To Request Is Human, To Retrieve Divine

Mat Kelly, PhD
Assistant Professor, Information Science, Drexel CCI

Drexel CCI Doctoral Student Association (DSA)
Research Showcase
May 28, 2024

CC BY SA

What?

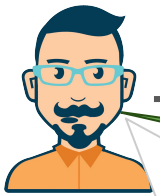
IETF RFC7089

- Web researcher, focusing on computational web archiving
  - Digital histories, encoded as bits
- Replicating an experience is complex!
  - Assembling bits does not a complete sufficient replication make
- Information Retrieval (IR) to web archives is rudimentary
  - **Time** + **location (URI)** as query parameters
- Want to ***empower*** end-user with rich, but intuitive tooling

H. Van de Sompel et al., "***HTTP Framework for Time-Based Access to Resource States -- Memento***", IETF RFC 7089, Dec. 2013, https://datatracker.ietf.org/doc/html/rfc7089

Time-Based Content Negotiation: Requesting

Time-Based Content Negotiation: Responding



HTTP Request

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

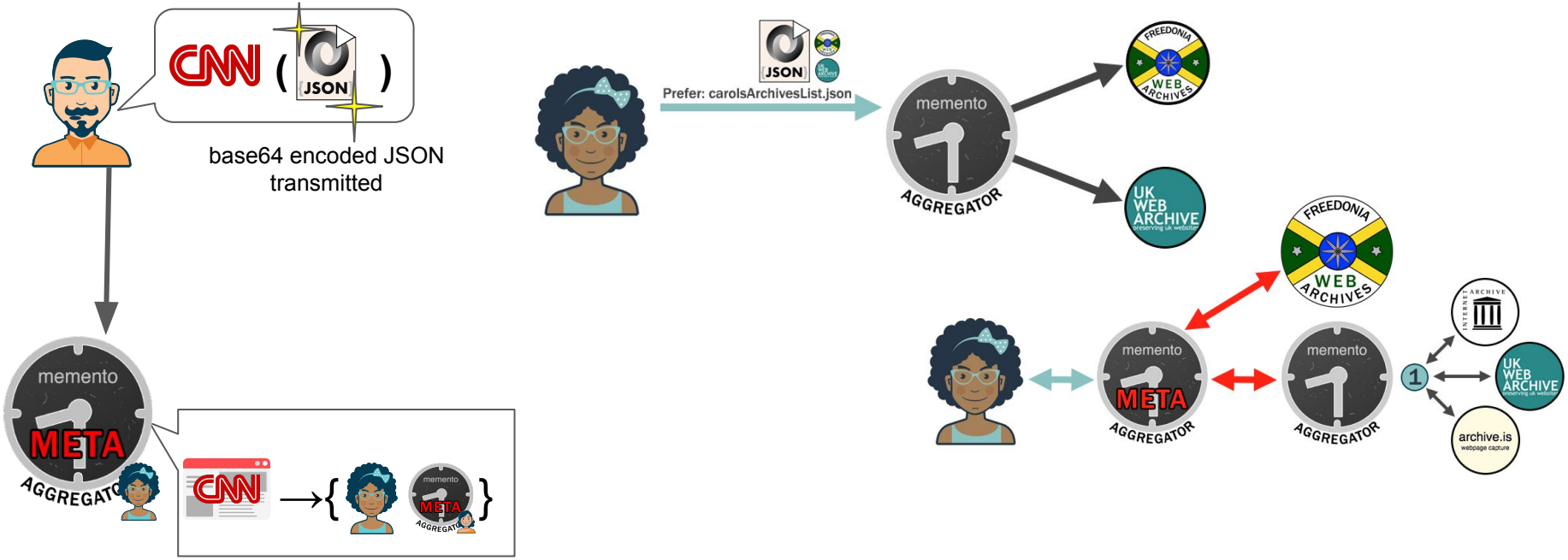
"Close enough"

Request `drexel.edu` at Aug 2, 2017 at 6:15pm EST

HTTP Response (302)

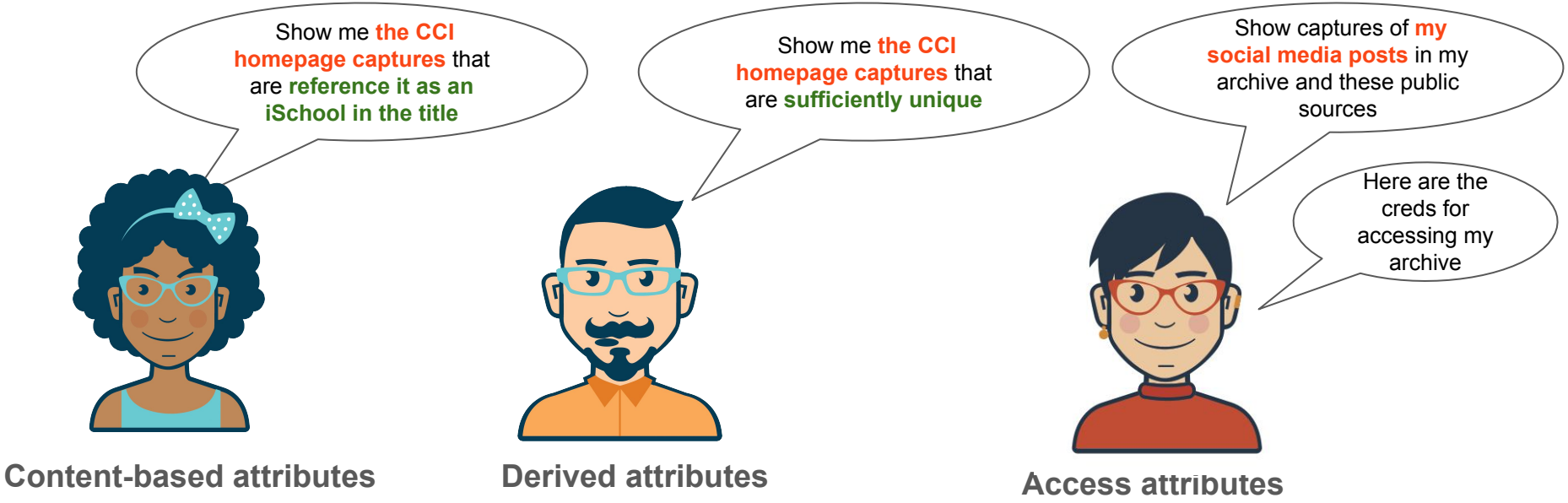
- Memento-Datetime: Wed, 02 Aug 2017 23:18:04 GMT
- Location: http://web.archive.org/web/20170802231804/http://www.drexel.edu/
- Link:

How About Just From These Repos?



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), Fort Worth, Texas, June 2018.

Attributes to Surface for Querying

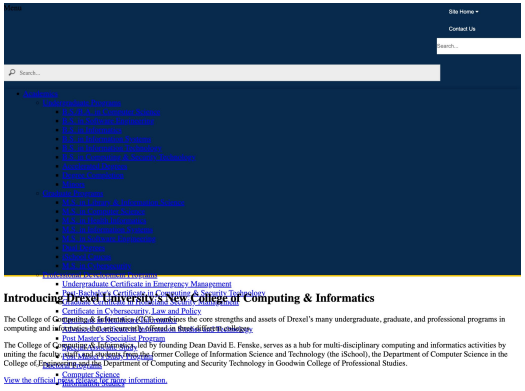


Threshold-based Retrieval

- Ability to request preference that meets some quantitative/qualitative criteria
- e.g., Memento Damage < 0.6

Captured May 28, 2024

Captured Aug. 2017
(missing image)



Captured Aug. 2015
(missing JS, CSS)

Justin F. Brunelle, **Mat Kelly**, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson, "***Not All Mementos Are Created Equal: Measuring the Impact of Missing Resources***," International Journal on Digital Libraries (IJDL), 16(3), pp. 283–301. May 2015.

Preference from Client's Perspective

HTTP Request

```
GET /tm/cdxj/https://drexel.edu HTTP/1.1
Host: somewebarchive.net
Prefer: damage="<0.4"
Prefer:archives="data:application/json;charset=utf-8;
                 base64,Ww0KICB7...NCn0="
Prefer: publicOnly
```

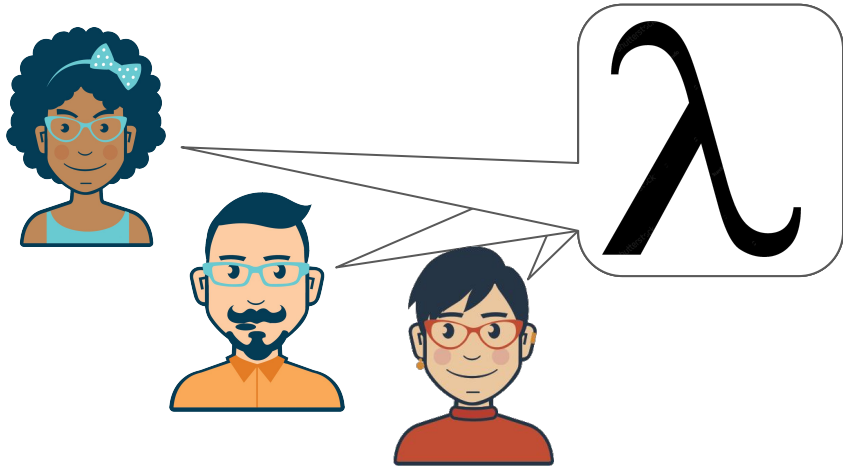
HTTP Response

```
HTTP/1.1 200 OK
Content-Length: 1207
Content-Type: application/cdxj+ors
Preference-Applied: damage="<0.4"
```

CDXJ
TM

Mat Kelly, Michael L. Nelson, and Michele C. Weigle, "A Framework for Aggregating Private and Public Web Archives," In Proceedings of the ACM/IEEE JCDL, 2018, pp. 273–282.

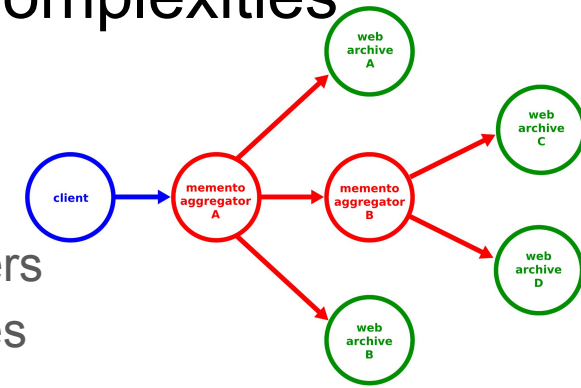
Lambda-based Functions



- User-supplied functions
  - Customizable criteria for retrieval
  - Analysis and filtering performed server-side
- Use Case (e.g.):
  - User requests all captures that are about topic healthcare reform that do not contain the mention of any of the current members of congress
  - Can AI/ChatGPT be used as the basis for parameterization? (obligatory)

Chaining Sources, Scoped IR, Querying Complexities

- Scoped sources
- Efficient querying
- Empowering user/client beyond fundamental parameters
- Integration of public, private, and personal web archives
  - With an eye toward access regulation
- Expressive thresholds
  - Currently solely based on count, but also could be sources of Damage < 0.4
  - Requires consideration of deferral, optimization, async communication, etc.



Mat Kelly, "***Exploiting the Untapped Functional Potential of Memento Aggregators Beyond Aggregation***," International Journal on Digital Libraries (IJDL), 25(1), pp. 93–104, March 2024.

Upcoming Work

- Criteria/classification of attributes for querying
- Efficient querying/resolution approaches for clients and servers
- Deploying a test bed of Memento aggregators for experimentation
  - prev. just w/ mocking
- Advance indexing approaches for web archives to facilitate richers information retrieval
- e.g., Which attributes do you surface if the function is lambda?