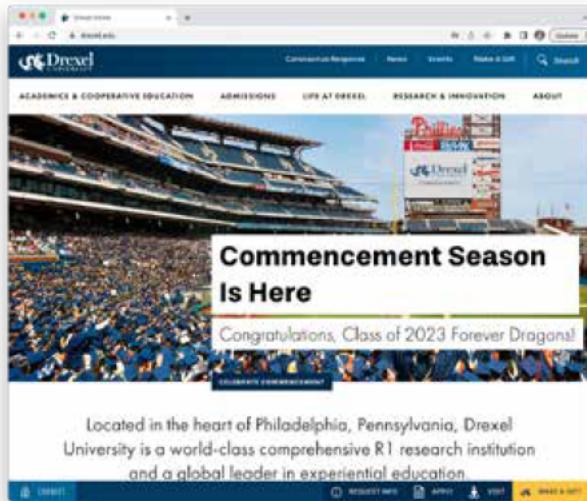


Now

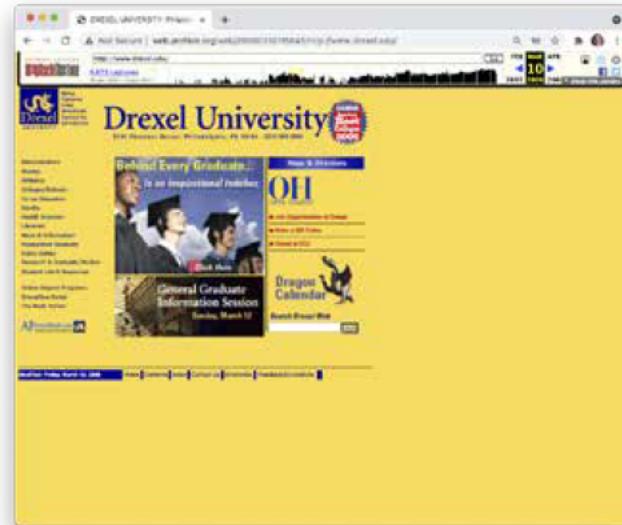


June 14, 2011

Web Archiving: View The Web of the Past

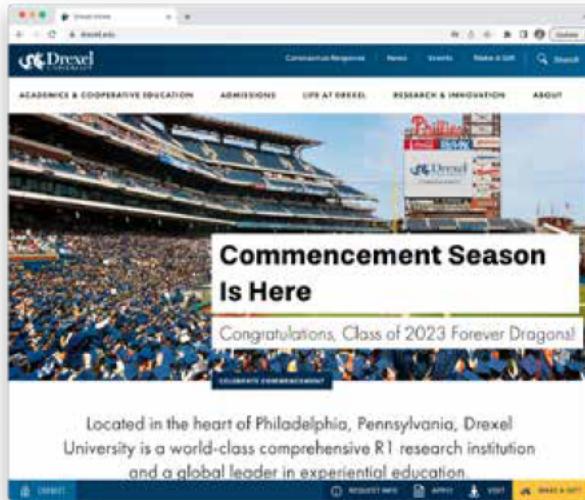


Now



March 10, 2006

Web Archiving: View The Web of the Past

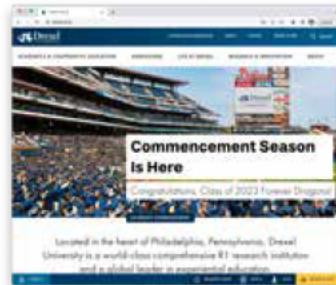


Now

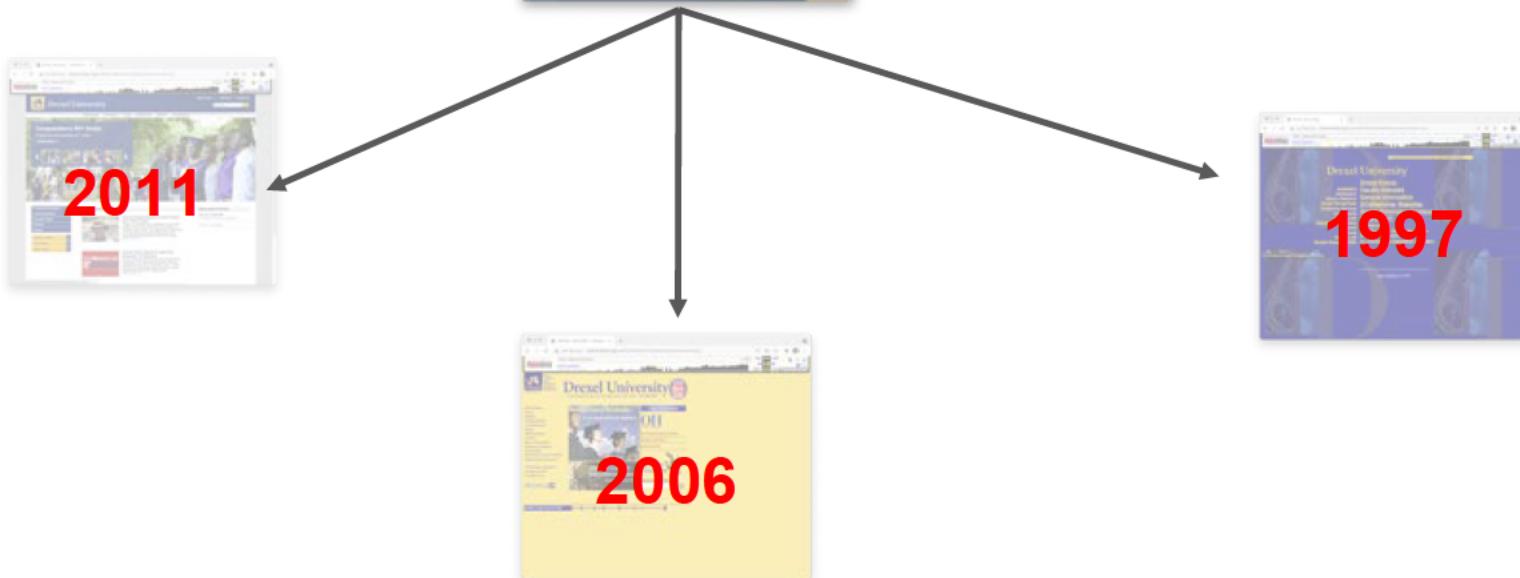


March 26, 1997

Web Archiving



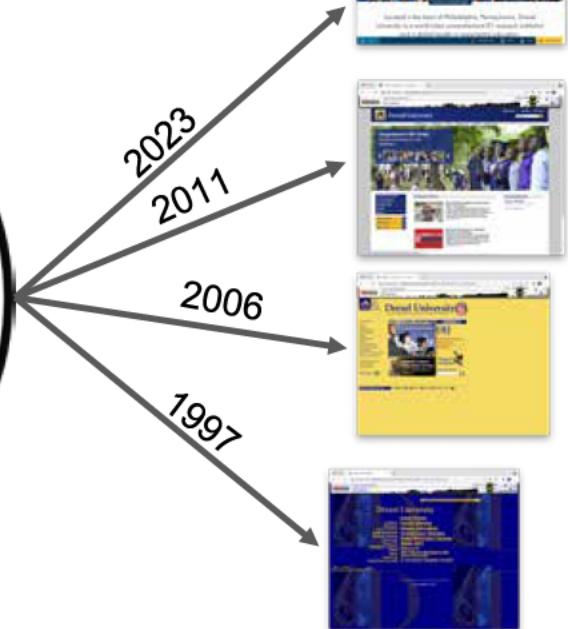
Associate a live Web URI
with their **archived representations**



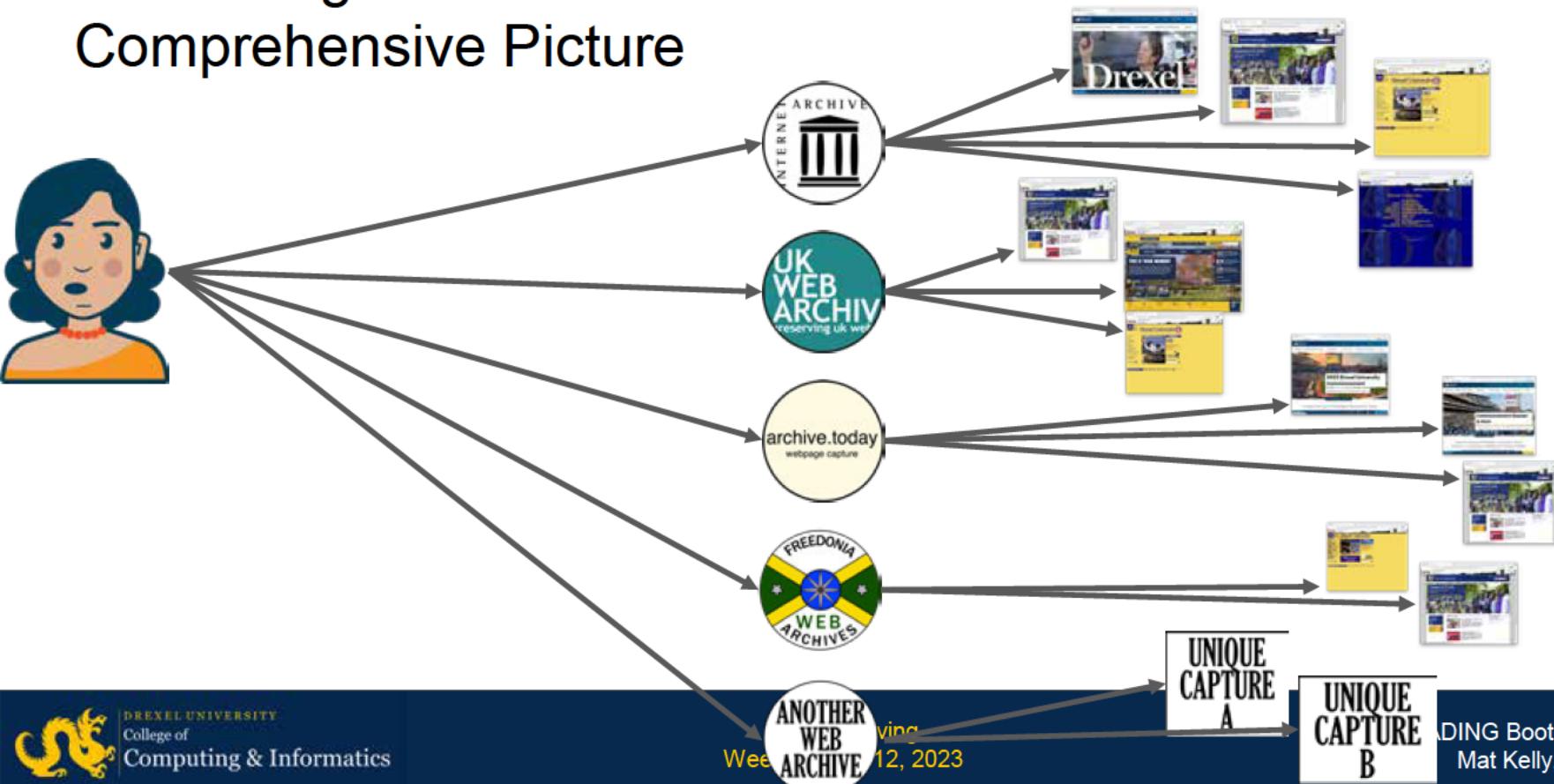
Web Archives Provide Access to the Web that was



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



Consulting More Archives Produces a More Comprehensive Picture



Even Then, Not Everything is Preserved

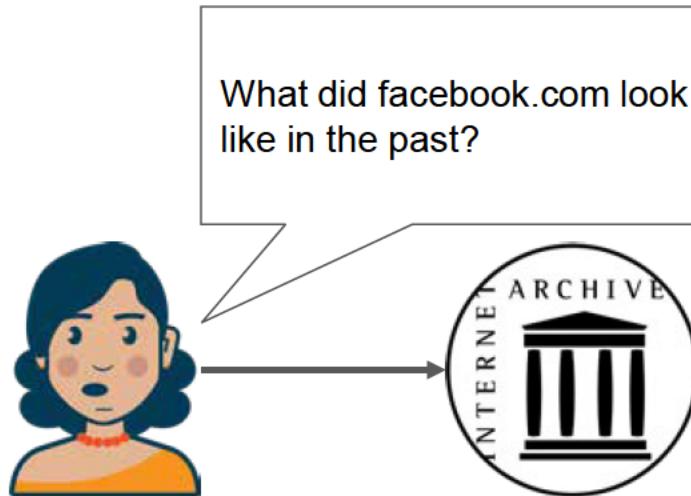
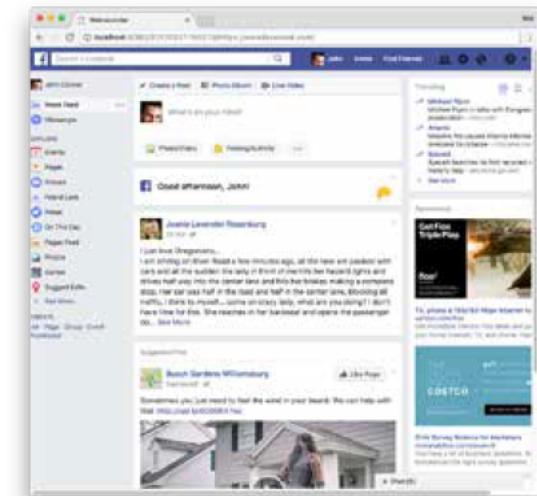


What did obscuresite.com look like in the past?

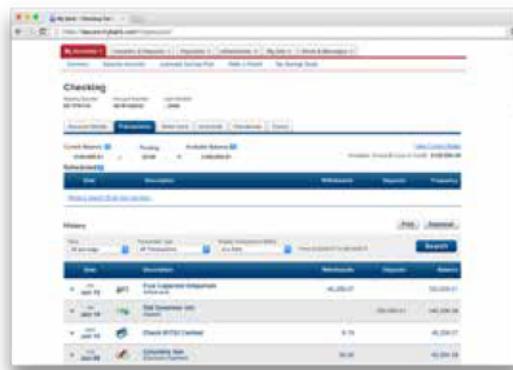


0 captures for
obscuresite.com

User Sees on Live Web May Not Be What is Captured



...And Oftentimes That is For the Best



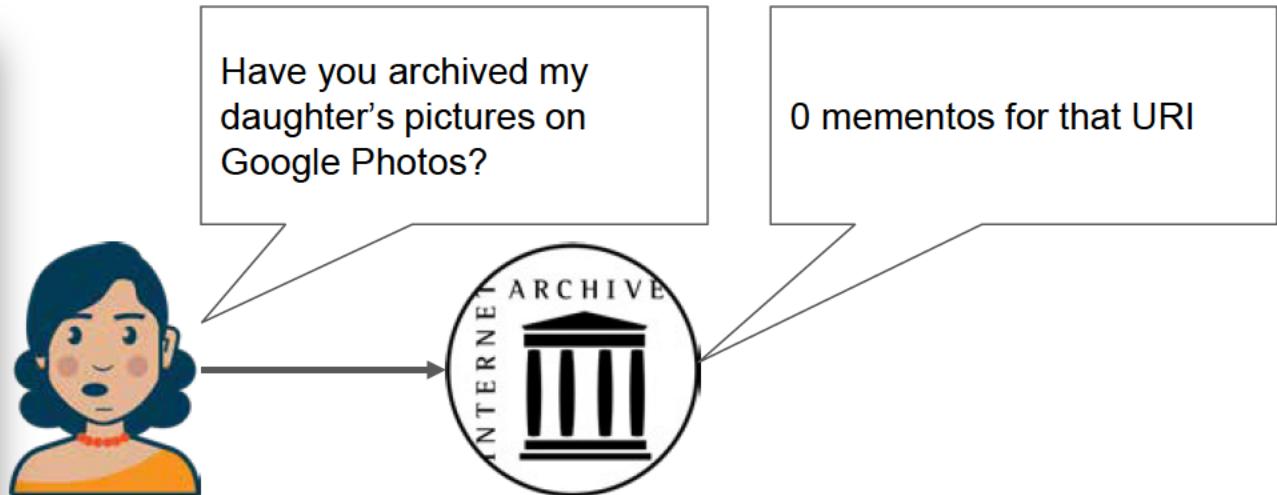
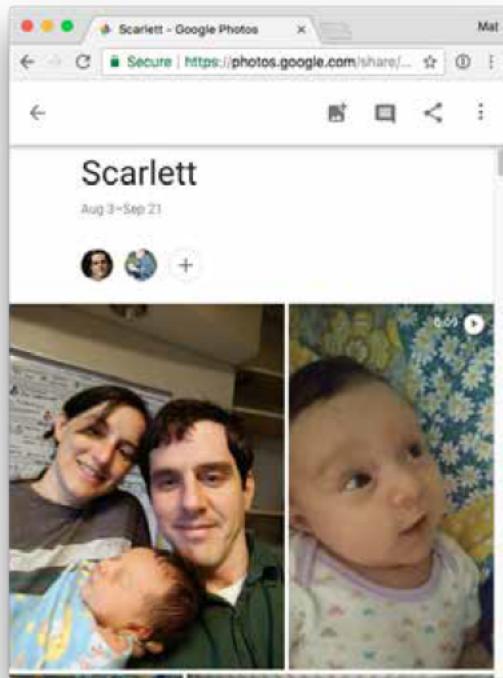
Have you preserved my
online banking?



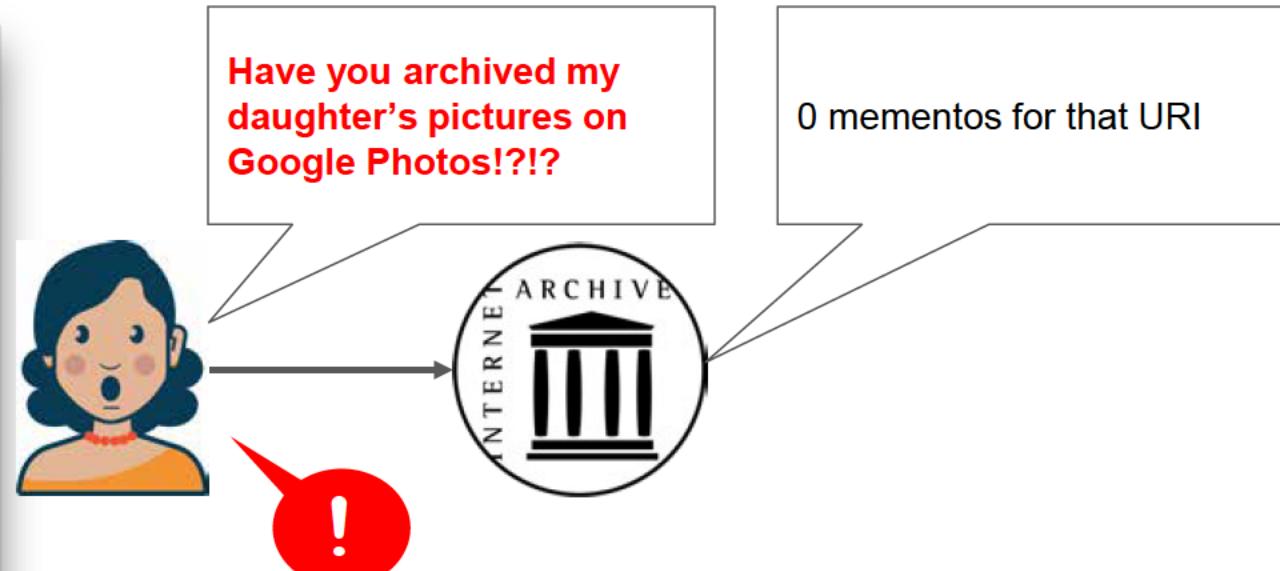
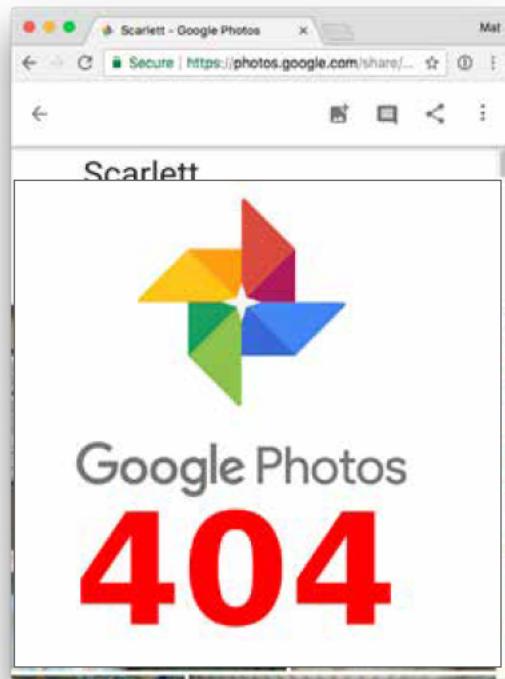
Online ID

Passcode

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



Save Page Now

<https://cci.drexel.edu/mrc/leading/>

SAVE PAGE

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

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

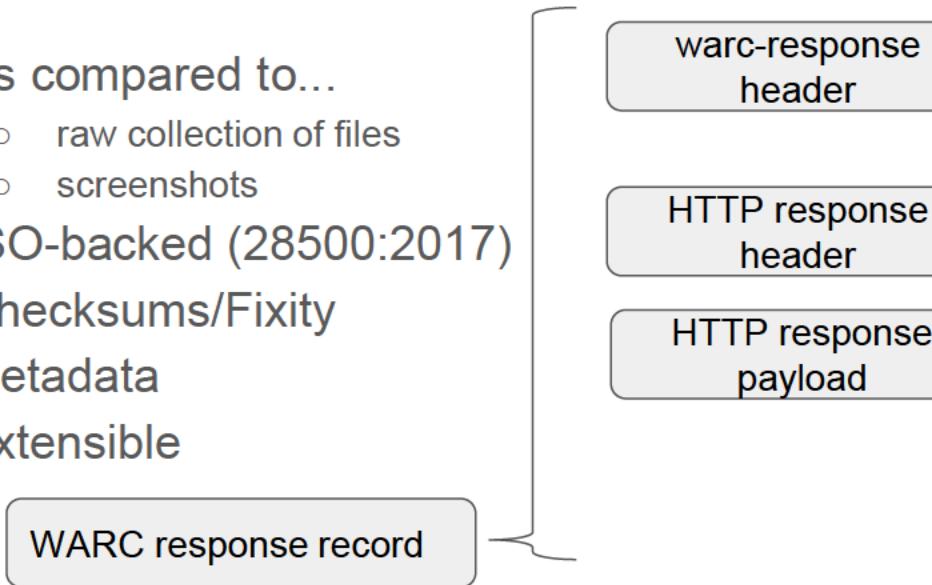
Archival Crawler

- 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



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:09f7761e-e6b4-d4c7-317b-49894413e6a5>
86 Content-Type: application/http; msgtype=response

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.com:8
0/>; rel="first memento"; datetime="Sun, 14 May 2006 12:35:11 GMT",
<http://web.archive.org/web/20060516213852/http://www.matkelly.com/">
; 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"
```

```
!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"}
```

Link (RFC 7089) TimeMap

Original URI (URI-R)

Other TimeMaps (URI-Ts)

TimeGate (URI-G)

Relative Relations

CDXJ TimeMap



College of
Computing & Informatics

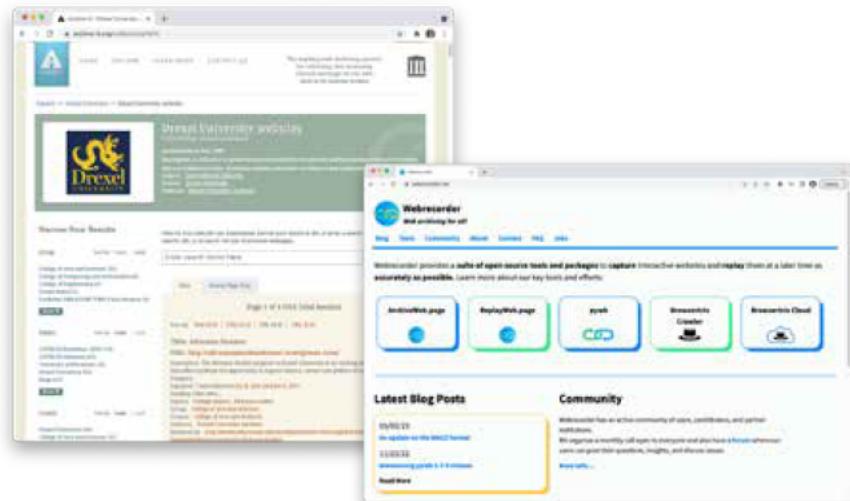
Web Archiving
Week 2: June 12, 2023

LEADING Bootcamp
Mat Kelly, PhD

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.net>
 - 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 a warning icon and the URL "Not Secure | 0.0.0.0:8443/engine". The main content area has a header "HERITRIX 3.2.0".

- Engine**: Shows memory usage (29145 KB used, 125952 KB current heap, 232960 KB max heap) and a "garbage collector" button.
- Job Directories**: Shows a list with "(0) detected" and "rescan" buttons. A link "Create new job directory with recommended starting configuration" is present.
- Add Job Directory**: A form field for entering a job directory path.

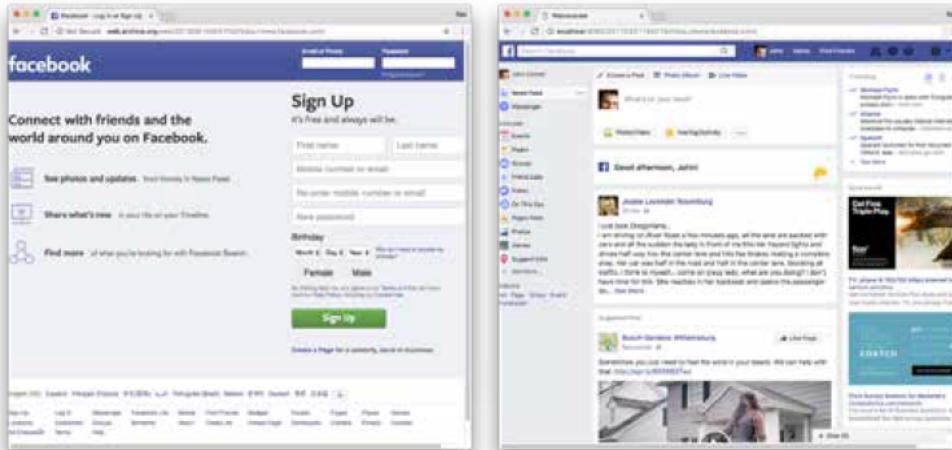
Heritrix web interface
For control control - not definition

Using Web Archives beyond simply Re-experiencing Web pages

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

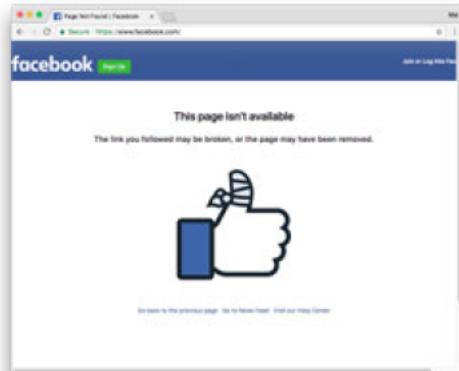
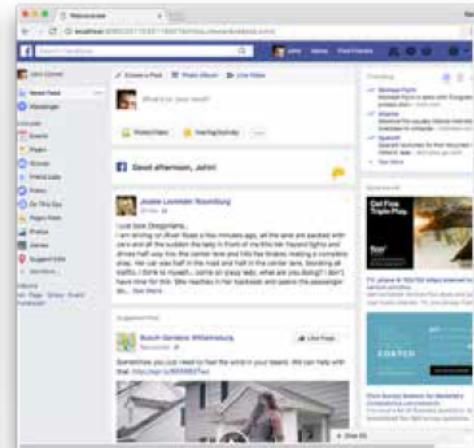
Personal Web Archiving?

- 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>



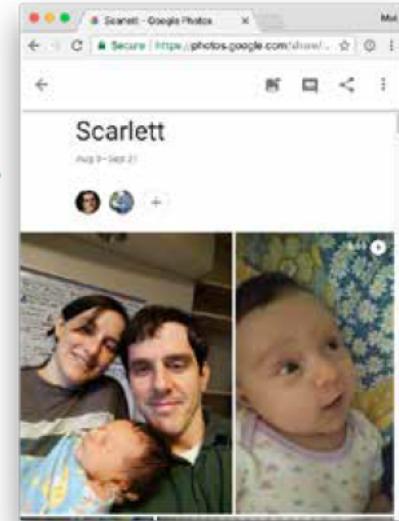
<https://github.com/machawk1/wail>

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
 - See:

Aturban M, Klein M, Van de Sompel H, Alam S, Nelson ML, et al. (2023) Hashes are not suitable to verify fixity of the public archived web. *PLOS ONE* 18(6): e0286879.
<https://doi.org/10.1371/journal.pone.0286879>

Published June 9, 2023



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.

