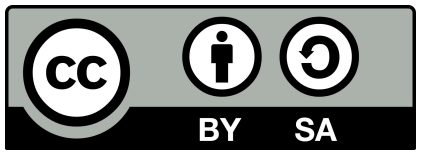
On Archival Negotiation Beyond Time

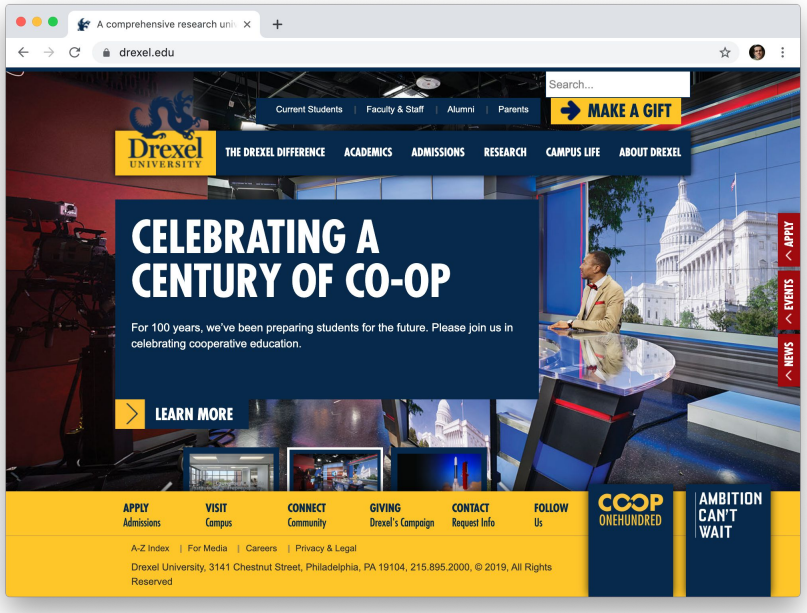
Dr. Mat Kelly
College of Computer and Informatics
Drexel University
mkelly@drexel.edu

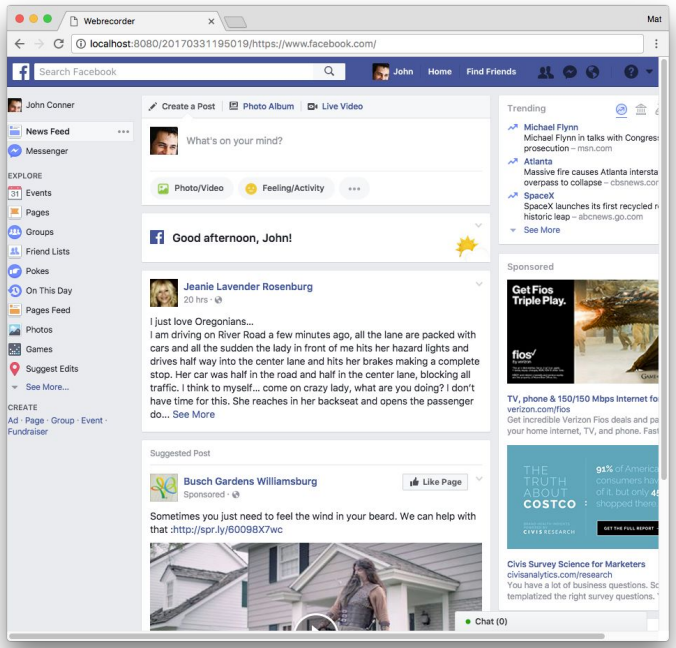
Guest lecture for
INFO 821: Foundations of Information Science
October 15, 2019

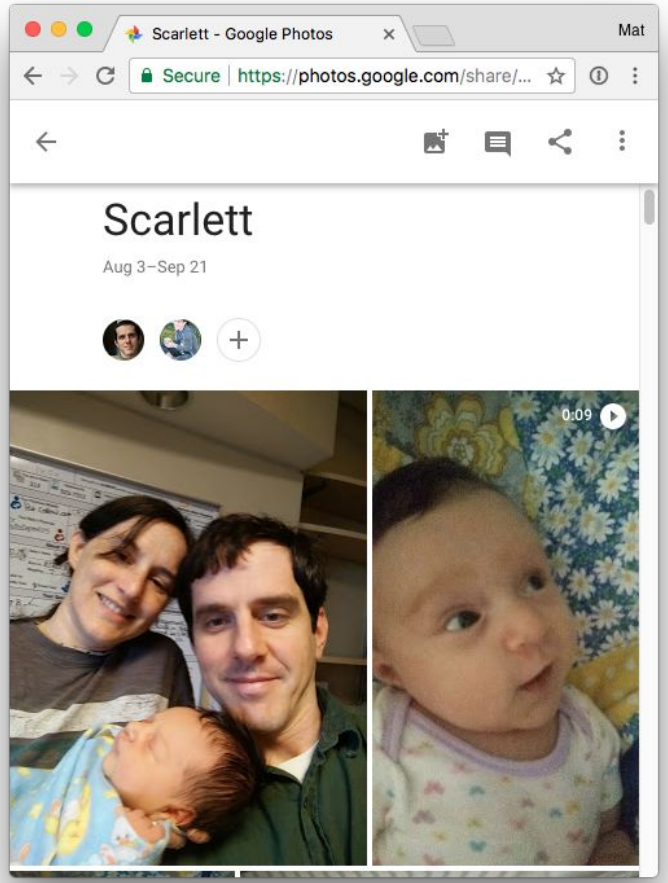
https://matkelly.com/2019info821



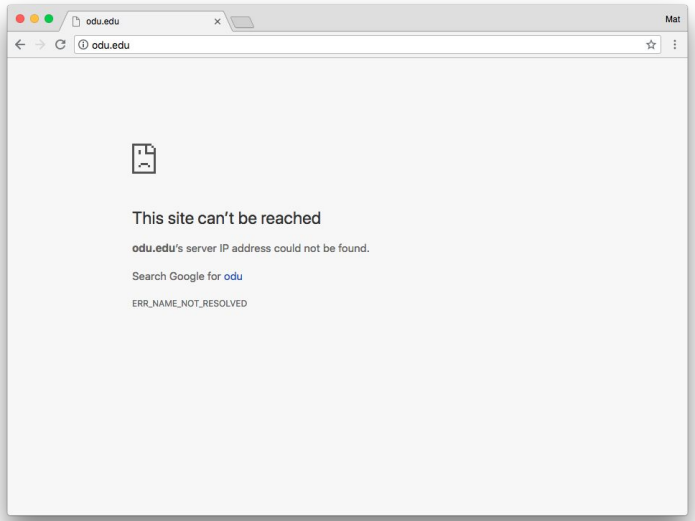
The (live) Web

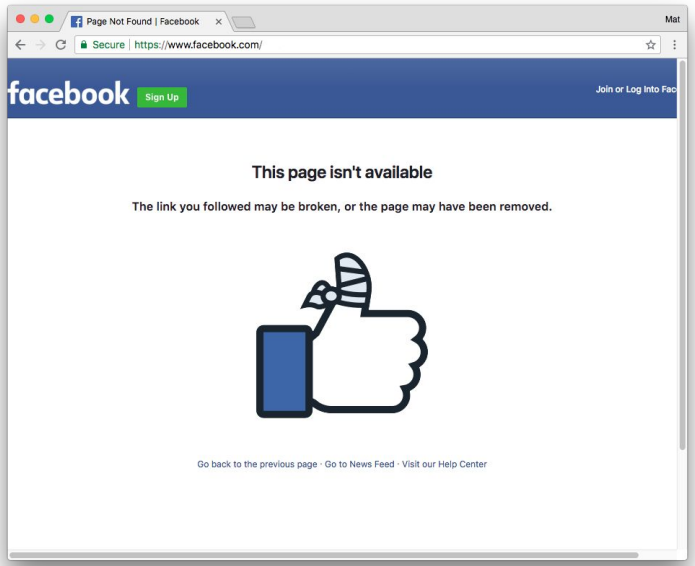

drexel.edu
Search...
Current Students | Faculty & Staff | Alumni | Parents
MAKE A GIFT
Drexel UNIVERSITY
THE DREXEL DIFFERENCE
ACADEMICS
ADMISSIONS
RESEARCH
CAMPUS LIFE
ABOUT DREXEL
CELEBRATING A CENTURY OF CO-OP
For 100 years, we've been preparing students for the future. Please join us in celebrating cooperative education.
LEARN MORE
APPLY
EVENTS
NEWS
APPLY Admissions
VISIT Campus
CONNECT Community
GIVING Drexel's Campaign
CONTACT Request Info
FOLLOW Us
COOP ONEHUNDRED
AMBITION CAN'T WAIT
A-Z Index | For Media | Careers | Privacy & Legal
Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, 215.895.2000, © 2019, All Rights Reserved

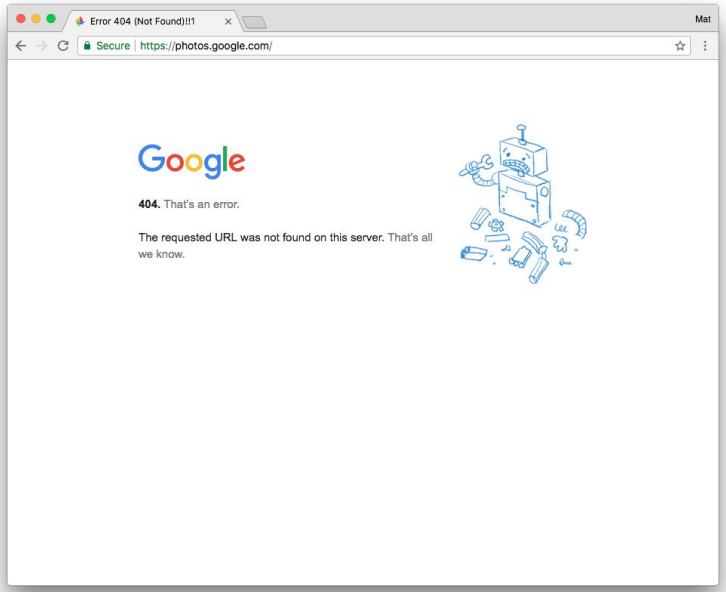

Webrecorder
localhost:8080/20170331195019/https://www.facebook.com/
Search Facebook
John Home Find Friends
John Conner
News Feed
Messenger
EXPLORE
Events
Pages
Groups
Friend Lists
Pokes
On This Day
Pages Feed
Photos
Games
Suggest Edits
See More...
CREATE
Ad · Page · Group · Event · Fundraiser
Create a Post | Photo Album | Live Video
What's on your mind?
Photo/Video
Feeling/Activity
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
Like Page
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 Congress prosecution – msn.com
Atlanta
Massive fire causes Atlanta interstate overpass to collapse – cbsnews.com
SpaceX
SpaceX launches its first recycled historic leap – abcnews.go.com
See More
Sponsored
Get Fios Triple Play.
TV, phone & 150/150 Mbps Internet
verizon.com/fios
THE TRUTH ABOUT COSTCO
Civis Survey Science for Marketers
civisanalytics.com/research
Chat (0)


Scarlett - Google Photos
Secure | https://photos.google.com/share/...
Scarlett
Aug 3–Sep 21
0:09

Web is Ephemeral

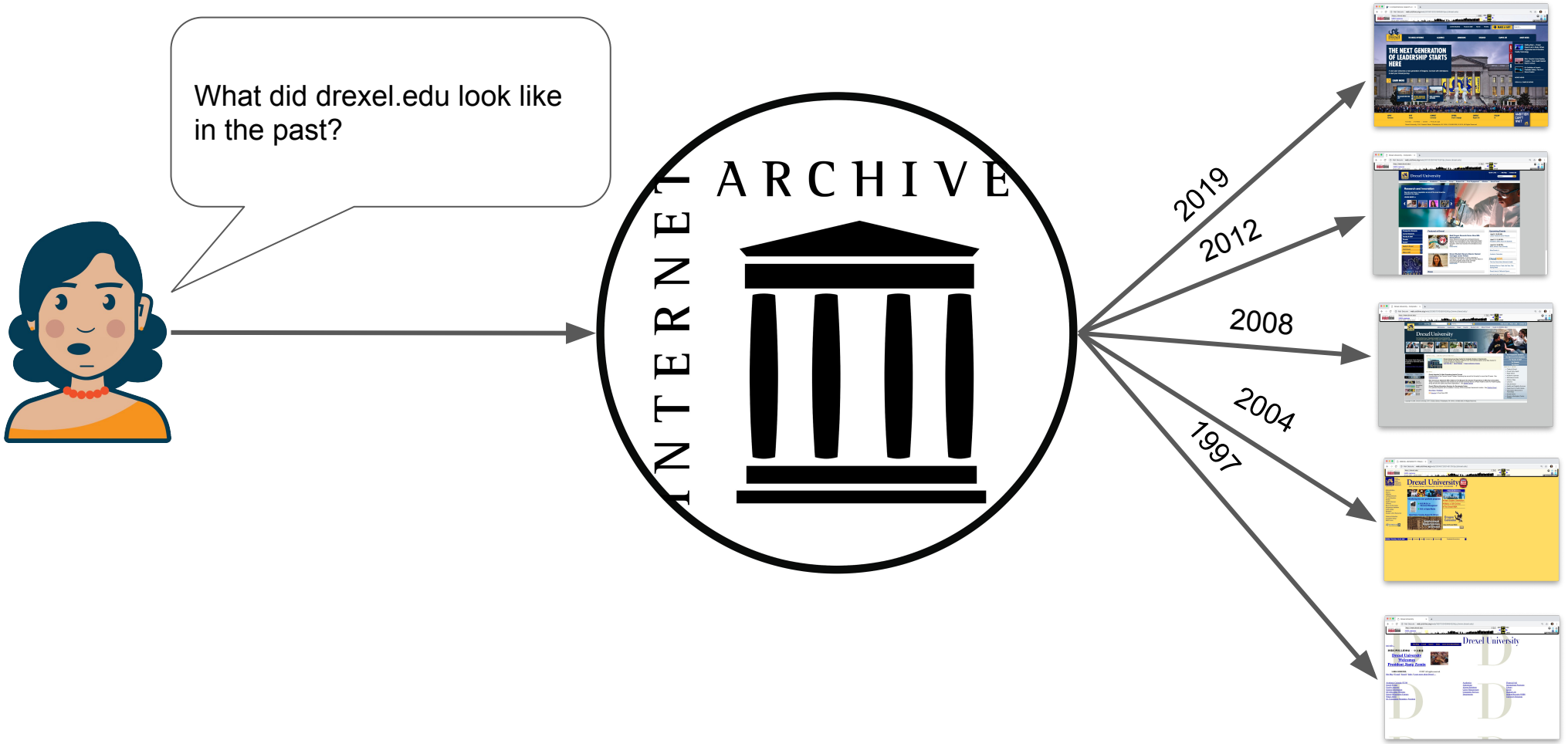






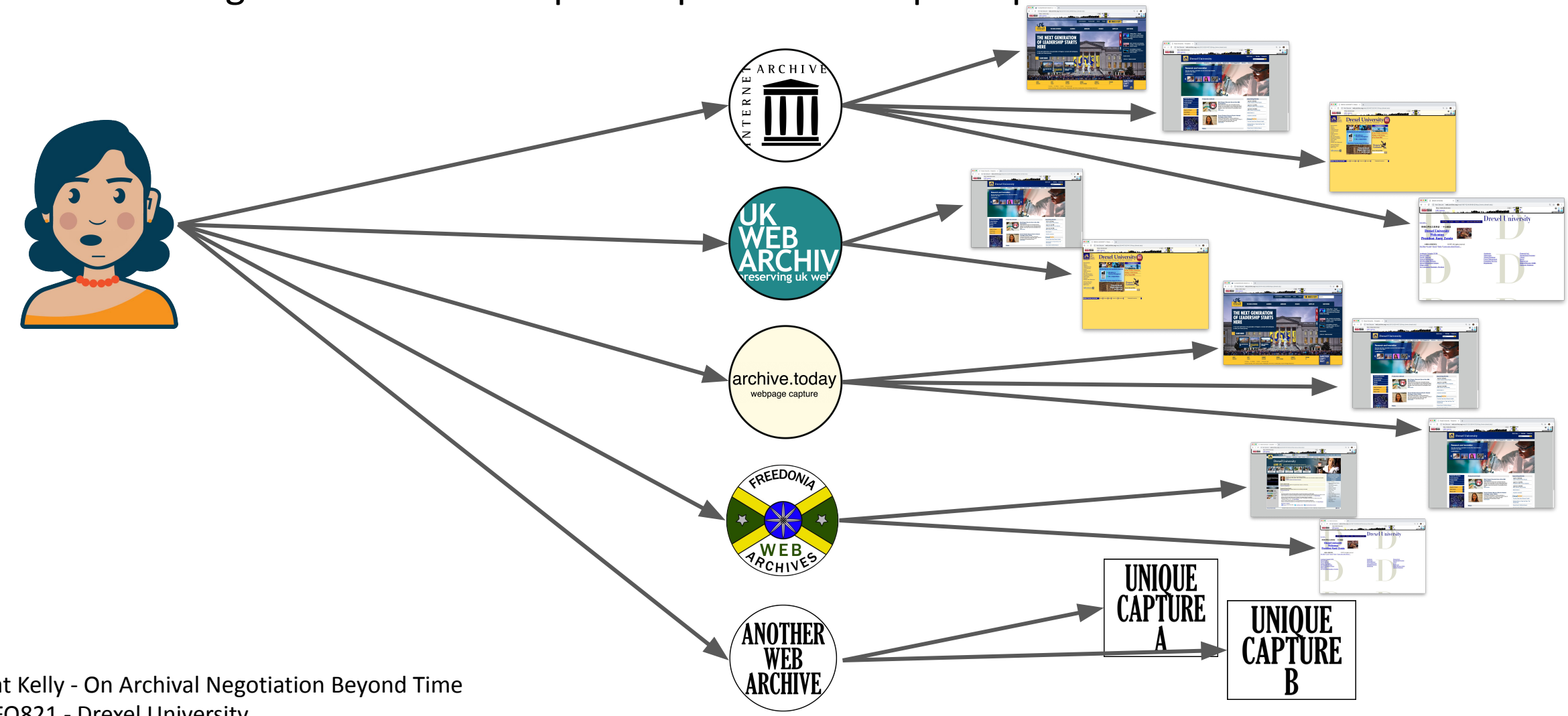
Internet Archive Has an Extensive History

but it is not comprehensive

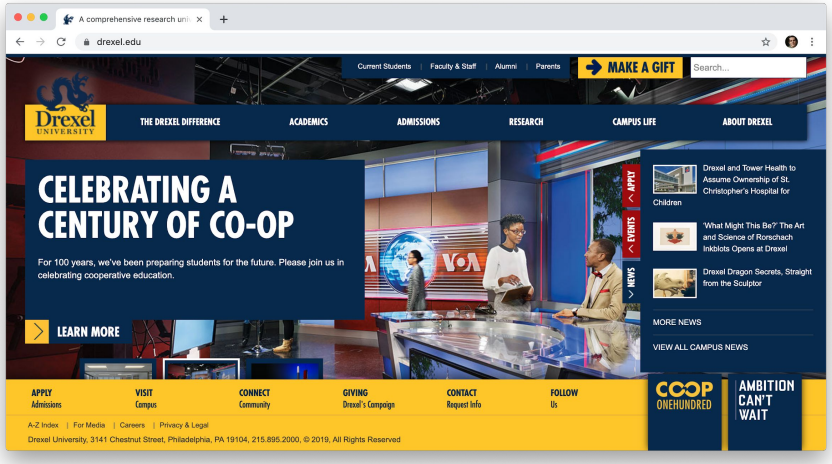


Multiple Archiving Efforts

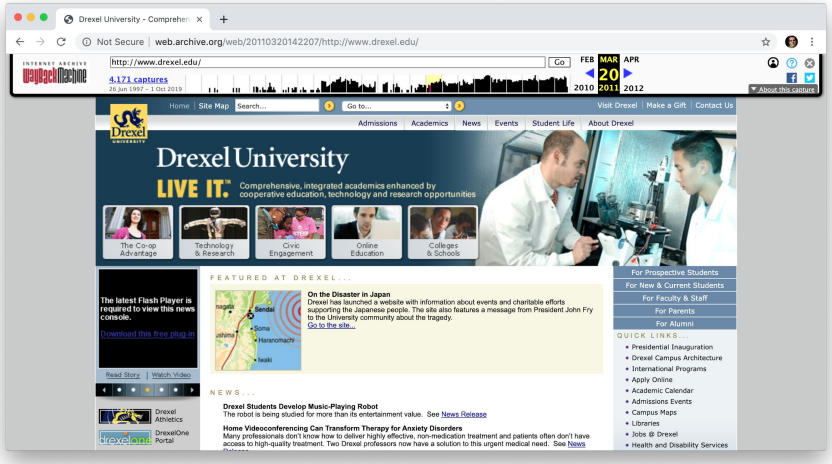
including more sources helps complete the temporal picture



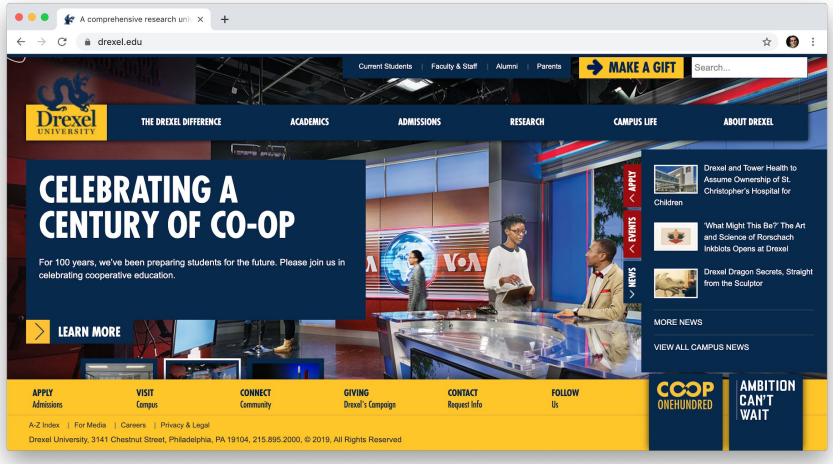
Accessing Web Archives



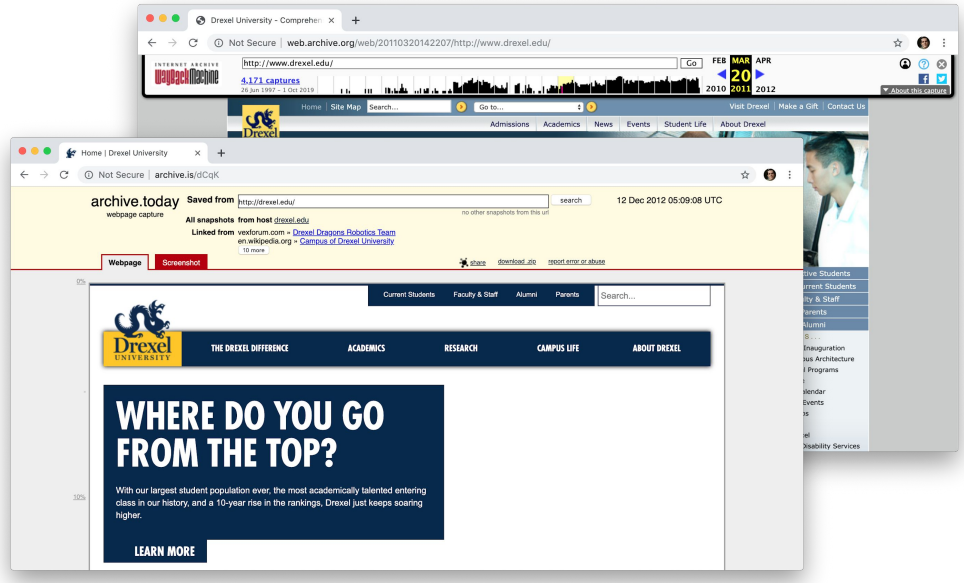
- Association with a live web URI
  - https://drexel.edu
- And an archived web URI
  - http://web.archive.org/web/20110320142207/http://www.drexel.edu/
- Should be straightforward



Accessing Web Archives



- Association with a live web URI
  - https://drexel.edu
- And an archived web URI
  - http://web.archive.org/web/20110320142207/http://www.drexel.edu/
- Should be straightforward
- Except URIs are opaque
  - Semantic should not be inferred



http://archive.is/dCqK

The Memento Framework

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

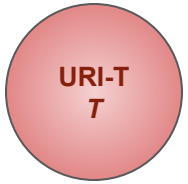
- Provides way to associate live Web URIs (URI-Rs)
  - https://drexel.edu
- With URIs of archived Web pages (URI-Ms)

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

  - http://archive.is/dCqK

TimeMaps & TimeGates

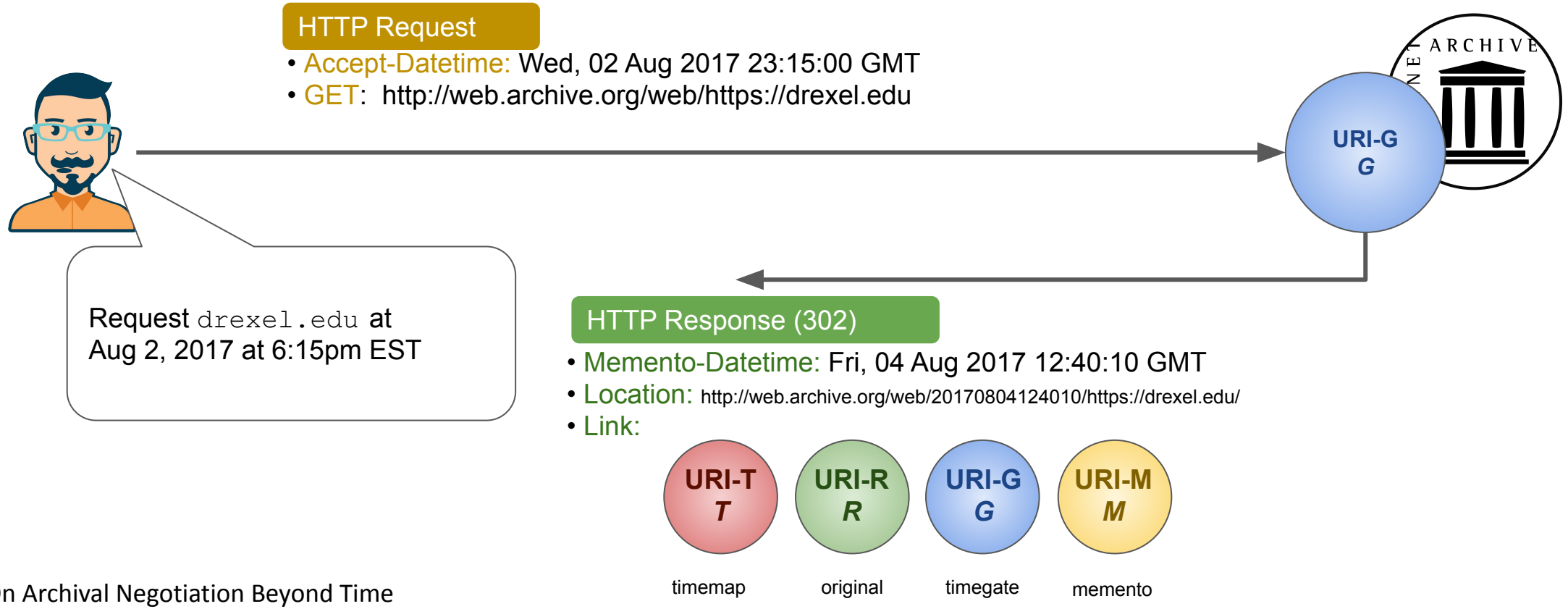
- TimeMaps – listing of URI-R, URI-Ms, and associated metadata
  - e.g., relative relation, datetime



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

Content Negotiation in Time

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



TimeMaps Show Limited Information

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

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

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

Same TimeMap Metadata in Multiple Formats

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

Memento entry in Link (RFC 7089) TimeMap

Memento entry in CDXJ TimeMap

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

Memento (URI-M)

Relative Relations

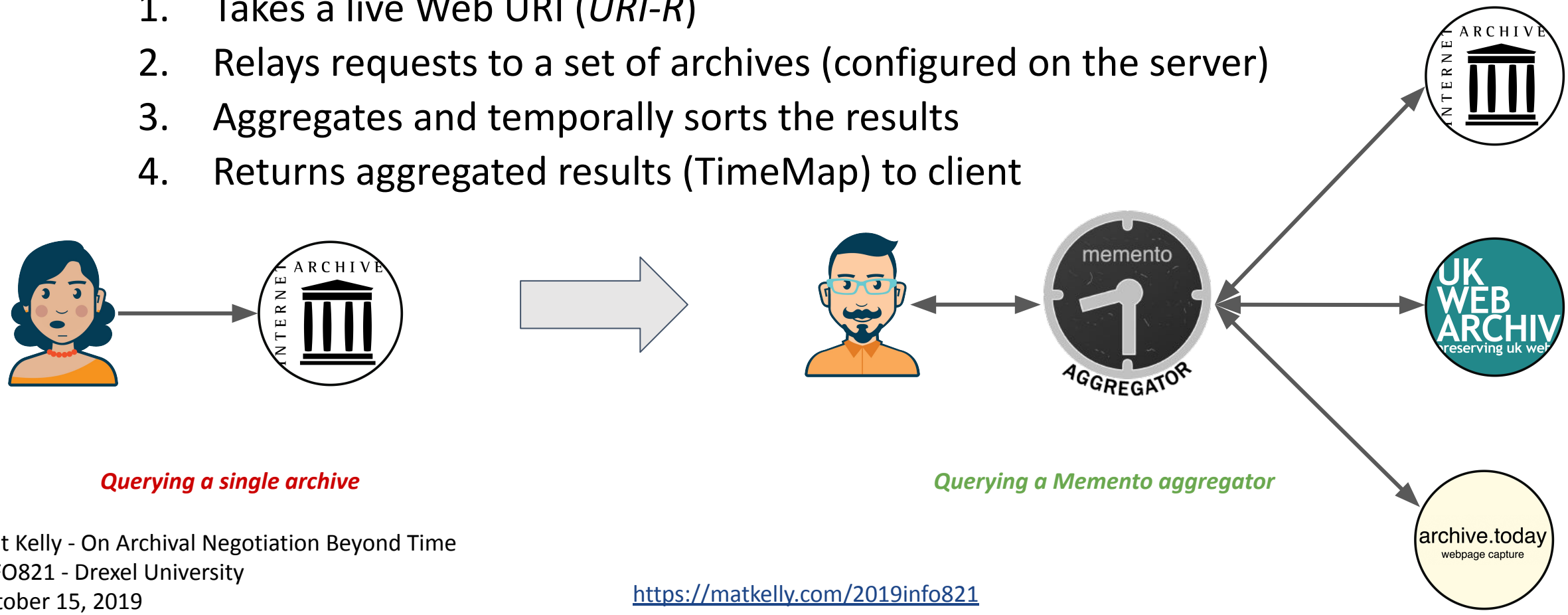
Memento-Datetime

Memento Aggregation

filling temporal gaps by using multiple sources

Memento aggregators are “queryable” Web services that:

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



Querying a single archive

Querying a Memento aggregator

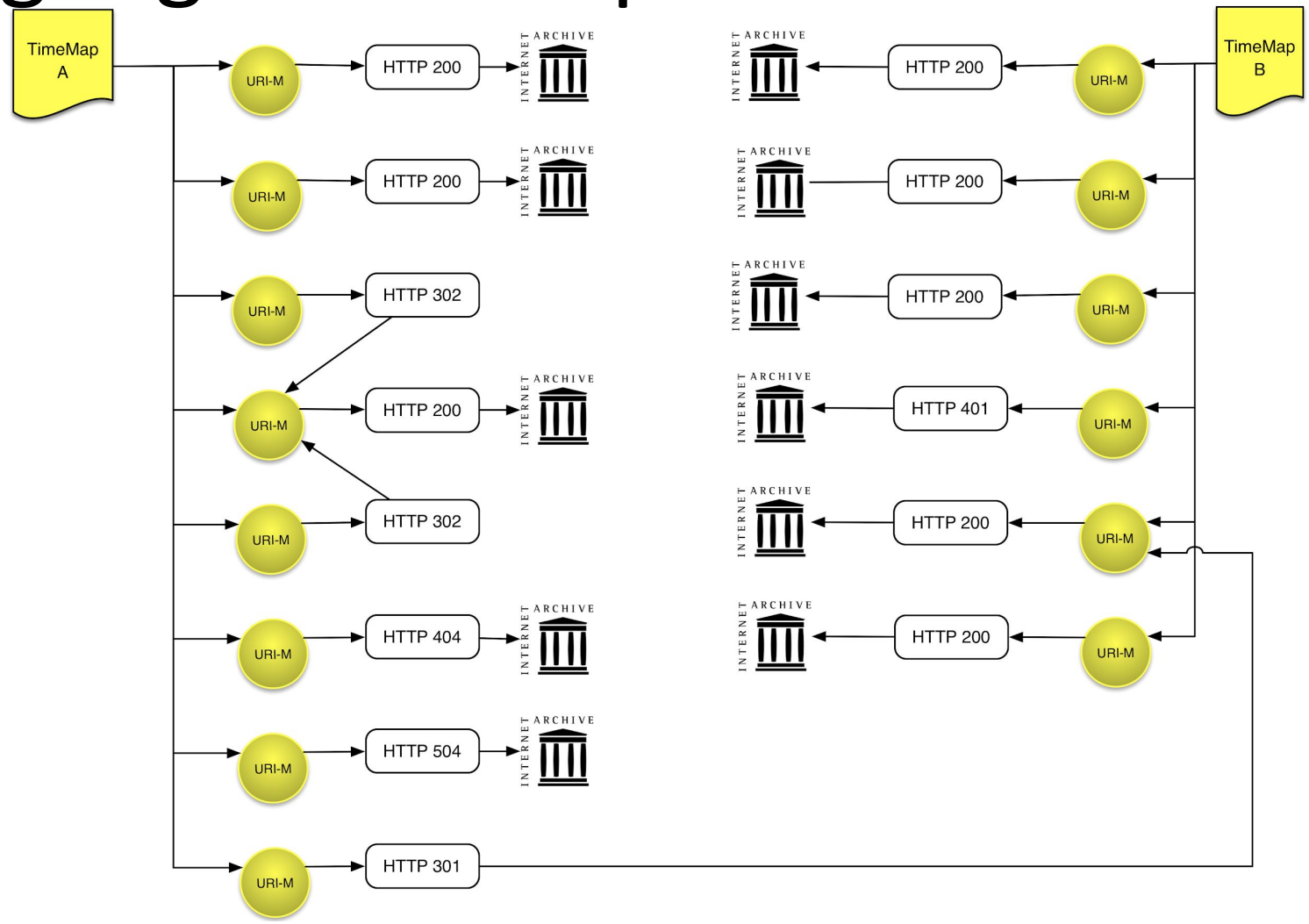
Which Archives are Queried?

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

more info, see:

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

80% of google.com Captures are Redirects



Kelly et al., “Impact of URI Canonicalization on Memento Count,” In Proc. of ACM/IEEE JCDL 2017, pp. 303-304.
(JCDL 2017, arXiv)

Content-based Attributes

- Require dereferencing URI-M
- If retained (e.g., status code) → more efficient archival usage
  - fewer wasted requests
  - more representative of non-redirecting capture quantity
- Surfaced attributes reduces retrieval cost

Derived Attributes

- Require calculations based on resource representations
  - Including embedded resources
- Expensive (time, compute, space)
- Caching Potential

apple.com over time

On Hashing

- Conventional hashing (e.g., md5) is sensitive to any change
  - Even a subtle change → drastically different hash
  - Hello INFO 821! → 17d89bfed2a1000eea8e2bb0cdcc2bdf
  - Hello INFO821! → bd9c94de002c46a984d15874e8c7d3fa
- SimHash indicates the representative partial change
  - Only the portion of the hash representative of the content changes
  - Hello INFO 821! → e7478cec8129**b159a561**c630a4c50261
  - Hello INFO821! → e7478cec8129**a90cb312**c630a4c50261

Archival Thumbnails: a SimHash Use Case

- Generating Thumbnails of **all** captures of a URI-R would take a long time, highly redundant
- Generating a SimHash for the HTML of each memento can serve as an indicator for significant change of a Web page
  - When a significant change has occurred, generate a thumbnail

- AlSum and Nelson[1] established a threshold Hamming distance of k=4 using a 64-bit SimHash

[1] AlSum and Nelson, Thumbnail Summarization Techniques for Web Archives, In Proc. of ECIR 2014, pp. 299-310.

Consider Memento Enrichment

what if TimeMaps could be richer?

- HTTP Status code – only view non-redirects
  - Associate with URI-M in TimeMap to make TM more representative
- Generated hash (to detect change)
  - e.g., some significant important value has changed
- SimHash – to detect degree of change
  - Sites that change beyond a threshold to indicate significance
- Generated screenshot URI
  - Potentially temporally & spatially expensive

Recall: Memento TimeMaps

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

Memento entry in Link (RFC 7089) TimeMap

- **Rigid Syntax/Semantics**
- **Well-supported**

Memento entry in CDXJ TimeMap

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

- **Extensible attributes**
- **Still new**

Memento (URI-M)

Relative Relations

Memento-Datetime

CDXJ “TimeMaps” for Richer Entries

```
19981212013921 {
 "uri": "http://localhost:8080/20101116060516/http://facebook.com/",
 "rel": "memento",
 "datetime": "Tue, 16 Nov 2010 06:05:16 GMT",
 "status_code": 200,
 "digest": "sha1:LK26DRRQJ4WATC6LBVF3B3Z4P2CP5ZZ7",
 "damage": 0.24,
 "simhash": "6551110622422153488",
 "content-language": "en-US", "access": {
  "type": "Blake2b", "token": "c6ed419e74907d220cdf0cc5647ef0a3..."
 }
}
```

- Is it a redirect/error?
- Has it changed at all?
- What’s the quality of the capture?
- Where and how drastic was the change?

These are just **sample addition attributes** that would be useful for further exploration using TimeMaps beyond what is **currently supported by Memento using Link format**.

Archival Content Negotiation in Dimensions Beyond Time

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

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

(note the potential for combinatorial complexity)

Relevance to Information Science

(points needing further research)

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

Contact with questions, comments, or any interest:

mkelly@drexel.edu