A Framework for Aggregating Private and Public Web Archives

Mat Kelly, Michael L. Nelson, and Michele C. Weigle
Old Dominion University
Web Science & Digital Libraries Research Group
{mkelly, mln, mweigle}@cs.odu.edu
@machawk1 • @WebSciDL

Joint Conference on Digital Libraries (JCDL)
June 5, 2018, Fort Worth, TX

OLD DOMINION UNIVERSITY

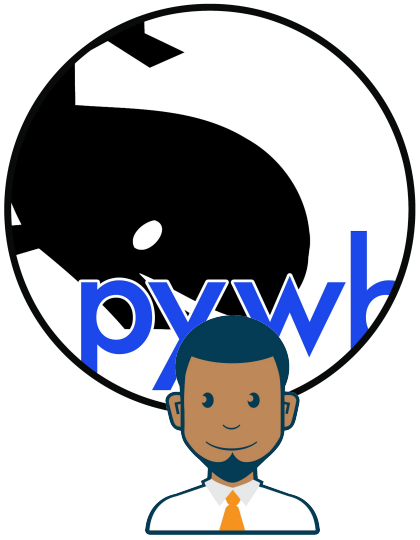
WSDL Old Dominion University

JCDL 2018
Ft. Worth, Texas

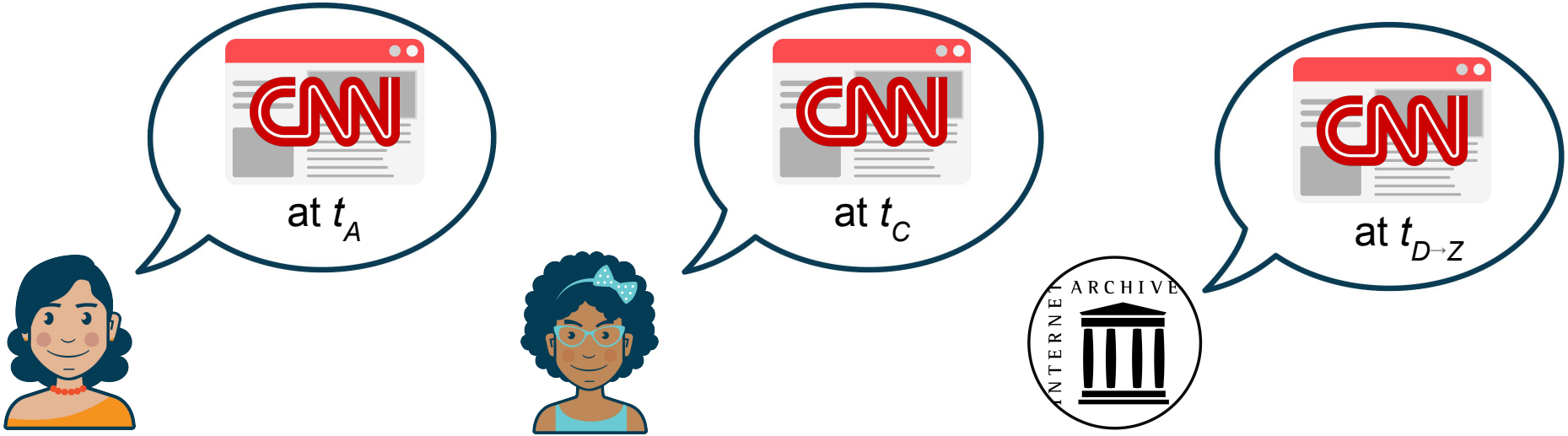
SIGIR Special Interest Group on Information Retrieval

Web Archiving

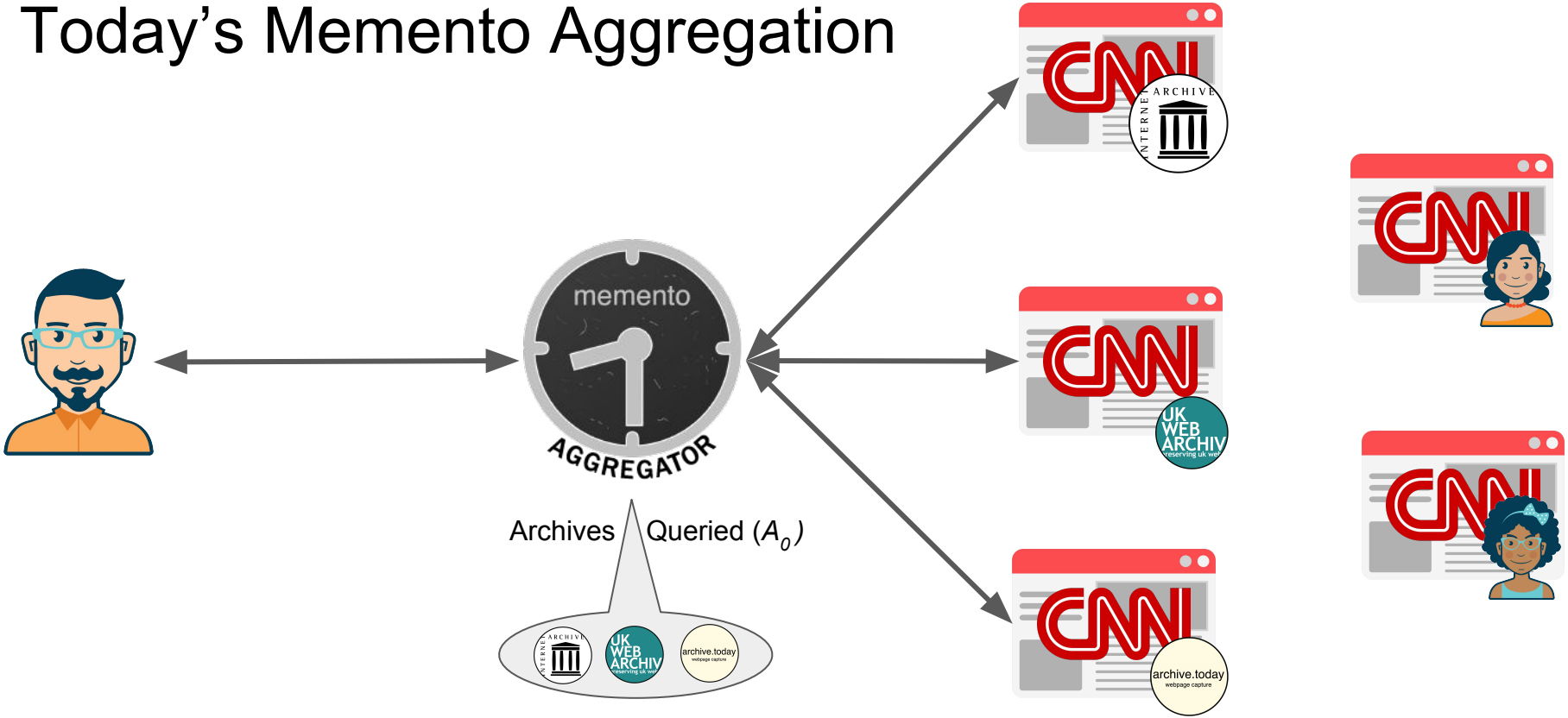
Personal + Private Web Archiving



Aggregation for a Better Picture



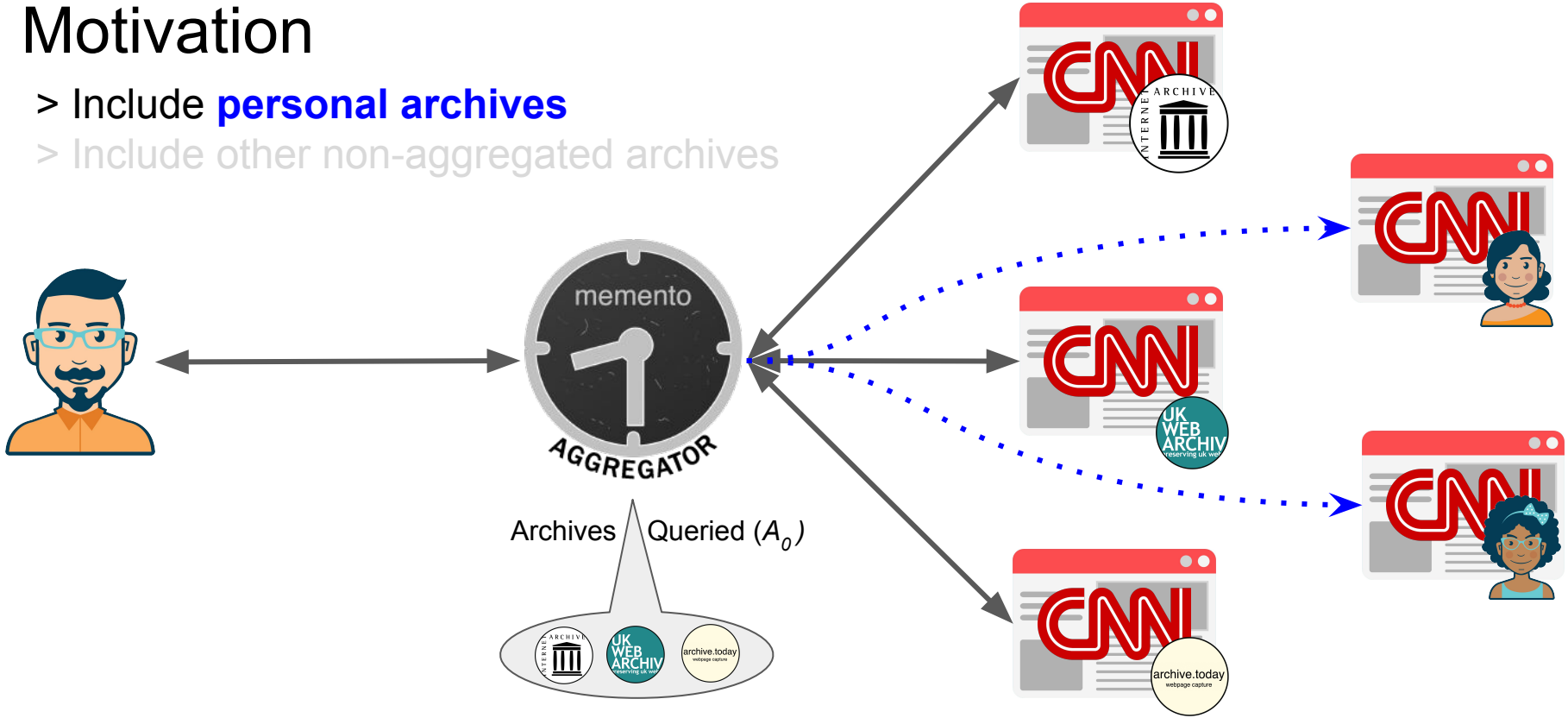
Today's Memento Aggregation



Motivation

> Include **personal archives**

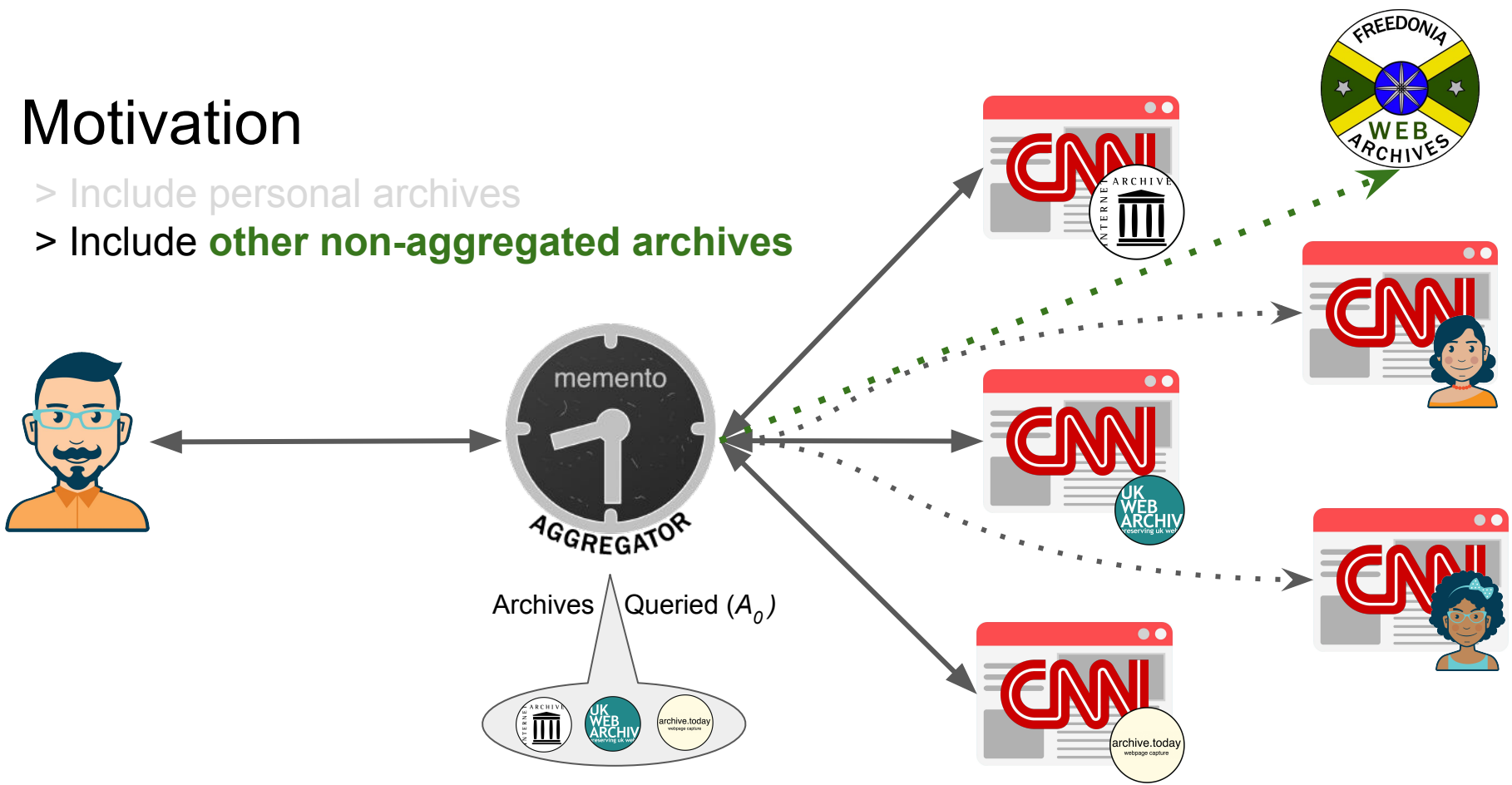
> Include other non-aggregated archives



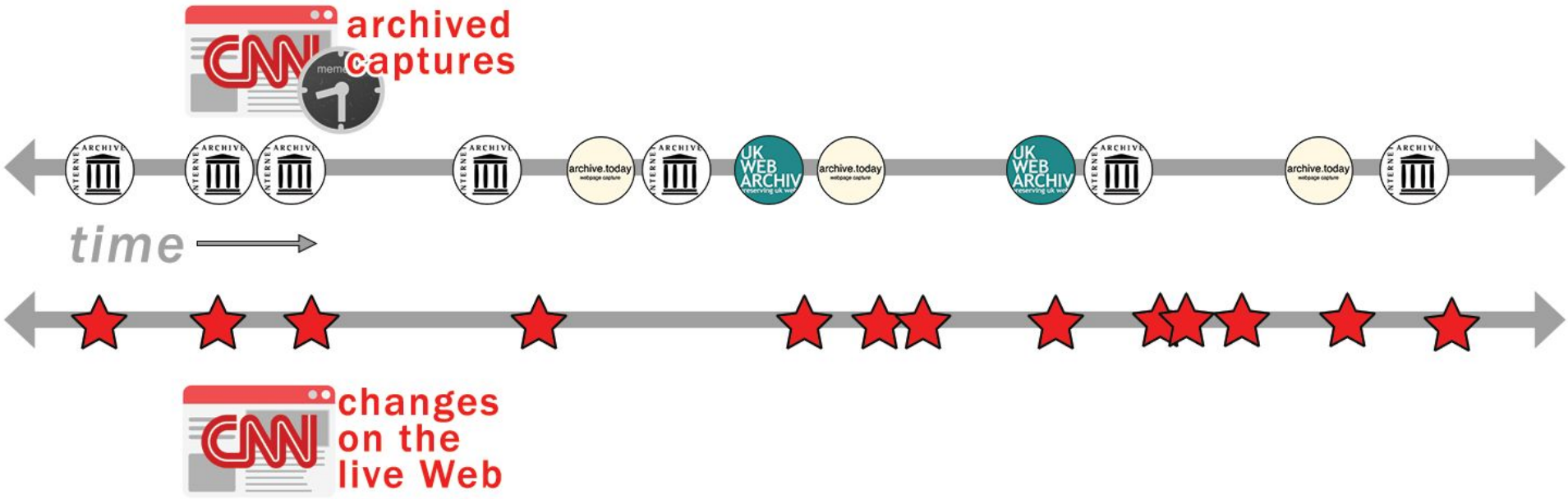
Motivation

> Include personal archives

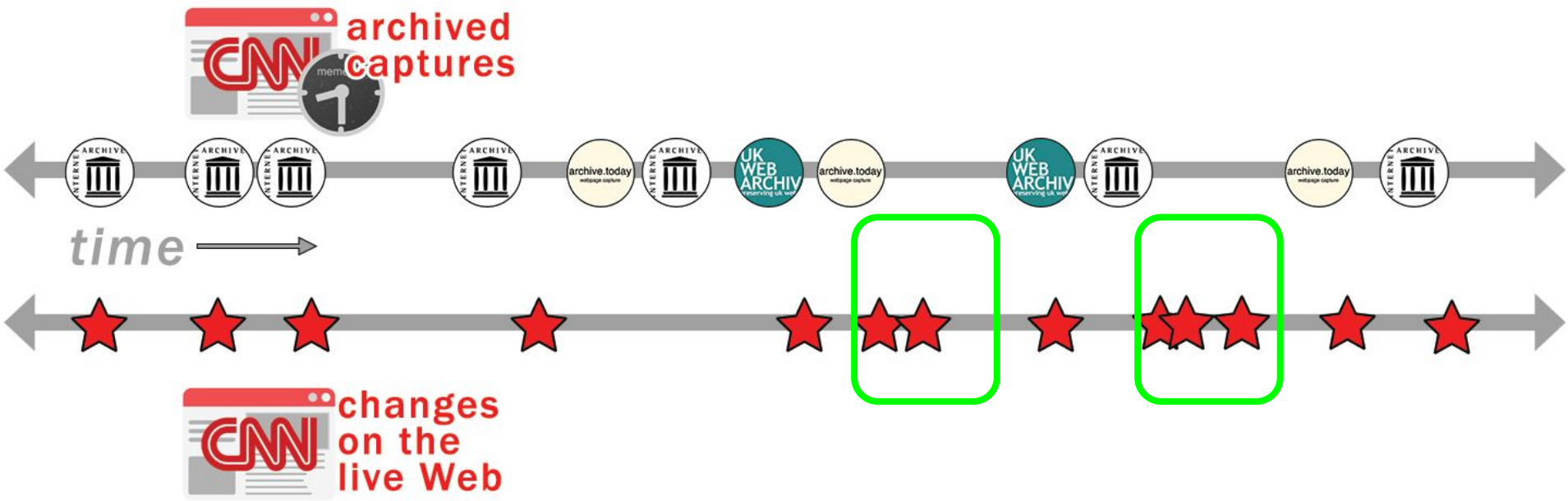
> Include **other non-aggregated archives**



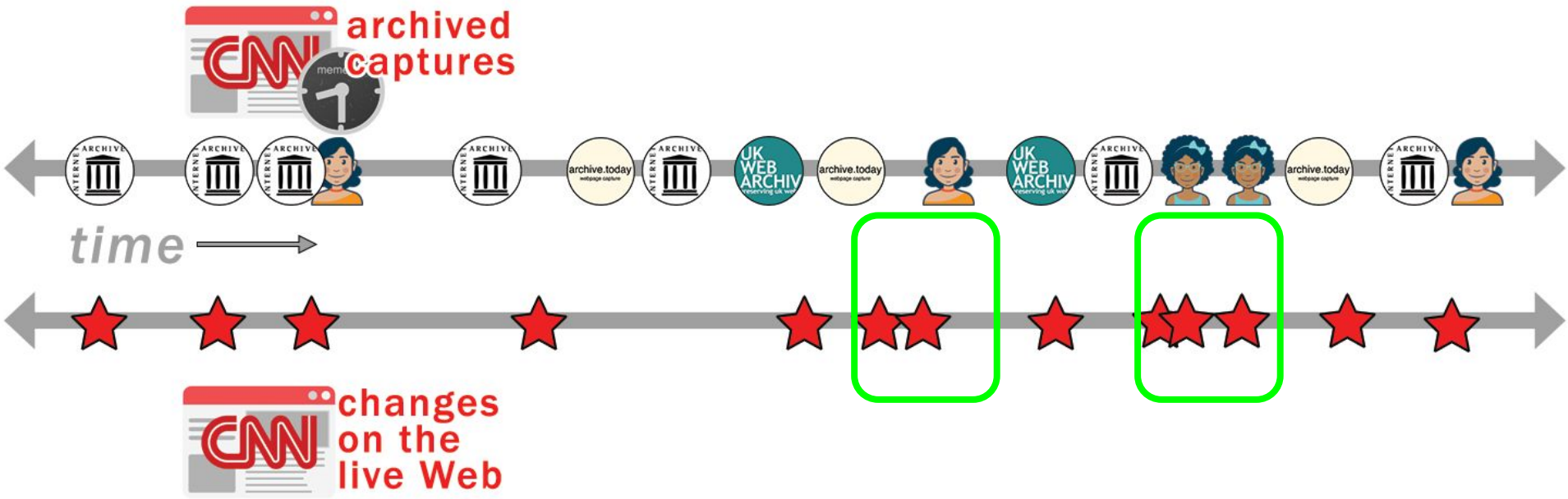
Rapidly Changing Pages May Not Be Comprehensively Captured



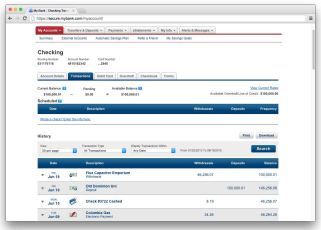
Missed Changes

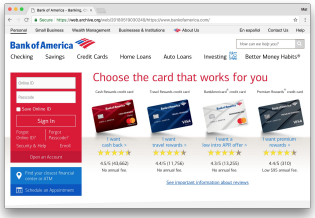


Archiving More Archives Provides a Better Picture of the Web

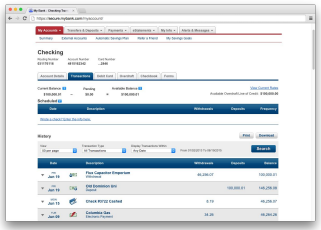


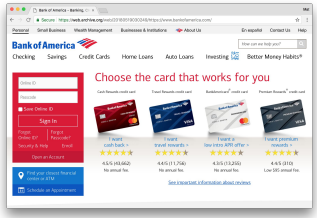
Private & Public Archives May Differ for the Same URI

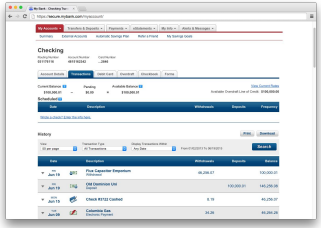




Should Public Archives *Really* Capture the Private Web?







A Framework for Aggregating Private and Public Web Archives

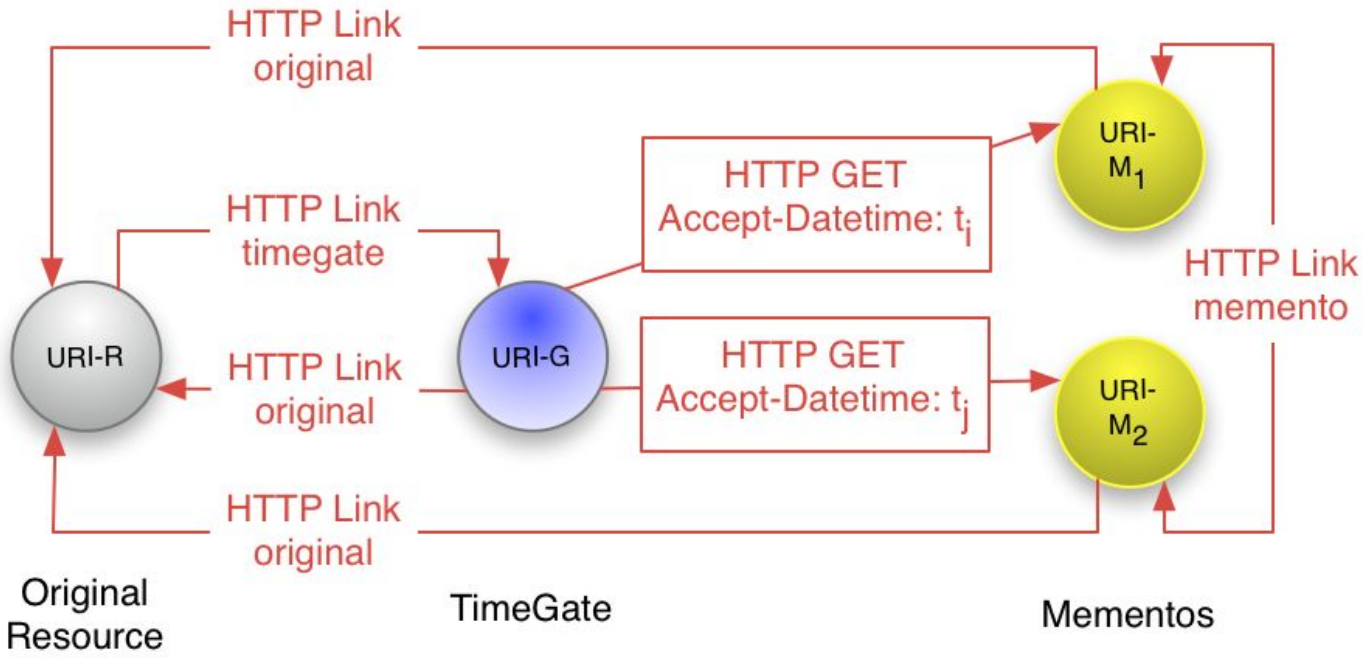
Outline

- Background and Related Work
- Memento Aggregation State of the Art
- More Expressive TimeMaps
- Query Precedence and Short-Circuiting
- Mementities & Mementity Dynamics
- Future Work and Conclusions

Outline

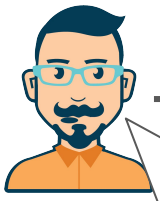
- Background and Related Work
- Memento Aggregation State of the Art
- More Expressive TimeMaps
- Query Precedence and Short-Circuiting
- Mementities & Mementity Dynamics
- Future Work and Conclusions

Background



Memento Guide: Introduction. http://www.mementoweb.org/guide/quick-intro/, January 2015.

Memento Request Example



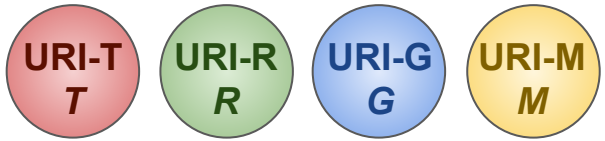
Request `cnn.com` at Sept 11, 2001 at 9am EST

HTTP Request

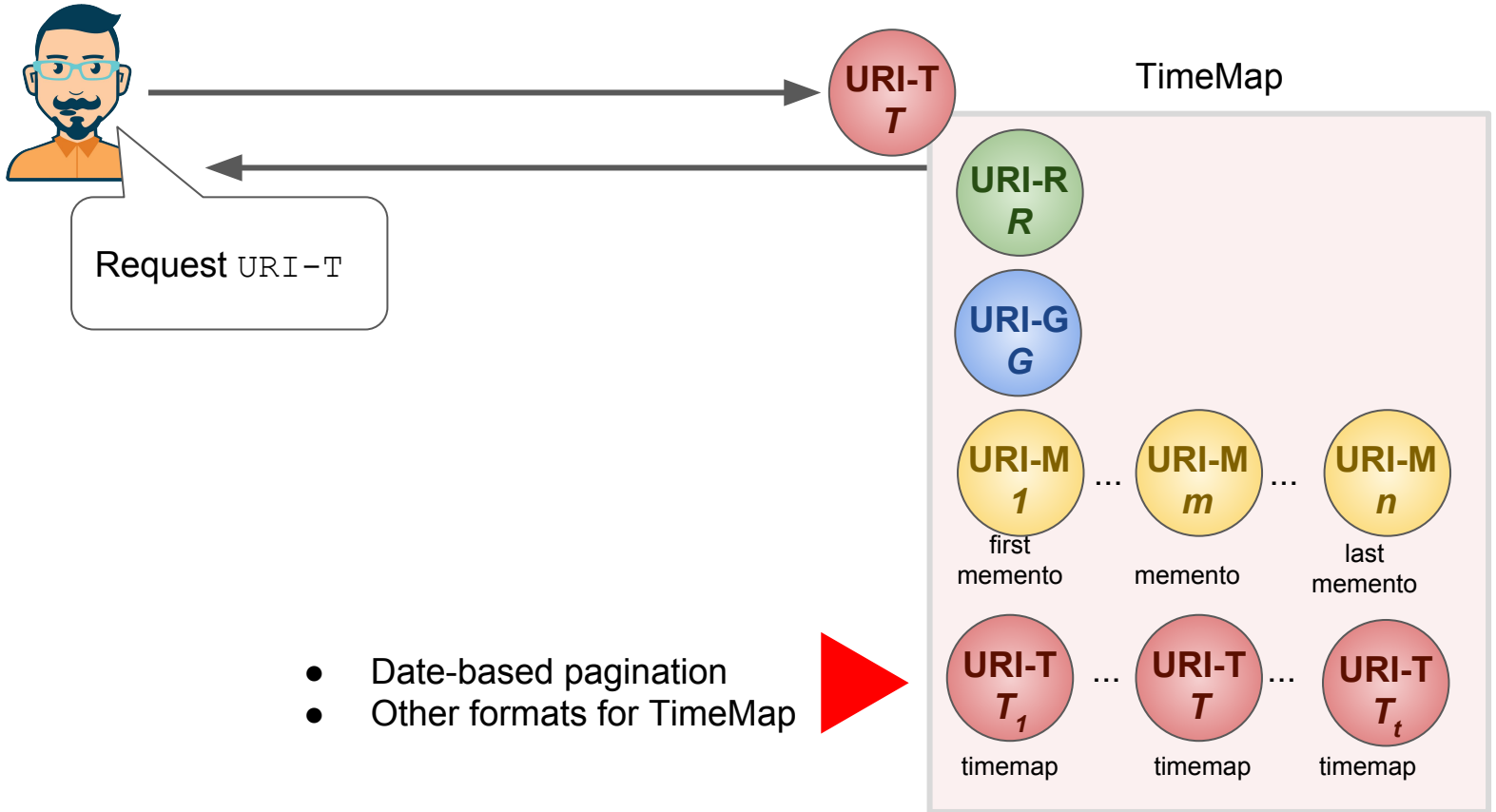
- Accept-Datetime: Tue, 11 Sep 2001 13:00:00 GMT
- GET: http://web.archive.org/web/http://www.cnn.com

HTTP Response (302)

- Memento-Datetime: Tue, 11 Sep 2001 20:03:18 GMT
- Location: http://web.archive.org/web/20010911200318/http://www.cnn.com:80/
- Link:



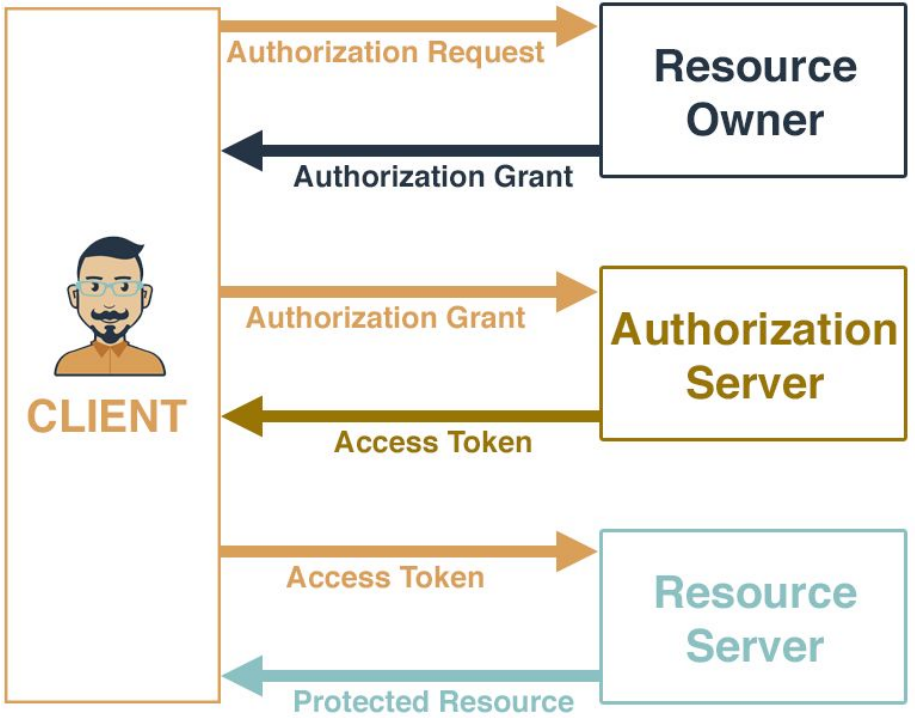
Dereferencing a TimeMap at URI-T



- Date-based pagination
- Other formats for TimeMap

Background - Privacy and Security

- Web users question trusting institutions to preserve private Web contents[1]
- OAuth 2.0 (RFC 6749) facilitates authentication cohesion of entities



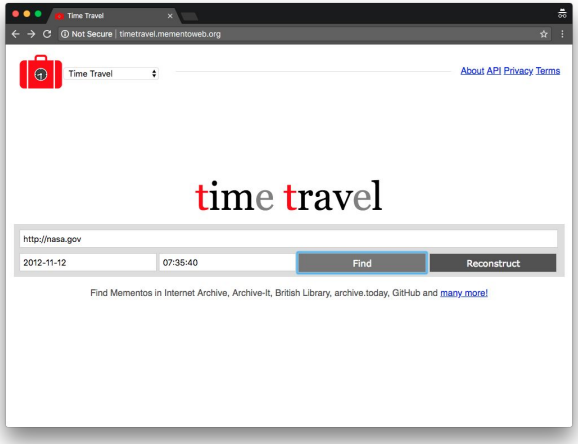
[1] Marshall and Shipman., "On the Institutional Archiving of Social Media", JCDL 2012

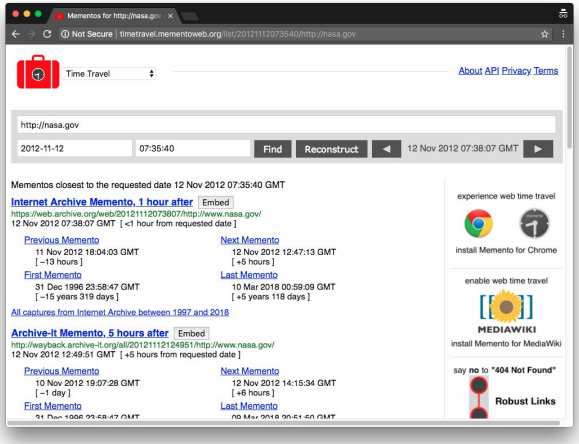
Outline

- Background and Related Work
- Memento Aggregation State of the Art
- More Expressive TimeMaps
- Query Precedence and Short-Circuiting
- Mementities & Mementity Dynamics
- Future Work and Conclusions

Memento Aggregation State of the Art

Memento Aggregation - MementoWeb



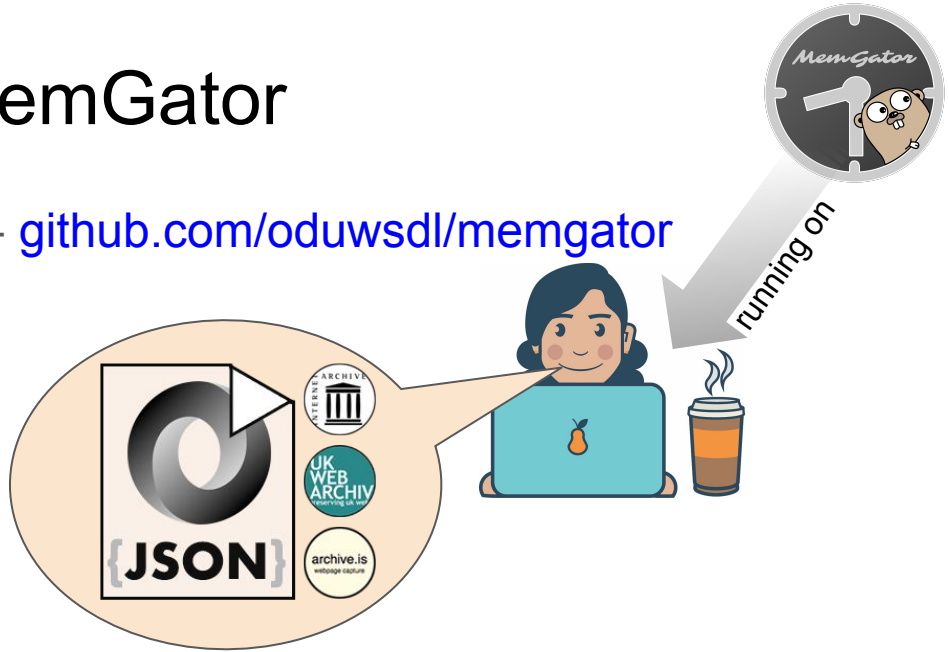


Also available via CLI:

```
$ curl http://timetravel.mementoweb.org/timemap/link/http://nasa.gov
```

Memento Aggregation - MemGator

- Open Source Memento Aggregator - github.com/oduwsdl/memgator
- Easy personal/local deployment
- Specify archive list on launch
  - Easily configurable **JSON** →
  - Use default collection if not specified
- TimeMap Formats:
  - Link
  - **JSON**
  - **CDXJ**



* Alam and Nelson, “MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go”, JCDL 2016

CDXJ: An Alternative TimeMap Format

```
<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.mat
kelly.com:80/>; rel="first memento"; datetime="Sun, 14
May 2006 12:35:11 GMT",
<http://web.archive.org/web/20060516213852/http://www.mat
kelly.com/>; rel="memento"; datetime="Tue, 16 May 2006
21:38:52 GMT",
...
<http://web.archive.org/web/20180128152125/http://matkell
y.com>; rel="memento"; datetime="Sun, 28 Jan 2018
15:21:25 GMT",
<http://web.archive.org/web/20180319141920/http://matkell
y.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

CDXJ TimeMap

Original URI (URI-R)

Other TimeMaps (URI-Ts)

TimeGate (URI-G)

Relative Relations

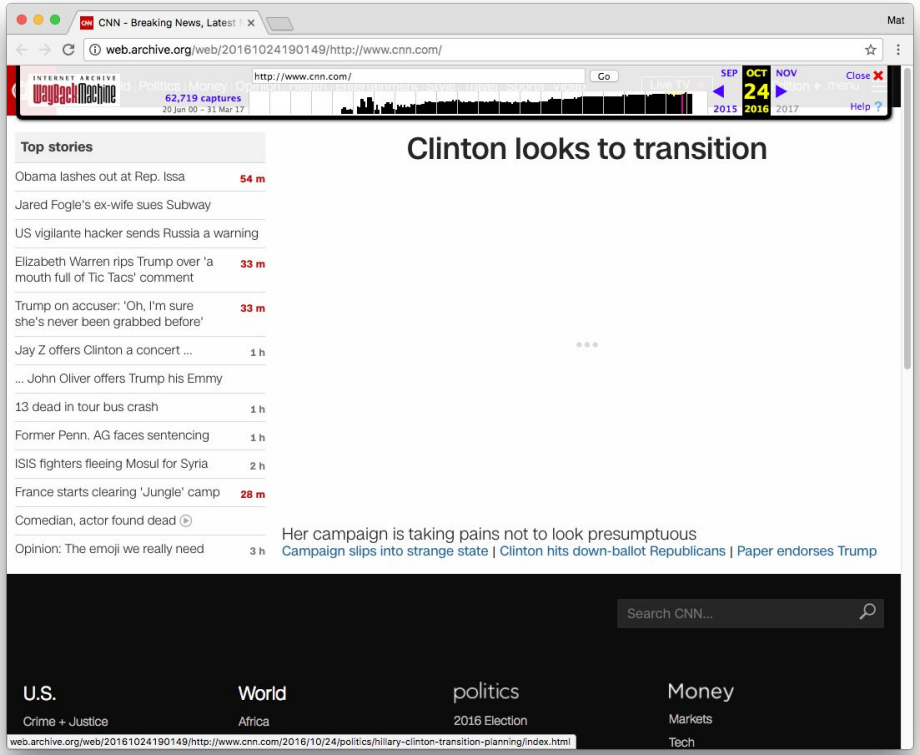
See Alam, "CDXJ: An Object Resource Stream Serialization Format", 2015

Outline

- Background and Related Work
- Memento Aggregation State of the Art
- More Expressive TimeMaps
- Query Precedence and Short-Circuiting
- Mementities & Mementity Dynamics
- Future Work and Conclusions

More Expressive TimeMaps

- Memento Quality (e.g., Damage)[1]
- How Many Captures?[2]
- How Many Are Identical?[2,3]
- Other Attributes of Mementos...



[1] Brunelle *et al.*, JCDL 2014, IJDL 2015
[2] Kelly *et al.*, JCDL 2017
[3] AlSum and Nelson, ECIR 2014

Additional TimeMap Attributes

Content-based Attributes

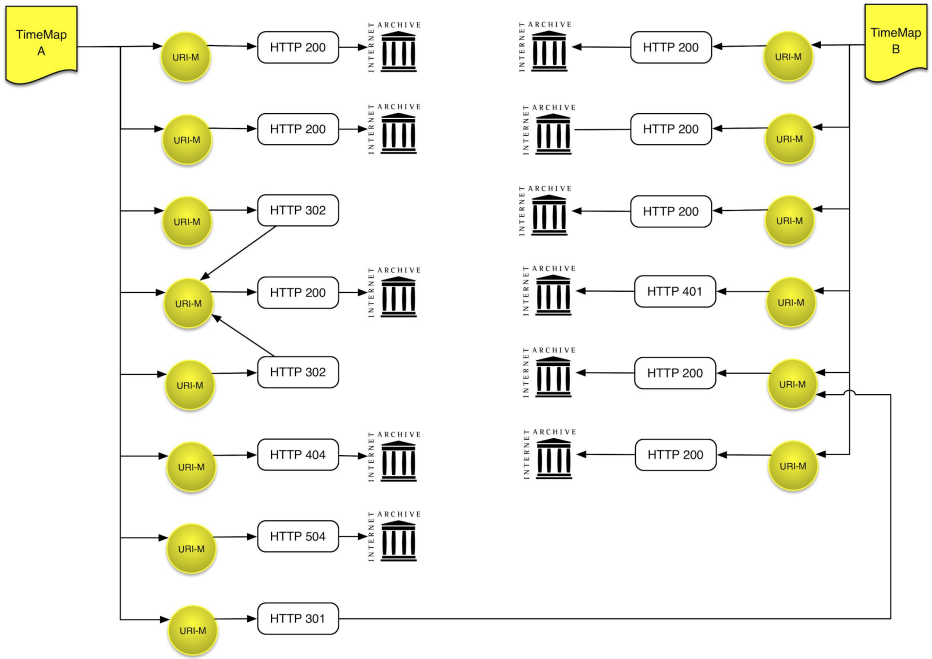
Derived Attributes

Access Attributes

TimeMap Enrichment: Content-Based Attributes

- Status Code[1]
- Content-Digest
  - In WARC & CDX
  - Not all archives expose CDX

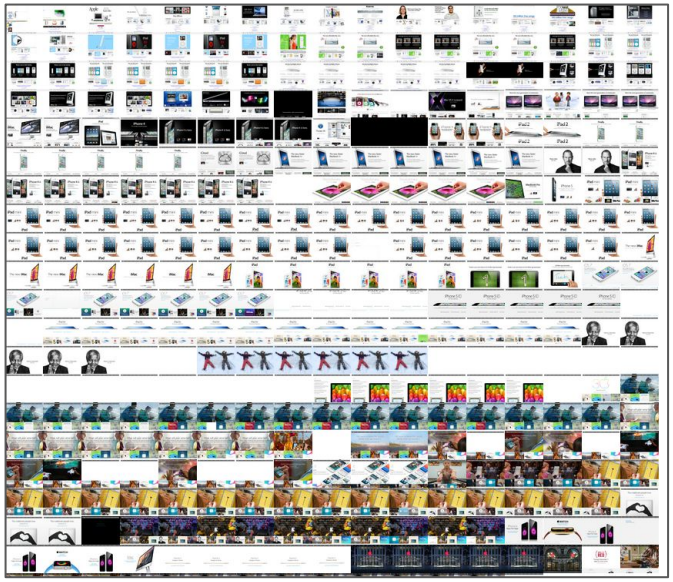
- Would allow more info about mementos without requiring comprehensive dereferencing



Kelly *et al.,* "Impact of URI Canonicalization on Memento Count", JCDL 2017, arXiv 1703.03302

TimeMap Enrichment: Derived Attributes

- Thumbnails (e.g, via SimHash)[1]
  - Calculation based on root memento's HTML
- Memento Damage (JCDL 2014, IJDL)[2]
  - Requires dereferencing embedded resources

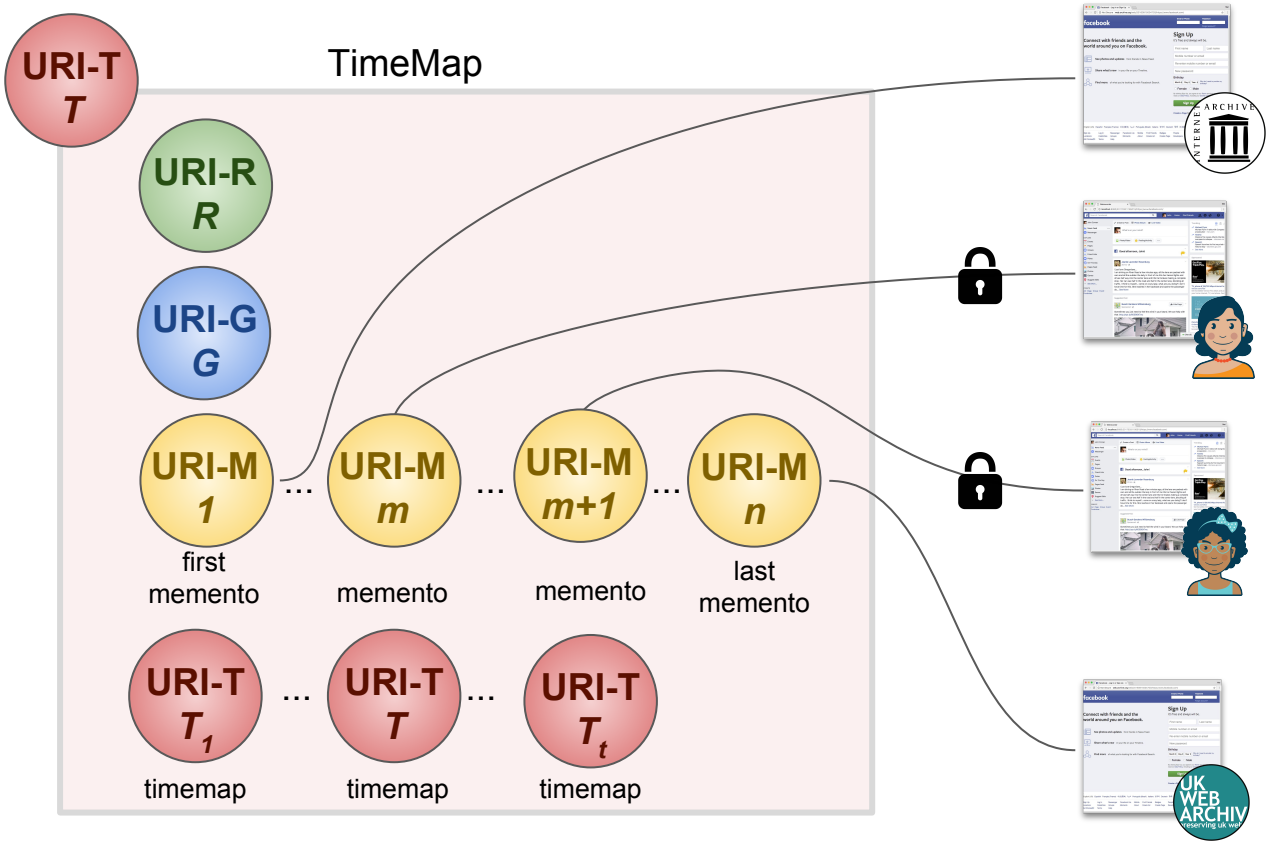


apple.com, many duplicate mementos!

[1] AlSum and Nelson, Thumbnail Summarization Techniques for Web Archives, ECIR 2014, pp. 299-310.
[2] Brunelle *et al.,* "The Impact of JavaScript on Archivability," IJDL, 17(2), pp. 95-117. January 2016.

TimeMap Enrichment: Access Attributes

- How to distinguish Private captures In a TimeMap?



TimeMap Enrichment - in a CDXJ TimeMap

Line breaks added for clarity, CDXJ records occupy a single line

```
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": "65511106224221534 88",
  "content-language": "en-US",
  "access": {
     "type": "Blake2b",
     "token": "c6ed419e74907d220c69858614d86...ef0a3a88a41"
  }
}
```

Content-based attributes **Derived Attributes** **Access Attributes**

Outline

- Background and Related Work
- Memento Aggregation State of the Art
- More Expressive TimeMaps
- Query Precedence and Short-Circuiting
- Mementities & Mementity Dynamics
- Future Work and Conclusions

Query Precedence

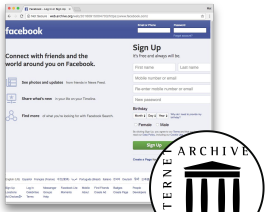
- More control of querying in series and parallel

1

2

3

3



"Check my archive first, then Carol's, then all public archives."

See Atkins', "Paywalls in the Internet Archive", March 2018

Query Precedence

- More control of querying in series and parallel

1

"Check my archive first, then Carol's, then all public archives."

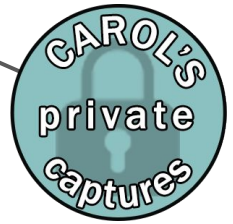
See Atkins', "Paywalls in the Internet Archive", March 2018

Query Precedence

- More control of querying in series and parallel

1

2



"Check my archive first, then Carol's, then all public archives."

See Atkins', "Paywalls in the Internet Archive", March 2018

Query Precedence

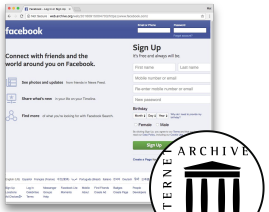
- More control of querying in series and parallel

1

2

3

3



"Check my archive first, then Carol's, then all public archives."

See Atkins', "Paywalls in the Internet Archive", March 2018

Query Short-Circuiting

- **May give priority to archive relevancy.**
- Series halt when threshold met.

1

1

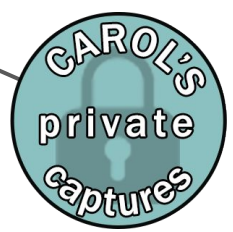
2

2

"Check private archives first. **Iff** you find no captures, only *then* check the public archives.

Query Short-Circuiting

- May give priority to archive relevancy.
- **Series halt when threshold met.**



1

1

2

2

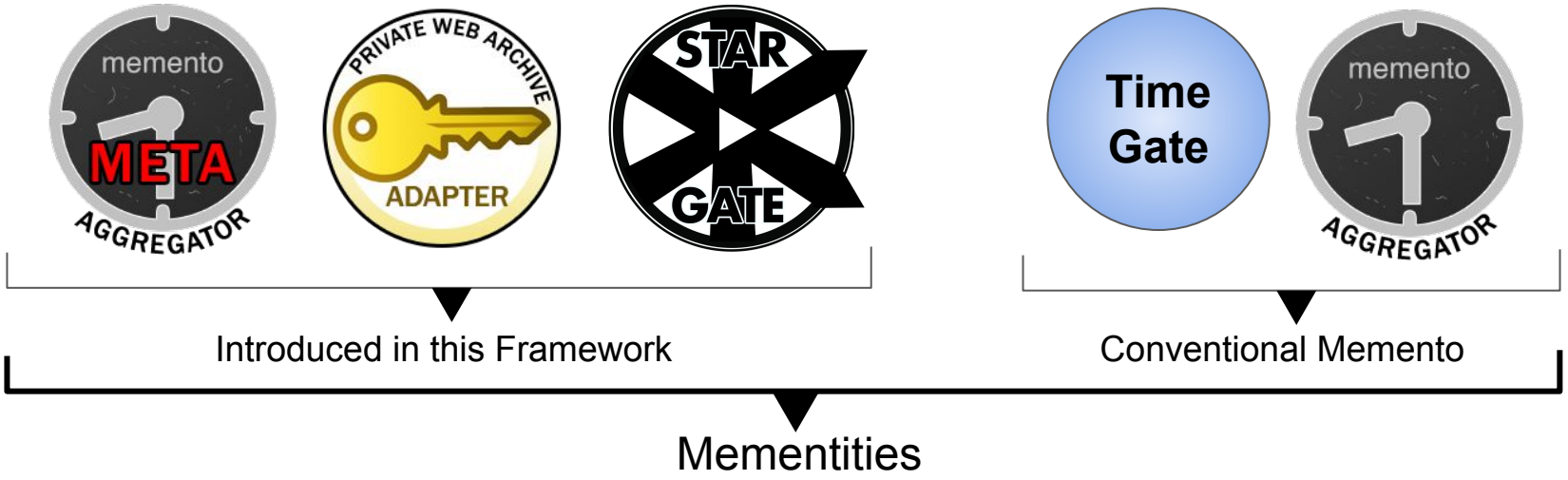
"Check private archives first. **Iff** you find no captures, only *then* check the public archives.

Outline

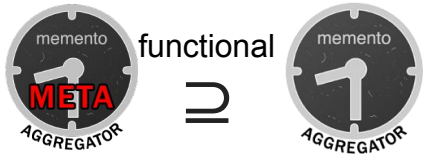
- Background and Related Work
- Memento Aggregation State of the Art
- More Expressive TimeMaps
- Query Precedence and Short-Circuiting
- Mementities & Mementity Dynamics
- Future Work and Conclusions

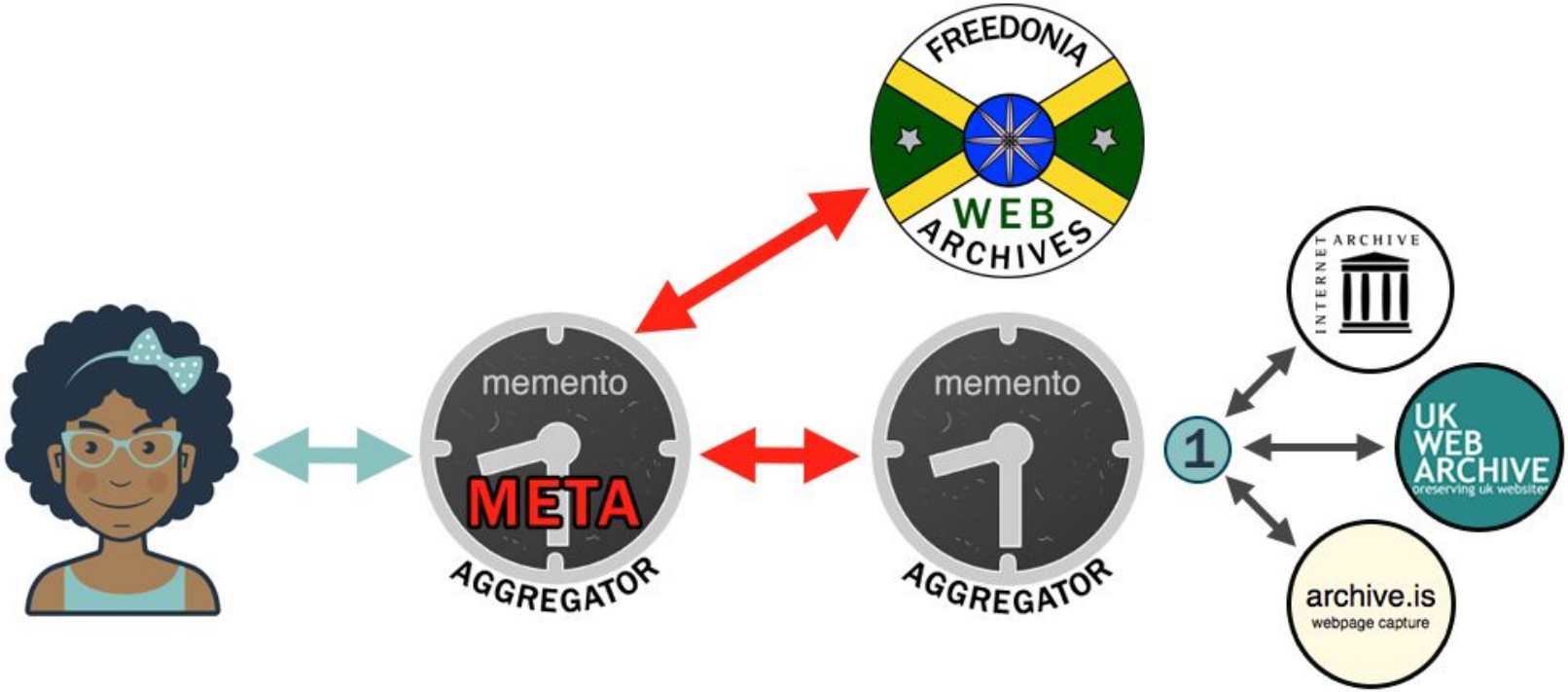
Mementities

- Memento + Entity (*entity* term already overused)



Memento Meta-Aggregator (MMA)





MMA: Archive Selection

MMA: User-Driven Archival Specification

MMA Aggregation sources

MMA_{α}:

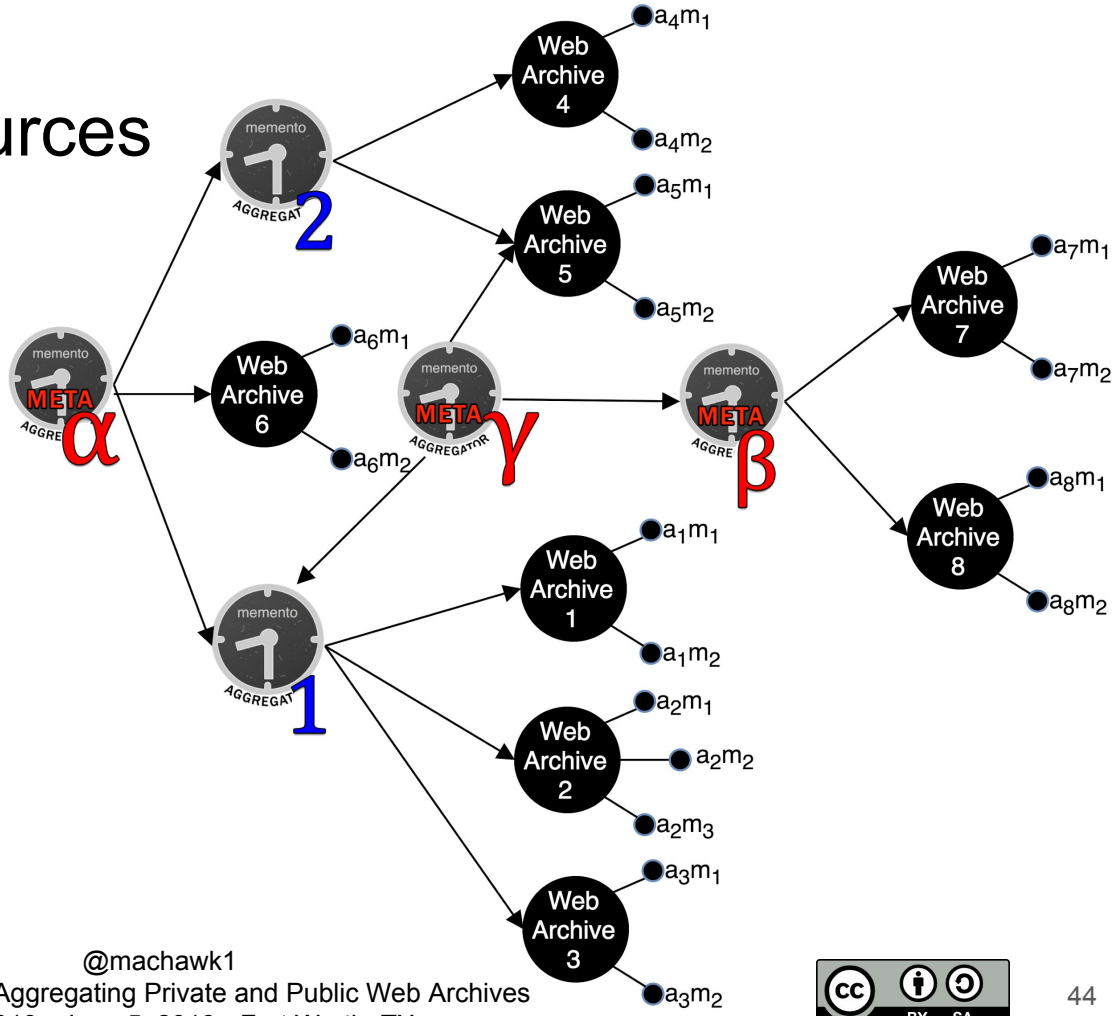
from MA_2, MA_1 and $\mathbf{WA_6}$

MMA_{β}:

from $\mathbf{WA_7}$ and $\mathbf{WA_8}$

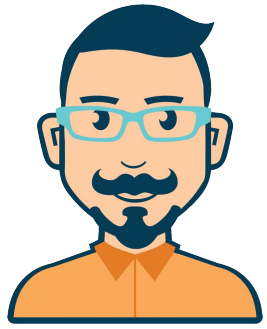
MMA_{γ}:

from MMA_{β}, MA_5, and $\mathbf{WA_1}$



MMA Dynamics By-Example

- Personal Archive Aggregation
- MMA Chaining
- Client-Side Aggregation Preference



MMA Dynamics - Personal Archive Aggregation

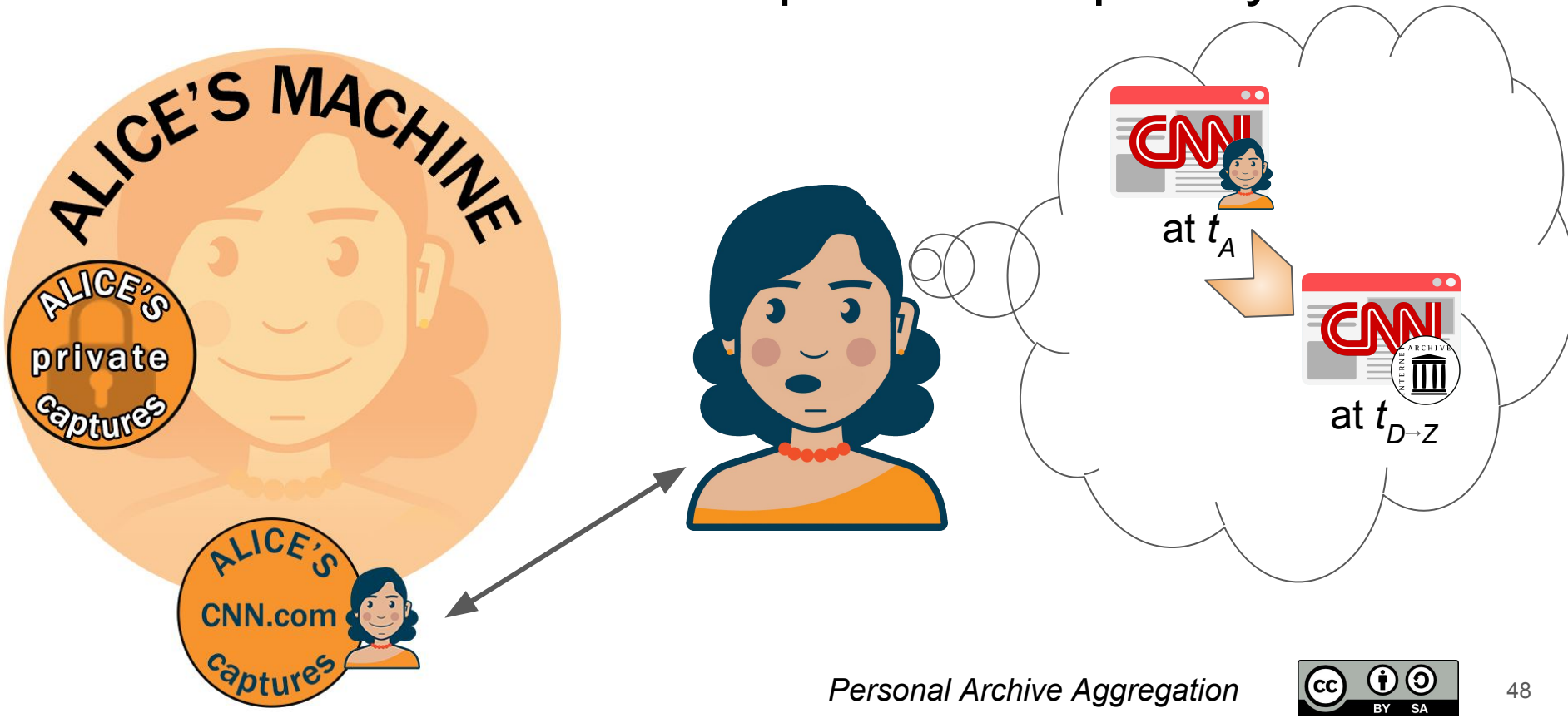
bbc | homepage

Public videos

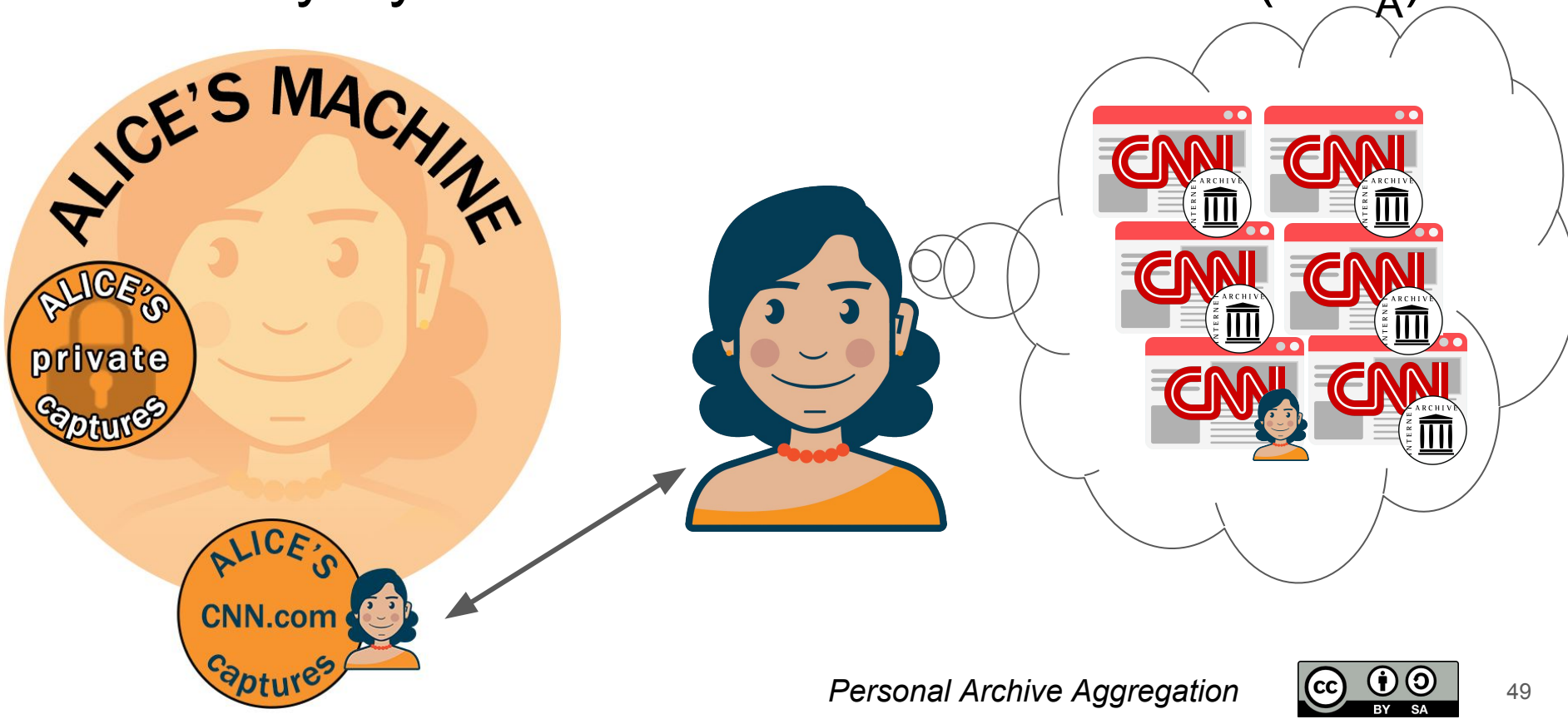
FB | bank | flickr

Alice Saves the Web

Alice Wants to See Her Captures Temporally Inline

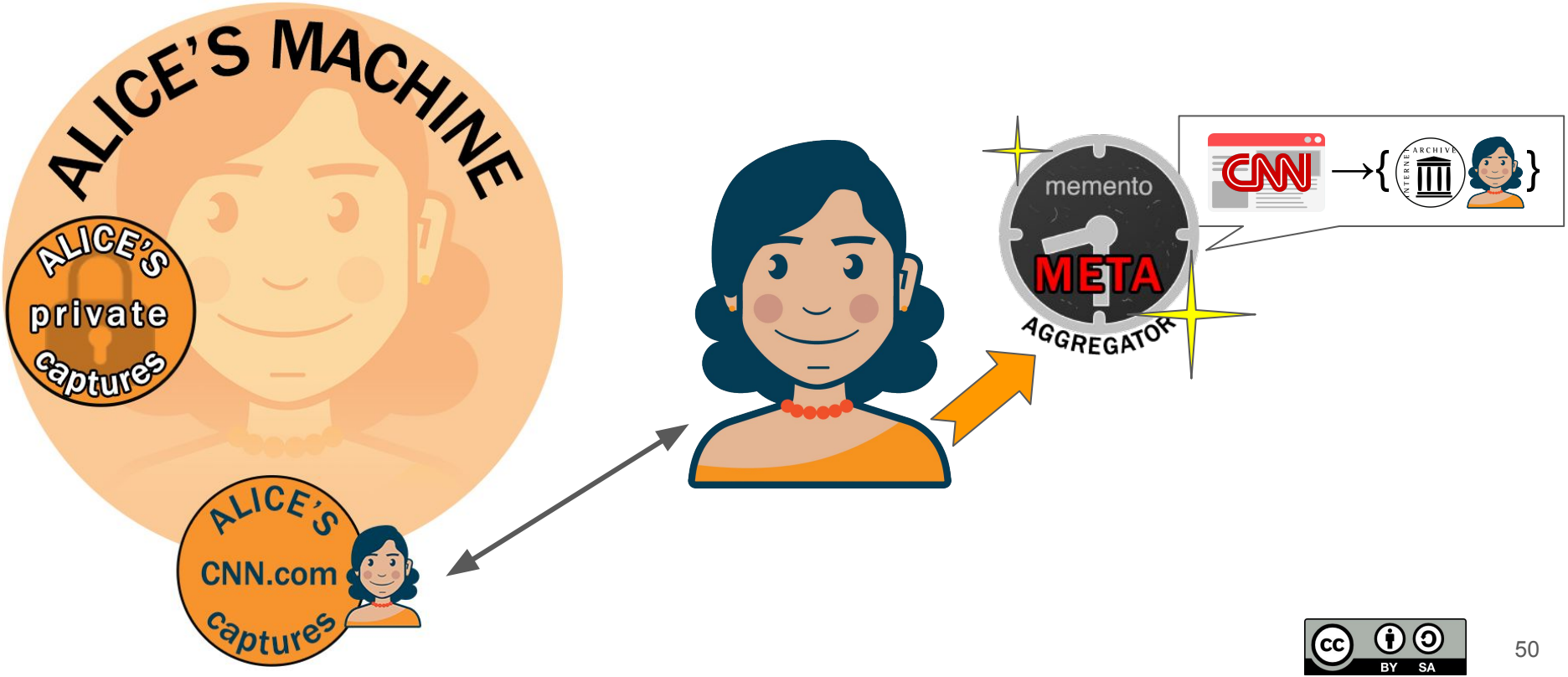


Mementity Dynamics - Alice & Her Archives (WA_A)

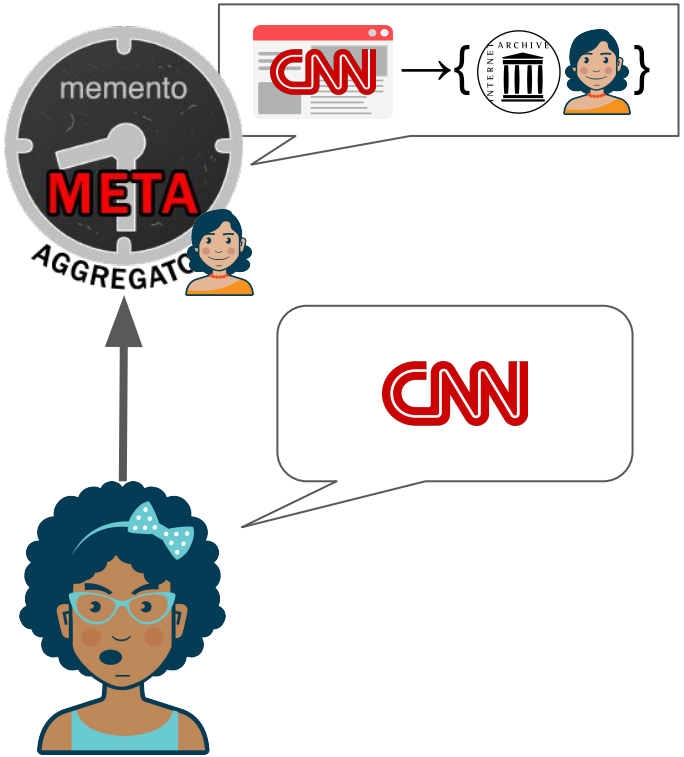


Personal Archive Aggregation

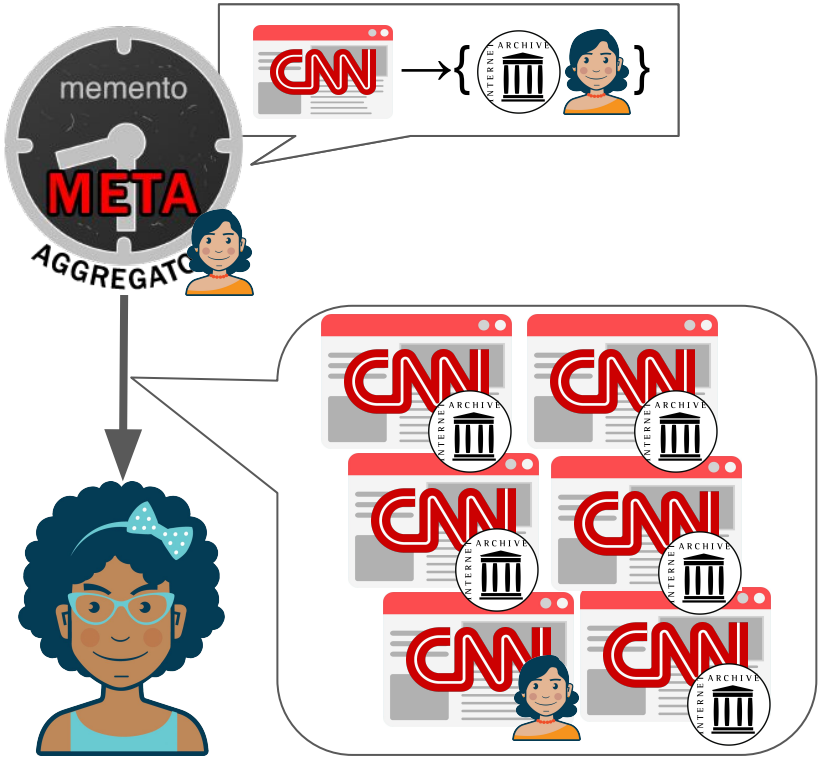
Alice Deploys MMA_A



Carol Asks MMA_A for CNN

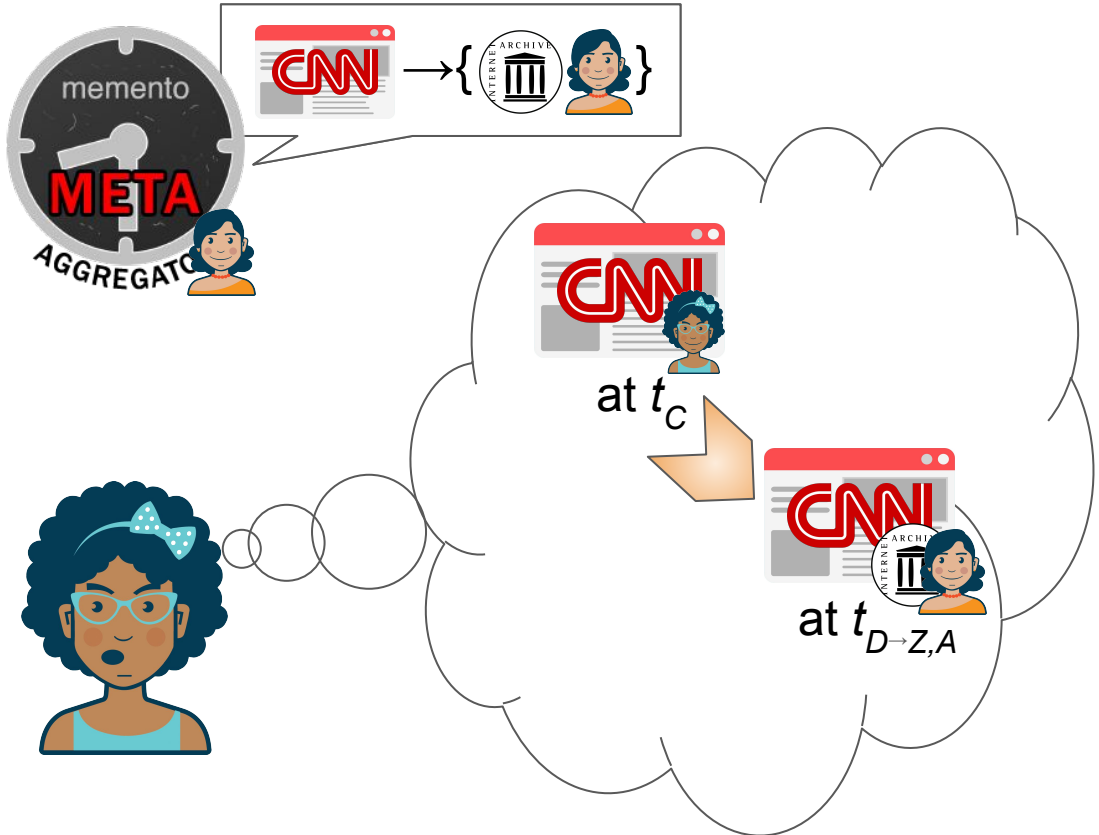


MMA_A returns CNN Memento $\{M_A, M_{IA}\}$

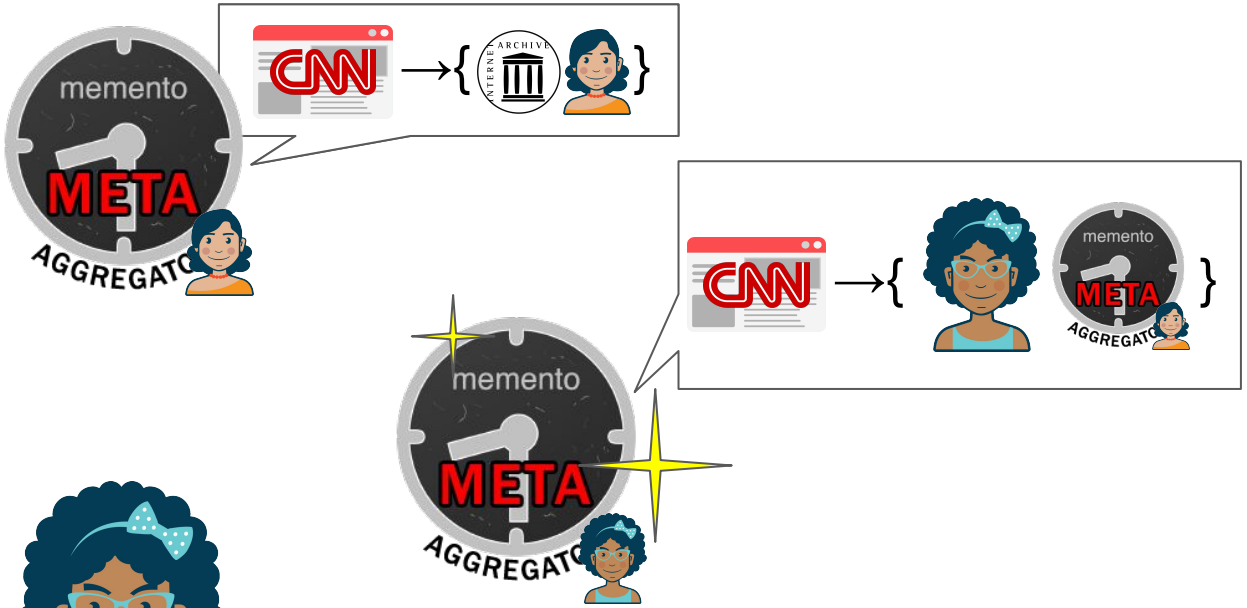


Carol Wants to Aggregate Her Own Captures

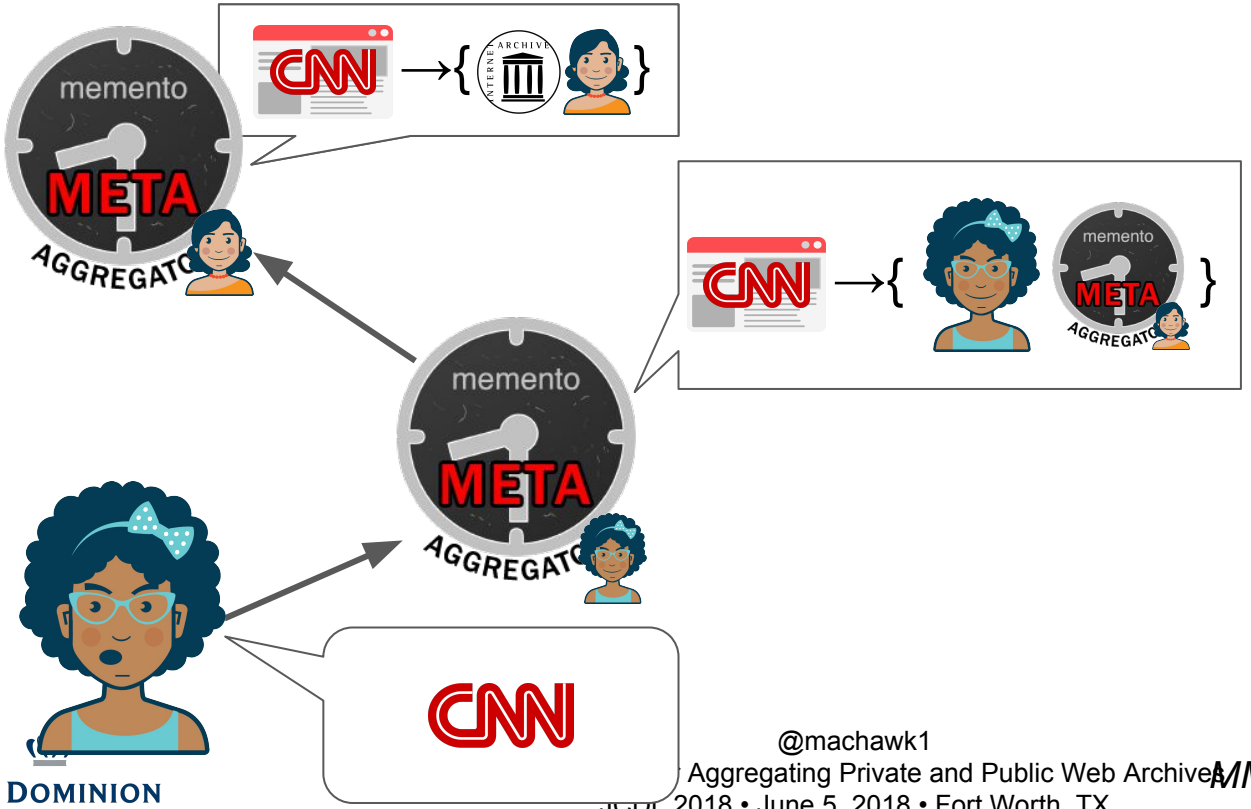
CNN(M(WA$_C$))



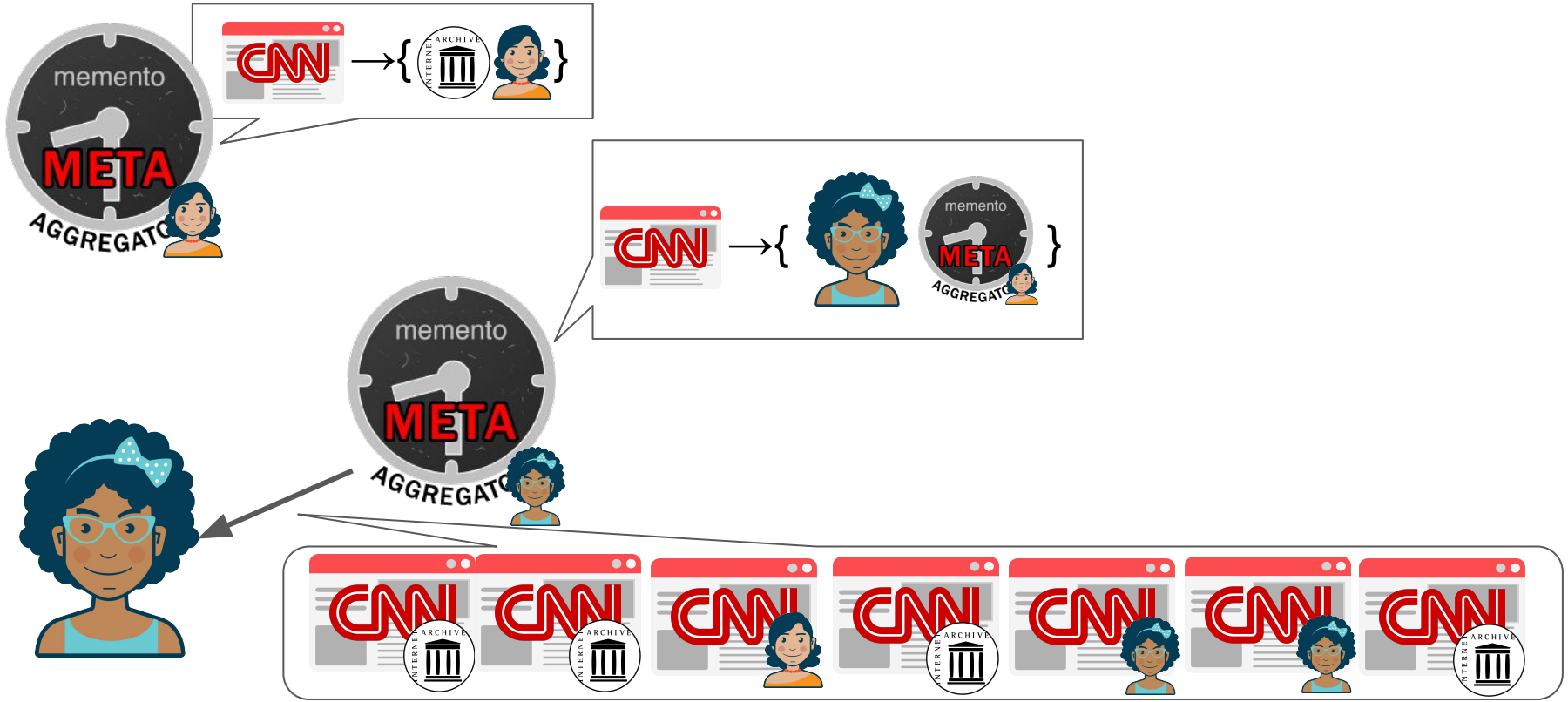
Carol Creates MMA_C to Access WA_C and MMA_A



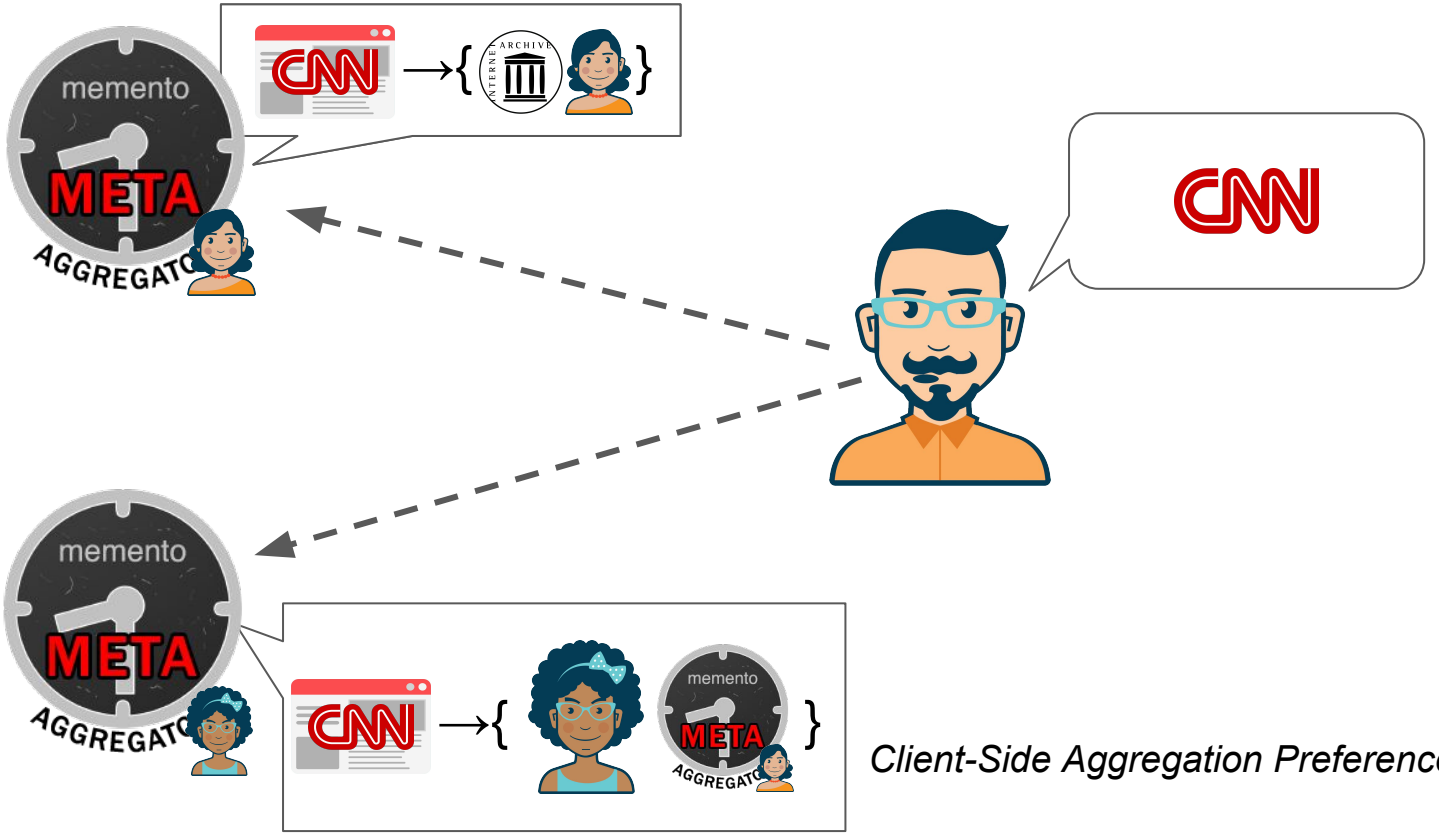
Carol Asks MMA_C For CNN



MMA_A returns CNN Memento $\{M_A, M_{IA}, M_C\}$



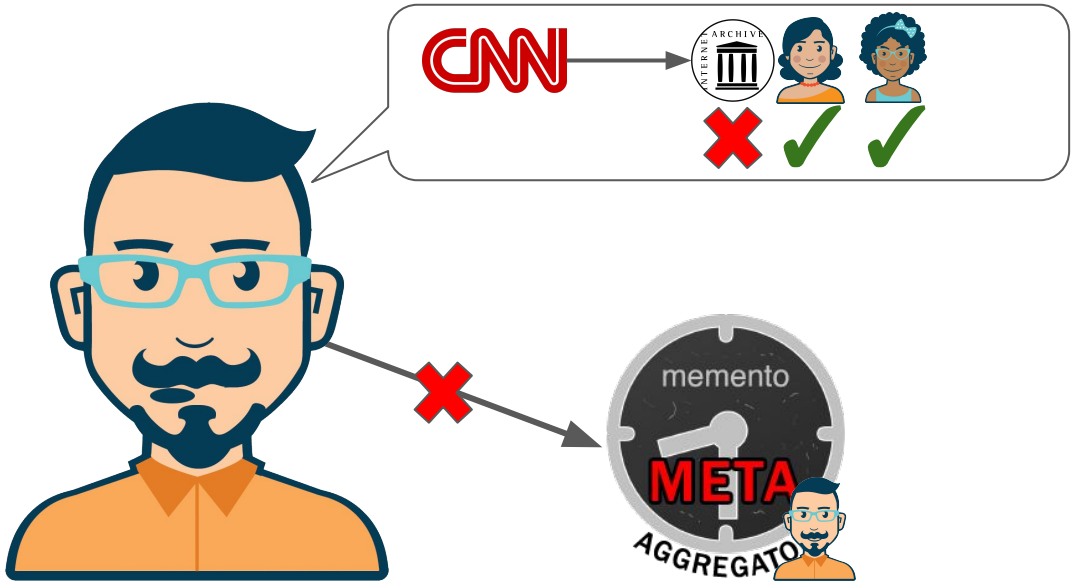
Bob May Request M(CNN) From MMA_A or MMA_C



Client-Side Aggregation Preference

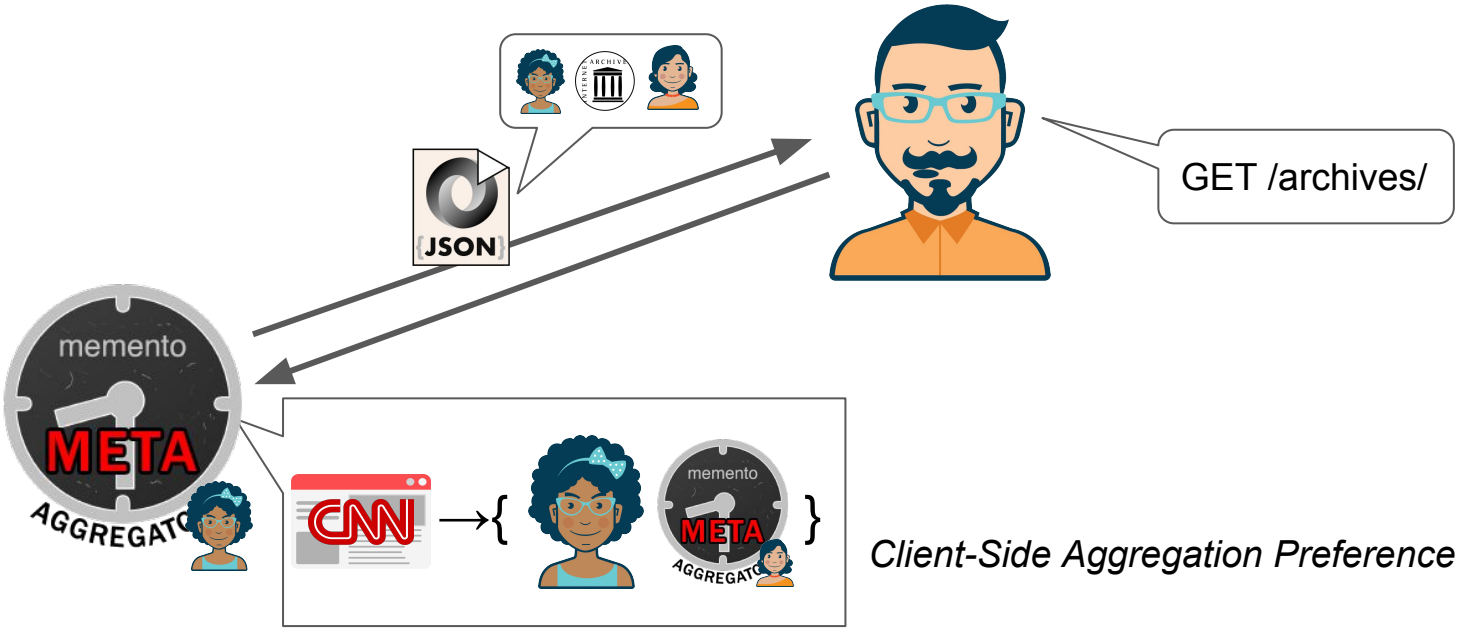
Bob Prefers to Exclude IA Captures

...and does not want to setup his own MMA



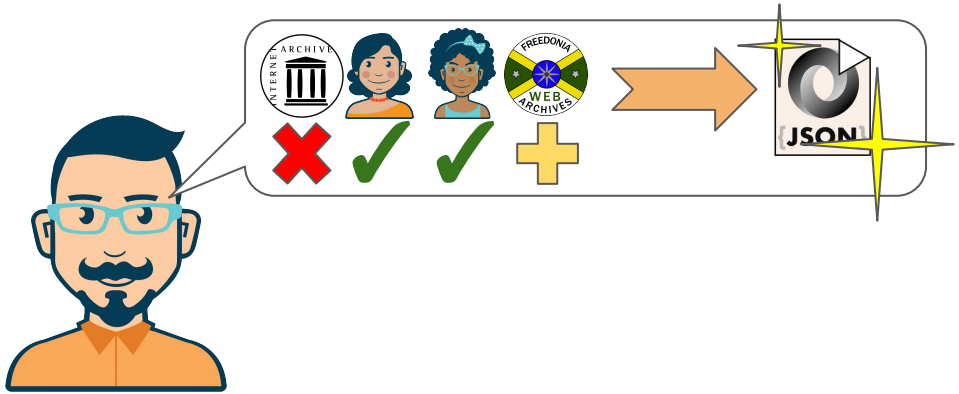
Client-Side Aggregation Preference

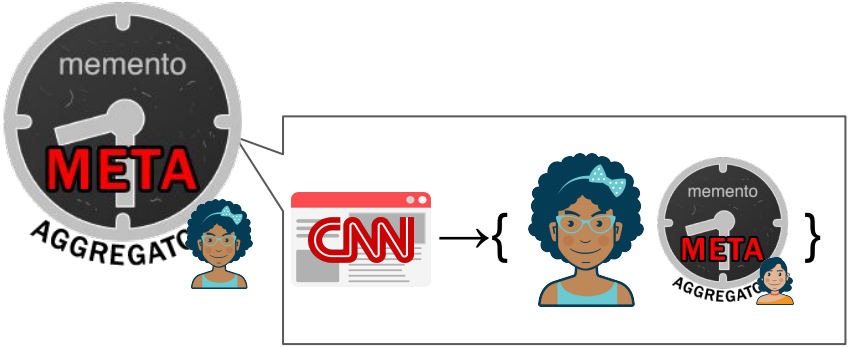
Bob Requests Supported Archives



Client-Side Aggregation Preference

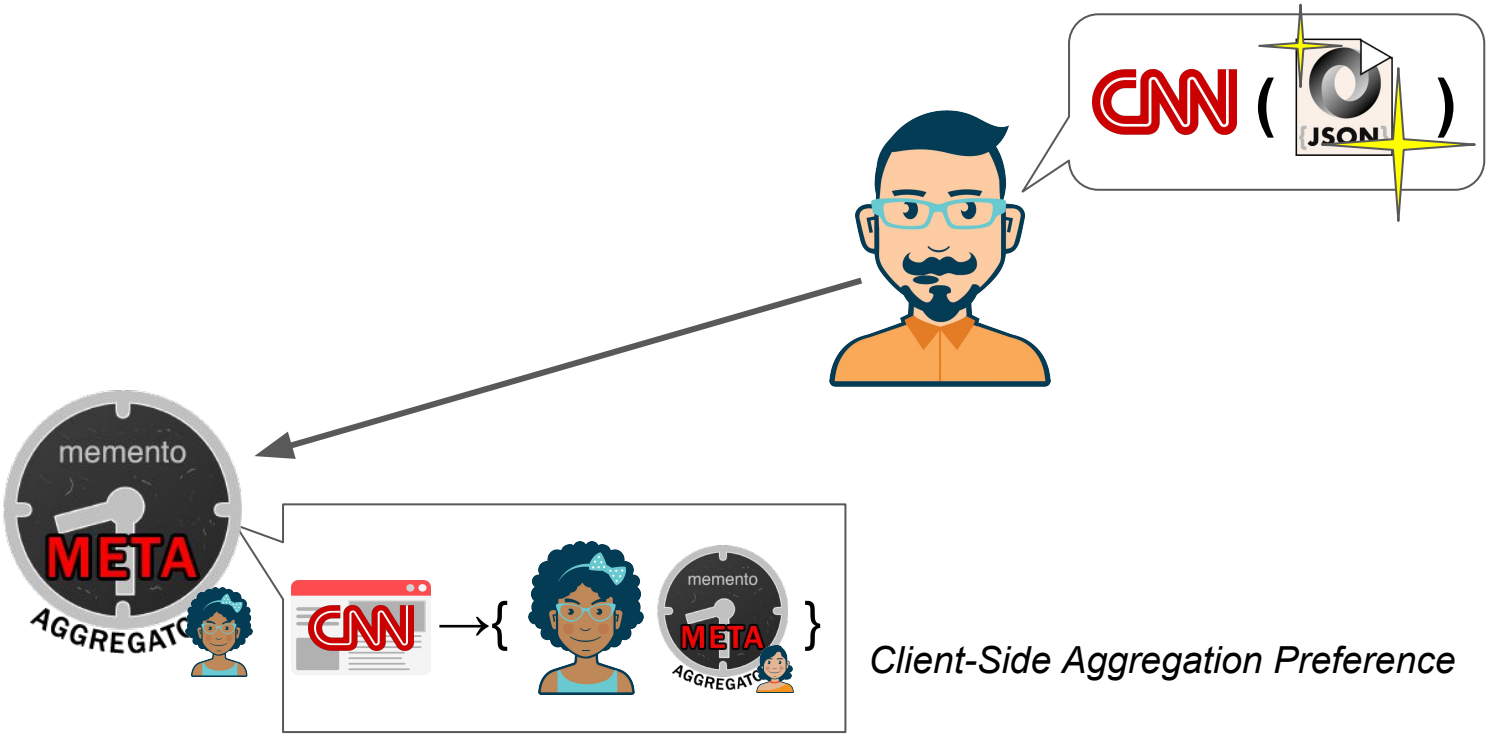
Bob Customizes the Set in the JSON





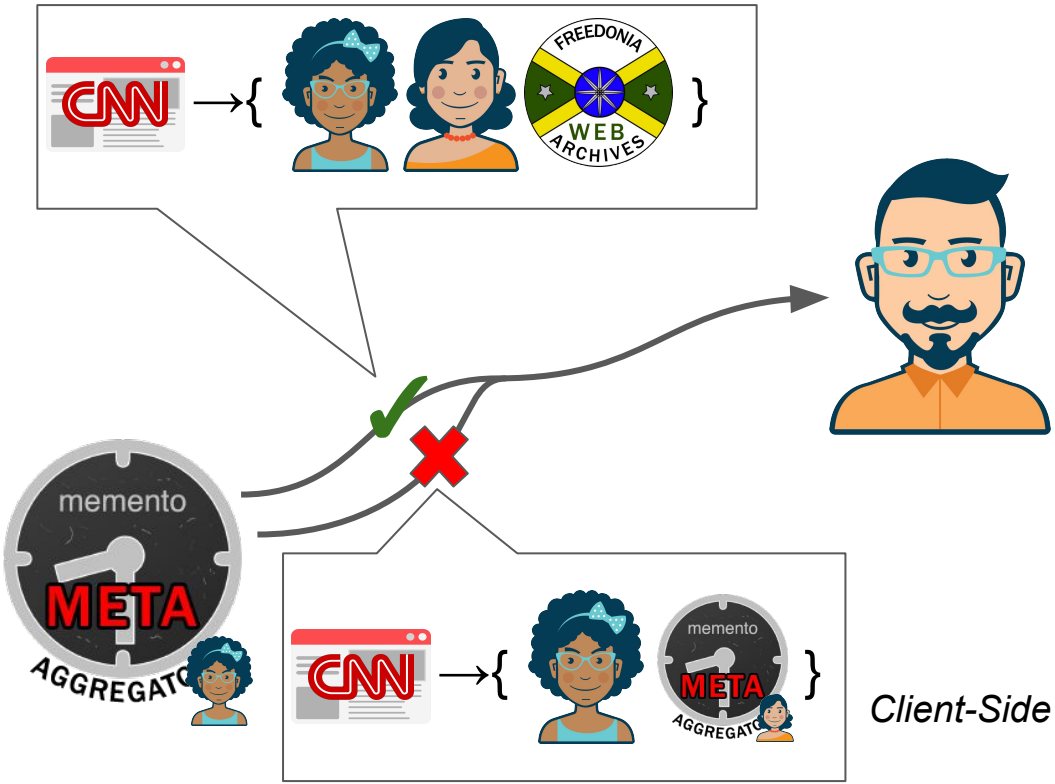
Client-Side Aggregation Preference

Bob Requests CNN for His Custom Set



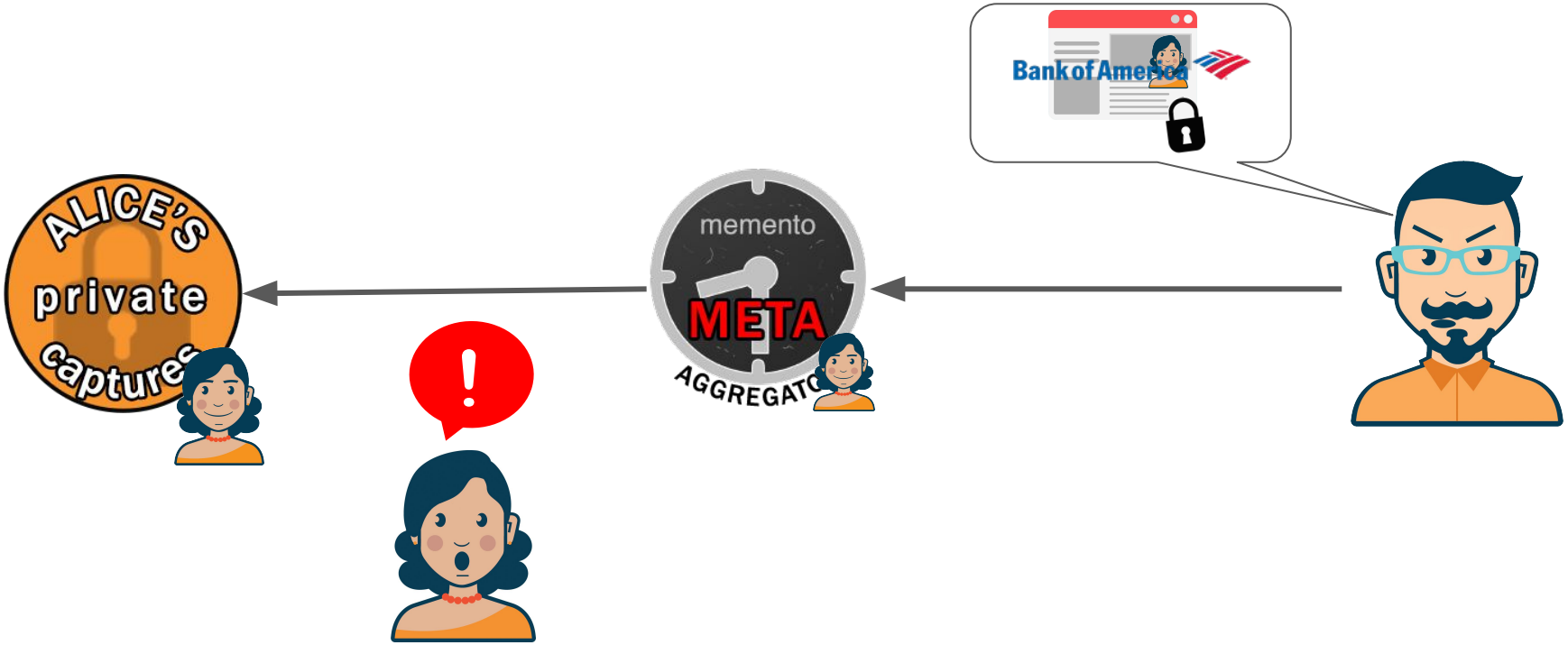
Client-Side Aggregation Preference

MMA Complies or Ignores Preference

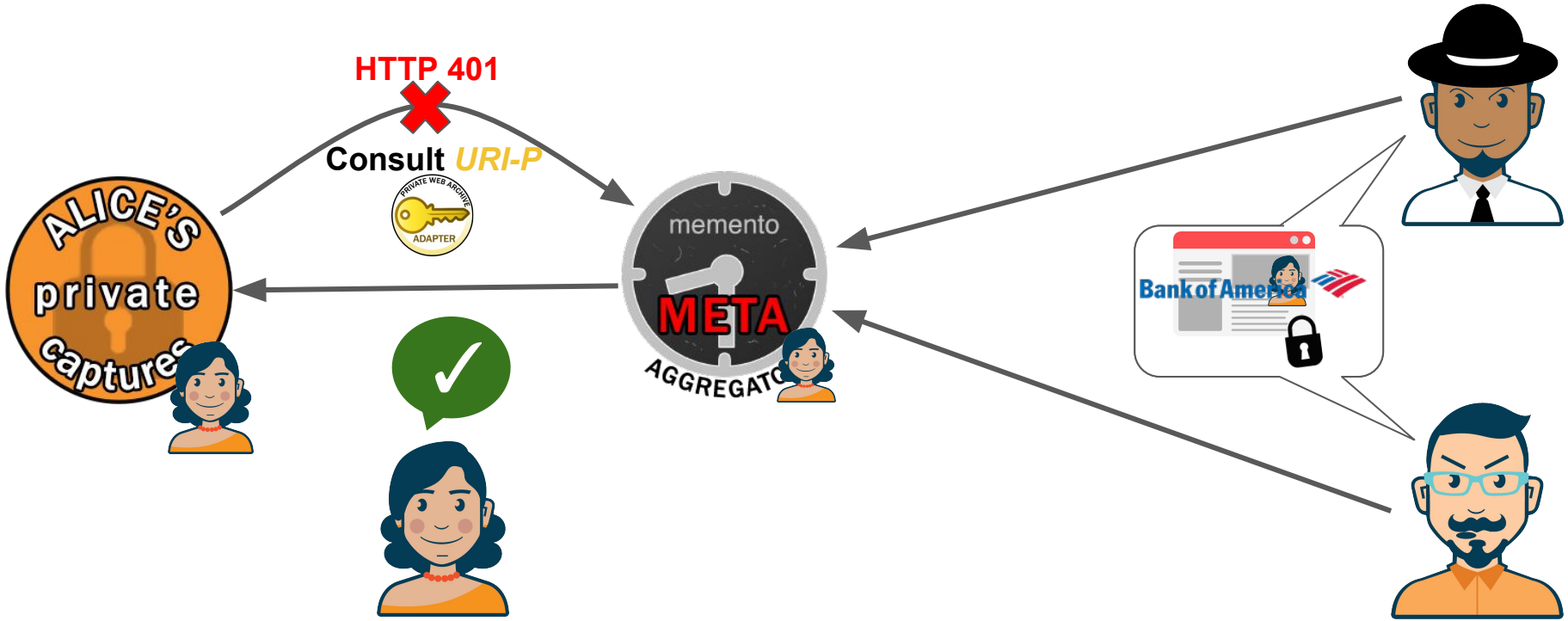


Client-Side Aggregation Preference

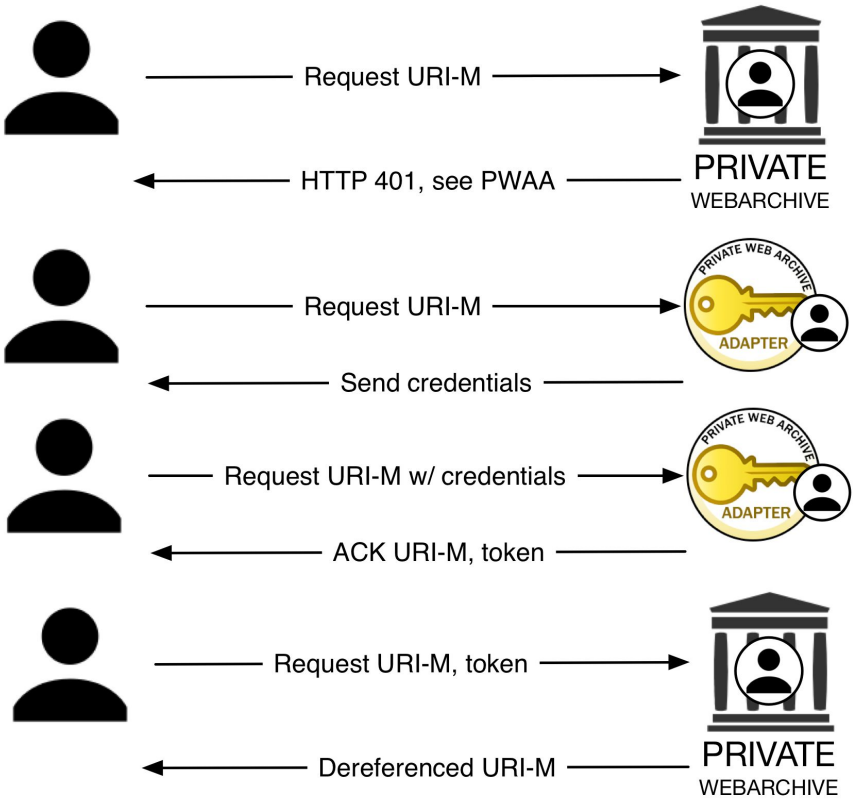
Hooray, Aggregation!



Hooray, Aggregation!

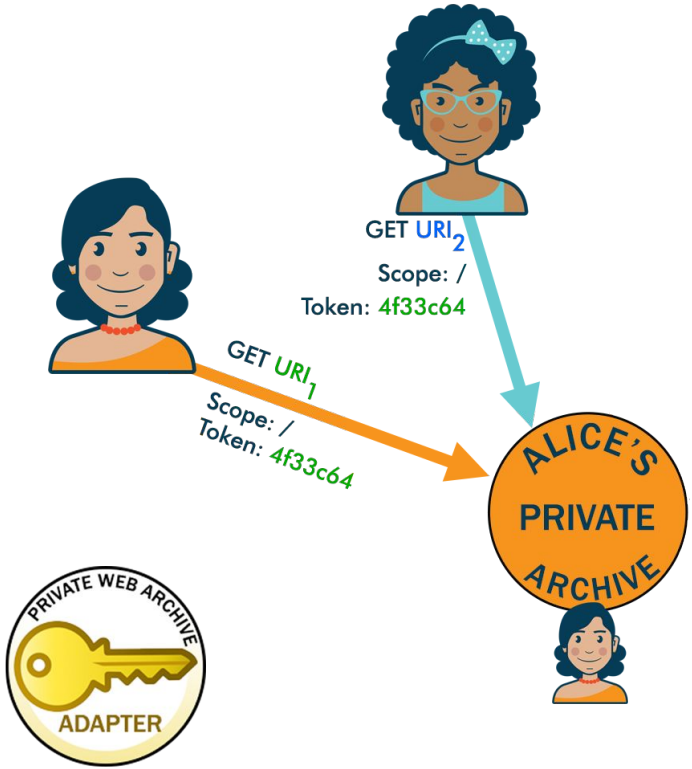


Private Web Archive Adapter (PWAA)

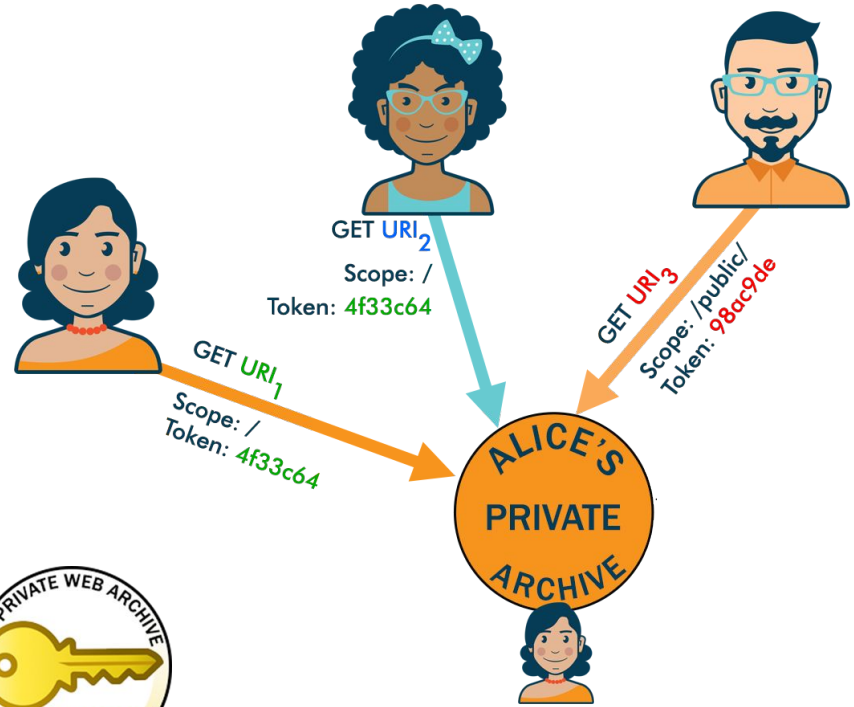


- Auth Layer for to encourage Private Web archive aggregation
- Typical OAuth 2.0 flow
- Auth role cohesive to PWAA
- Persistent access through tokenization

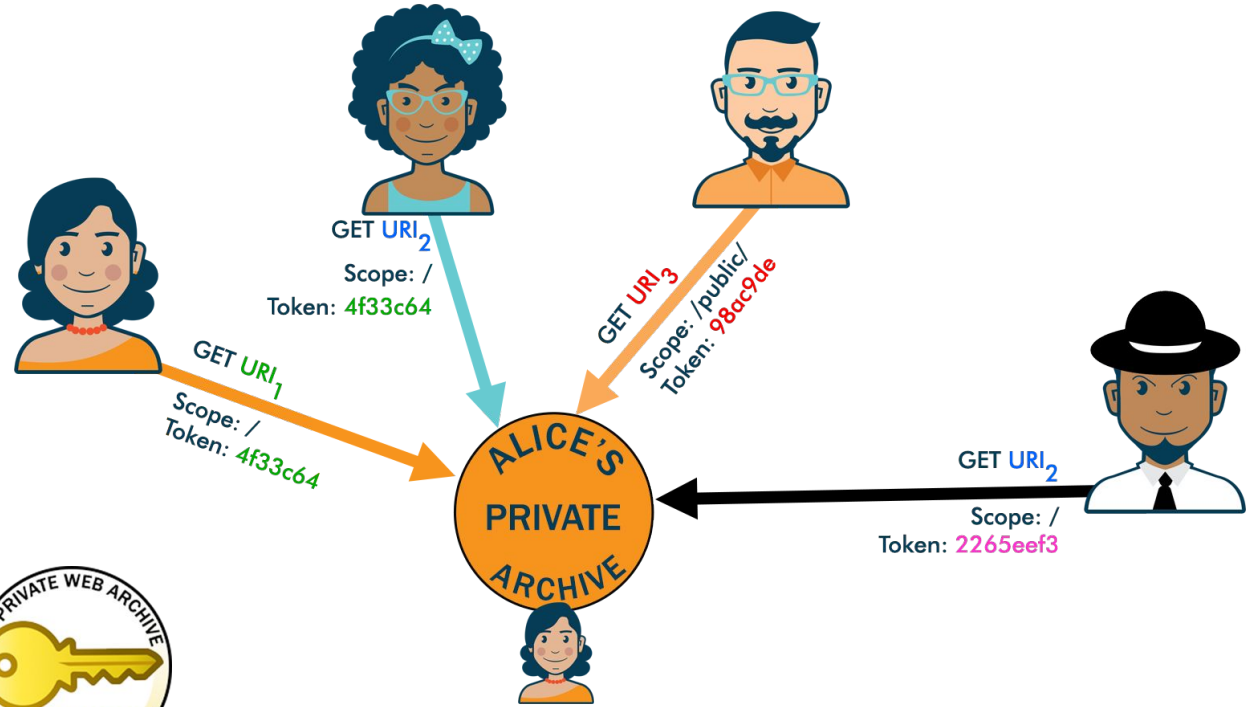
PWAA - Sharing Tokens



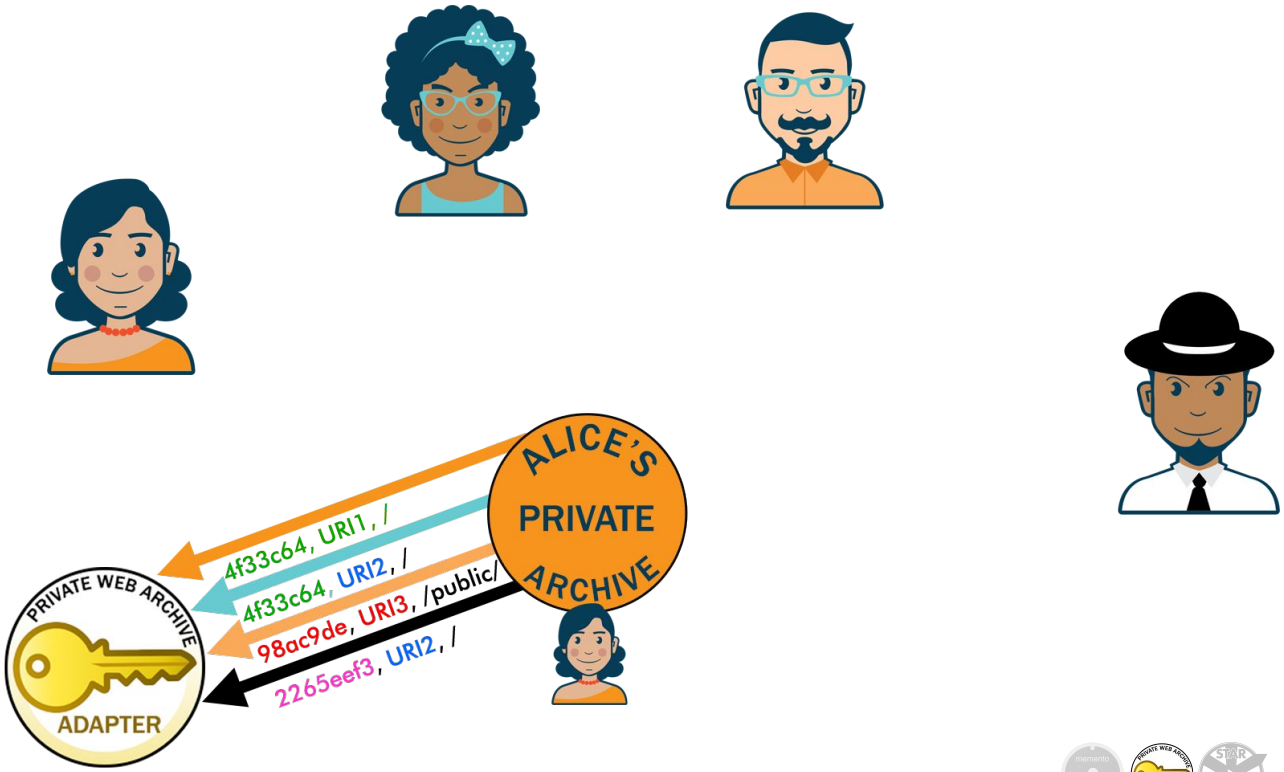
PWAA - Previously Authorized



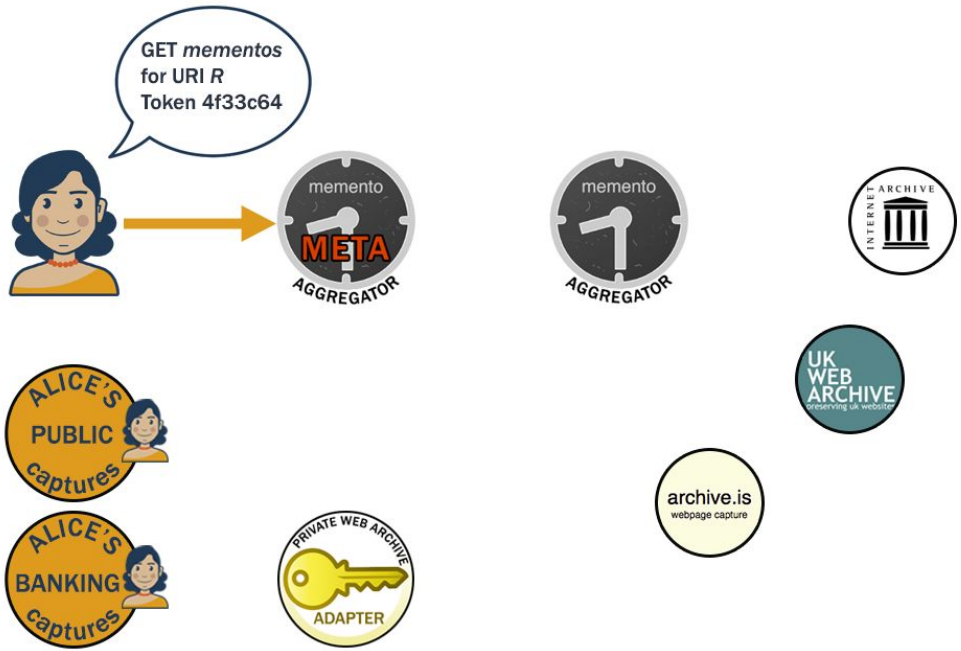
PWAA - Unauthorized Request



PWAA - Sharing Tokens

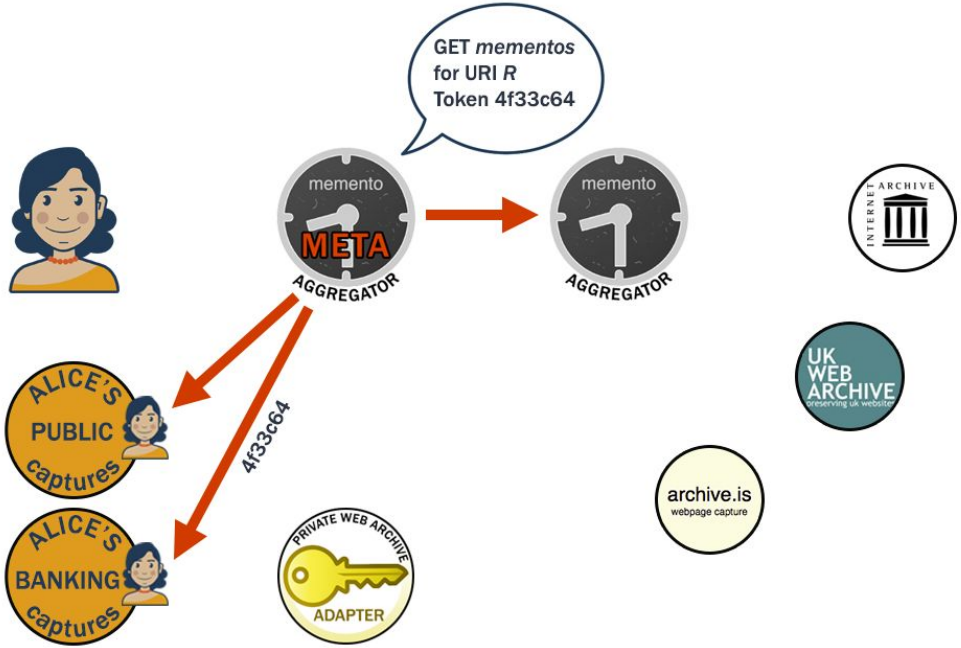


Alice Passes Associative Token to MMA

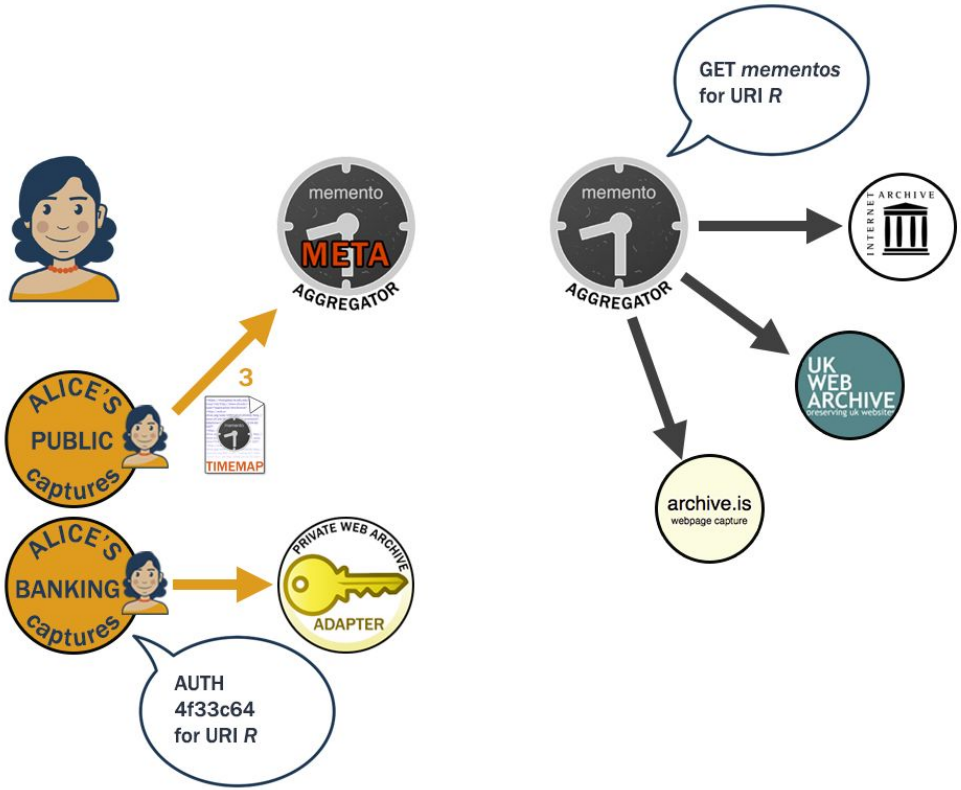


MMA requests URI-R...

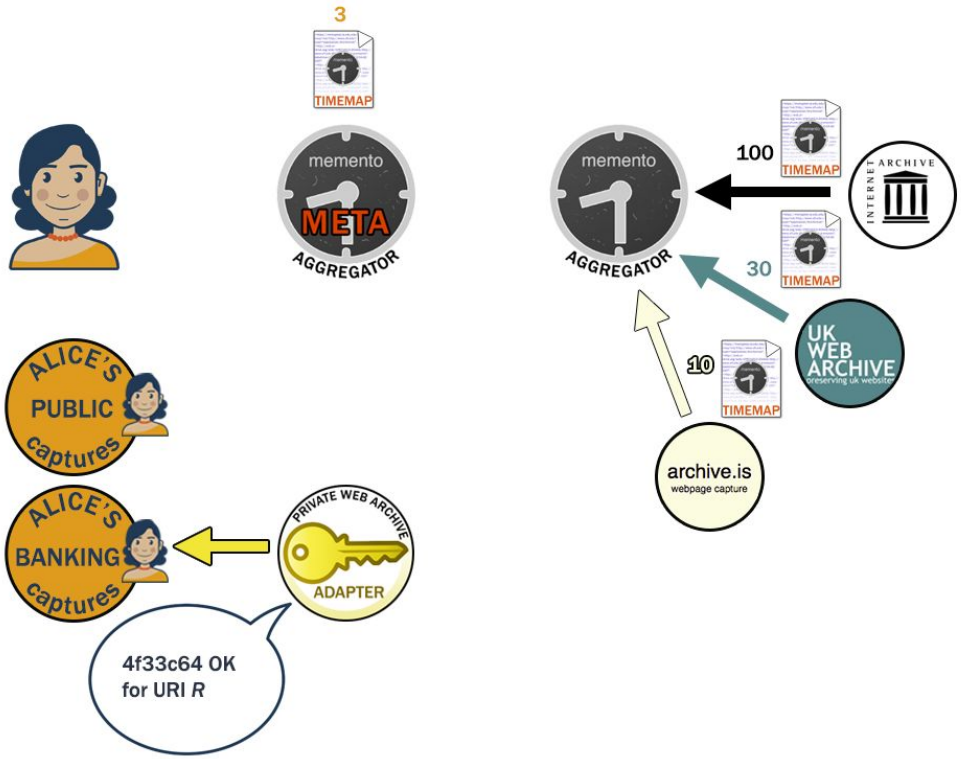
...relays token where applicable



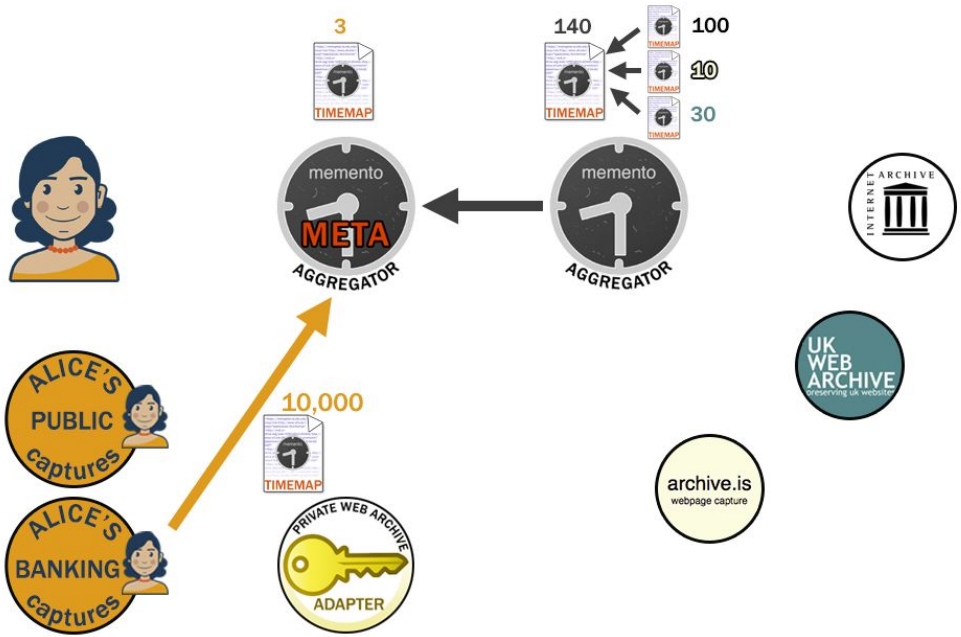
Private Archive Validates with PWAA



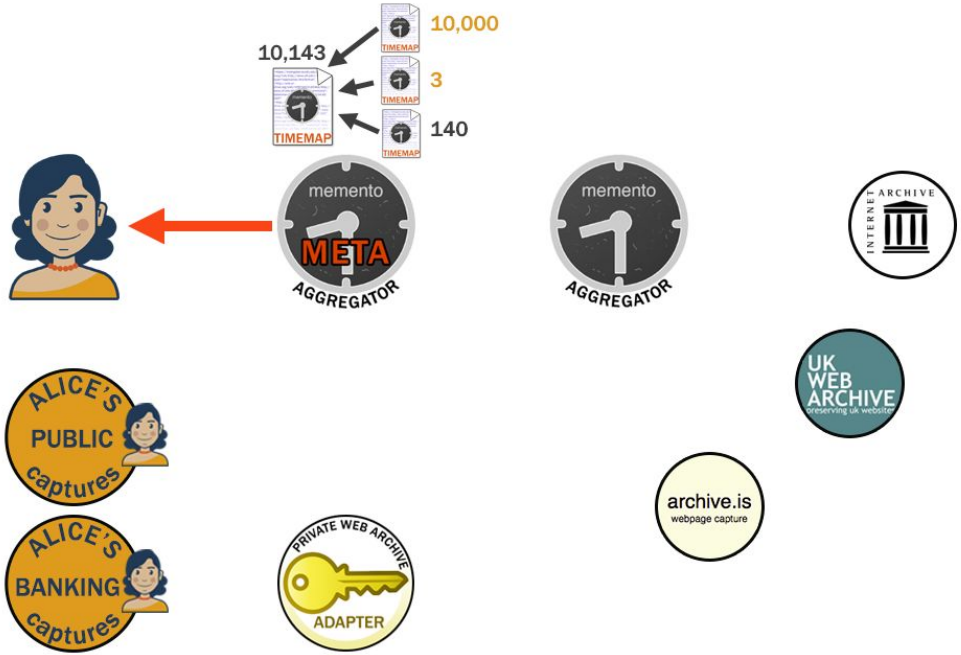
PWAA Confirms Token



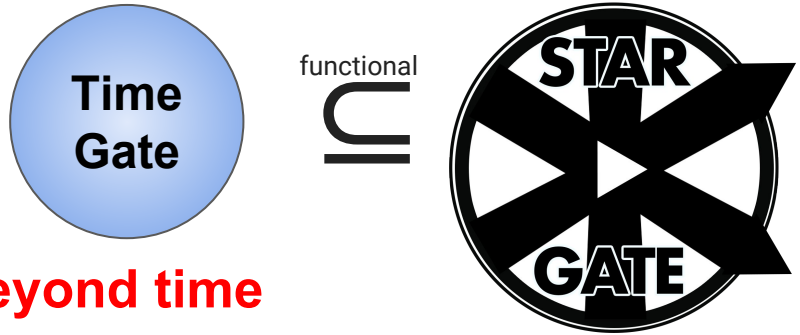
Private Archive Returns Captures



MMA Aggregates, Associates Token

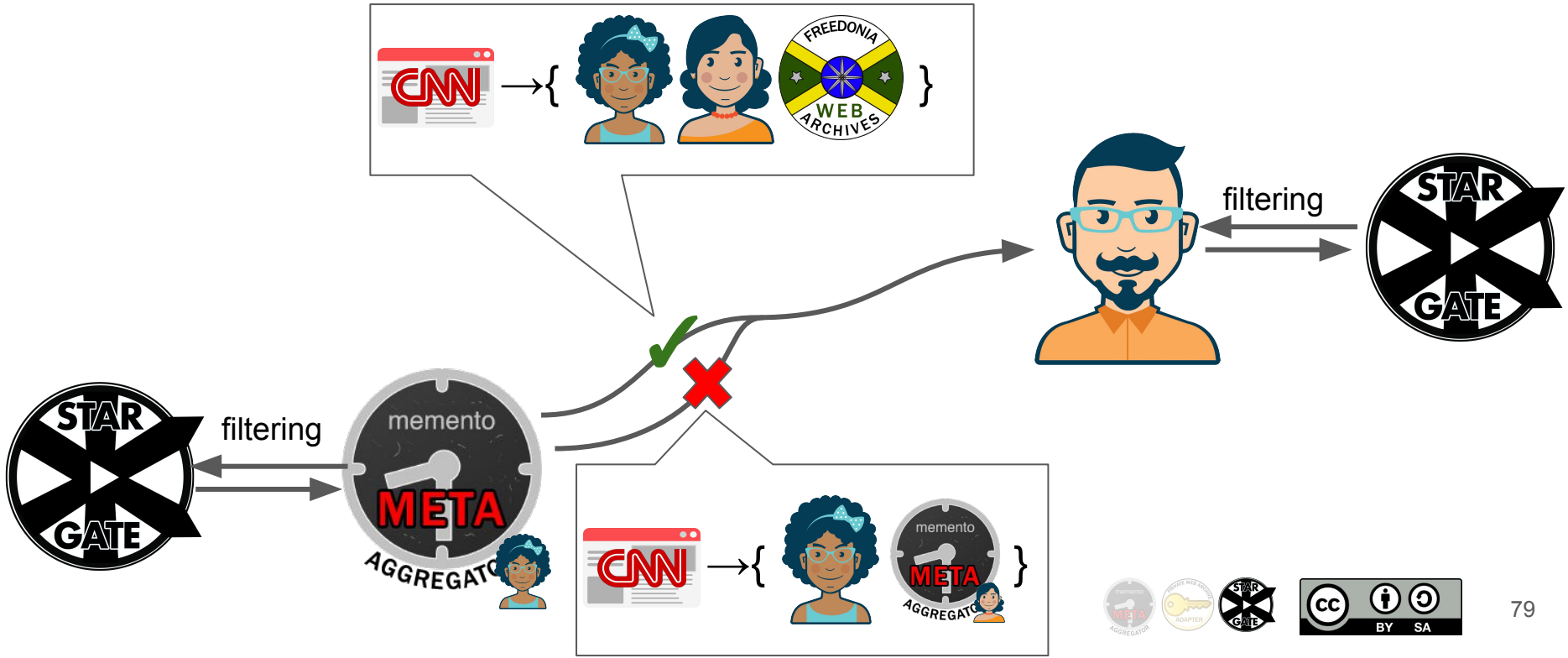


StarGate



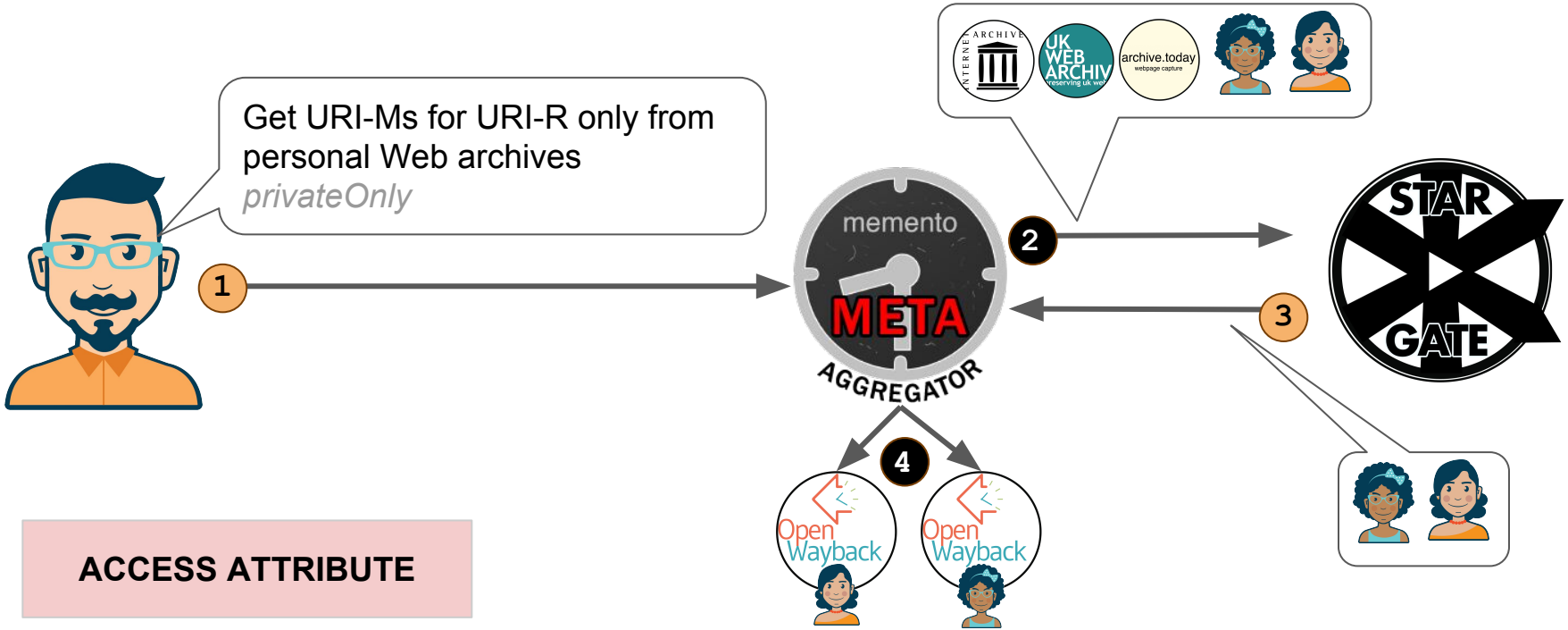
- Content negotiation in Web archives **beyond time**
- "Star" ~ wildcard (*) → **any** dimension of negotiation
- Allow for queries like: Only show me memento…
  - That are not redirects (*content-based attribute* HTTP Status ≠ 3XX)
  - Of a sufficient quality (*derived attribute* Memento Damage < 0.4)
  - Are from personal Web archives (*access attribute* indicate Facebook.com memento is not a login page)

Implicit Filtering via MMA or Directly (a la TG)



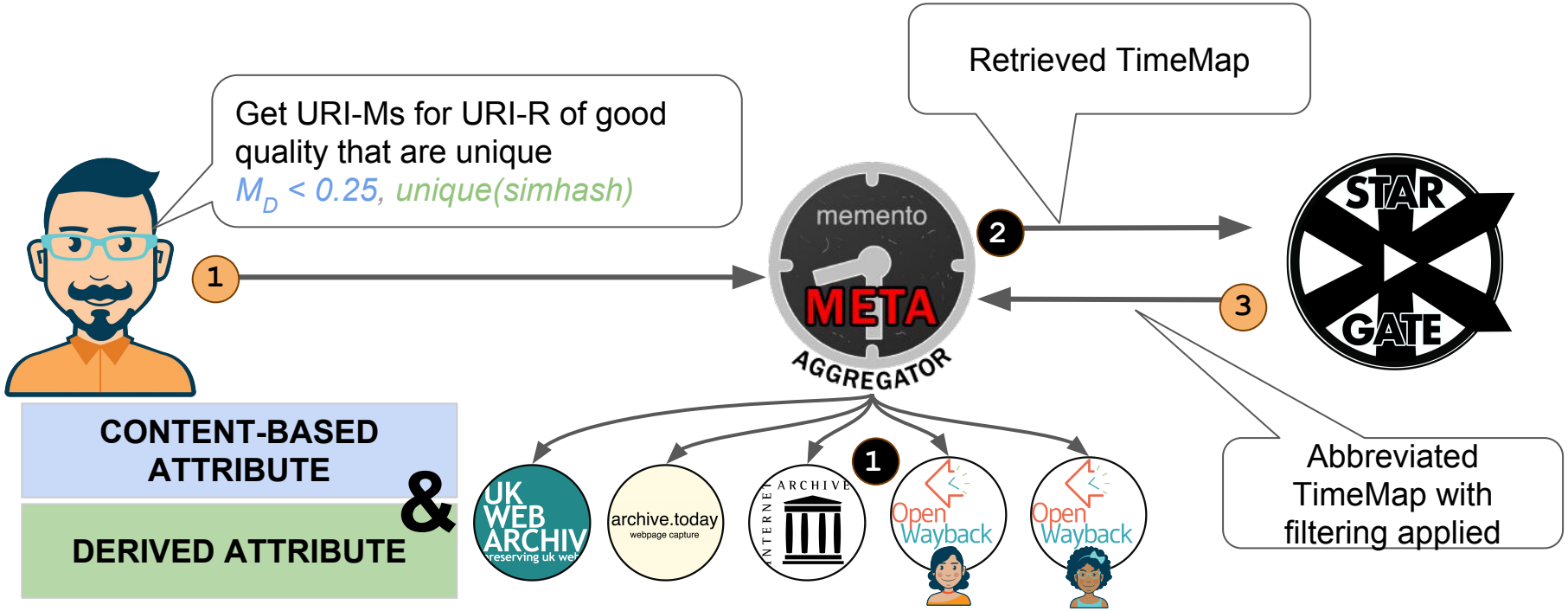
Negotiation in the Privacy Dimension

(via short circuiting)



Negotiation on Content-Based or Derived Attributes

(with response filtering)



Future Work and Conclusions

- **Aggregation with private web archives**
- **Client-side archive specification**
- ***TimeMap Caching Ramifications***

- **Authentication layer to systematically interface with private Web archives**
- ***Password-less approaches***

- **Archival negotiation in dimensions beyond time**
- ***Time/Space Complexity***
- ***Elegance of Expression***

Ongoing Research Supported By...

- NEH grant #HK-50181-14
- IMLS grant #RE-33-16-0107
- SIGIR Travel Grant

Some artwork based on Agata Krych
CC BY-SA 4.0, derivatives available at https://github.com/machawk1/jcdl2018-artwork

A Framework for Aggregating Private and Public Web Archives

Mat Kelly, Michael L. Nelson, and Michele C. Weigle
Old Dominion University
Web Science & Digital Libraries Research Group
{mkelly, mln, mweigle}@cs.odu.edu
@machawk1 • @WebSciDL

Joint Conference on Digital Libraries (JCDL)
June 5, 2018, Fort Worth, TX

OLD DOMINION UNIVERSITY

WSDL Old Dominion University

JCDL 2018 Ft. Worth, Texas

SIGIR Special Interest Group on Information Retrieval