Aggregator Reuse and Extension for Richer Web Archive Interaction

Mat Kelly
Drexel University
mkelly@drexel.edu
@machawk1

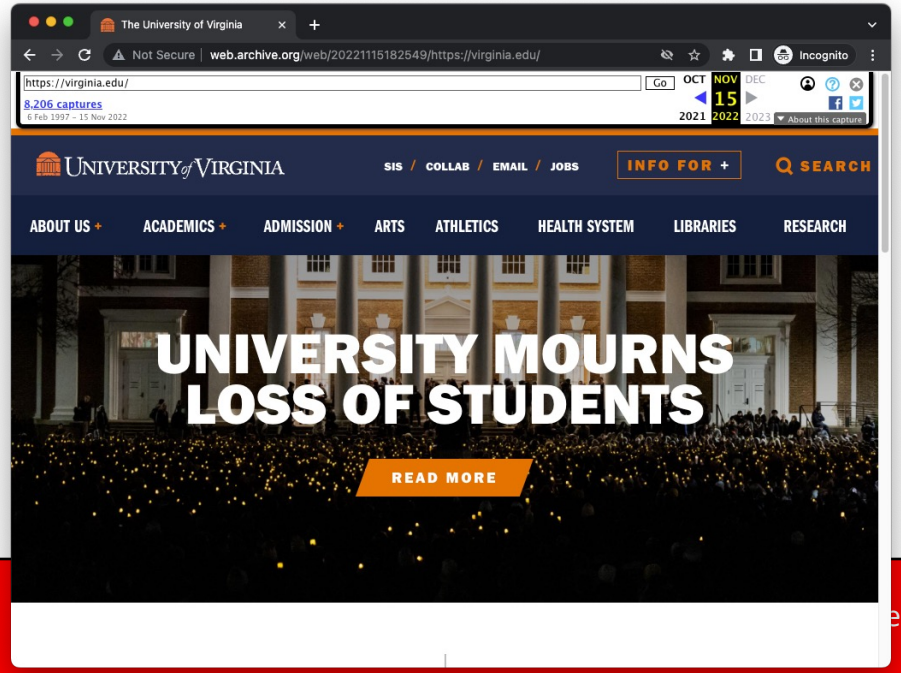
The 24th International Conference on Asia-Pacific Digital Libraries (ICADL 2022)
November 30 - December 2, 2022

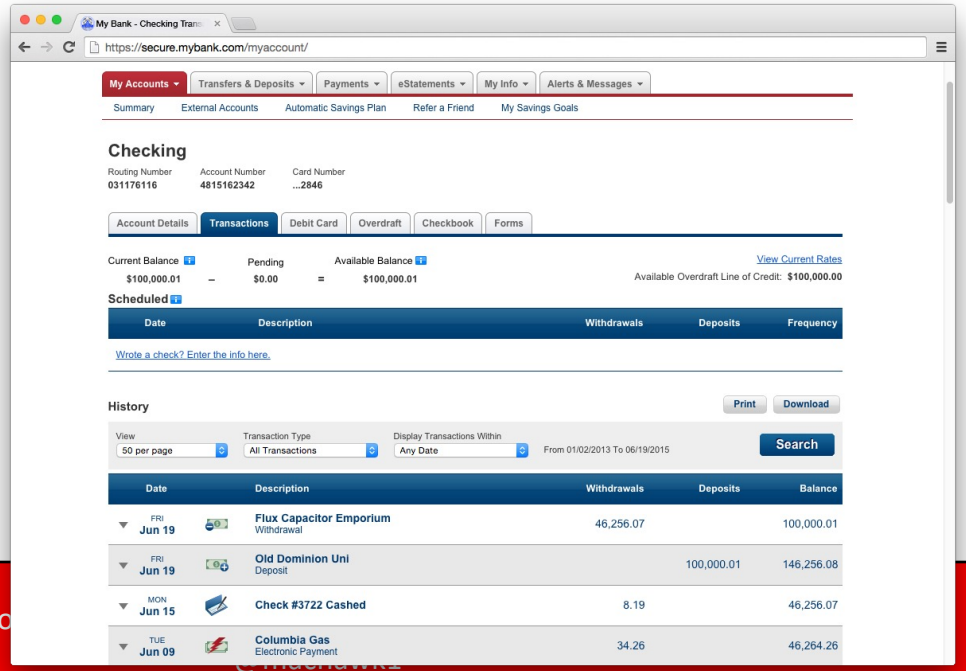
The Web as History

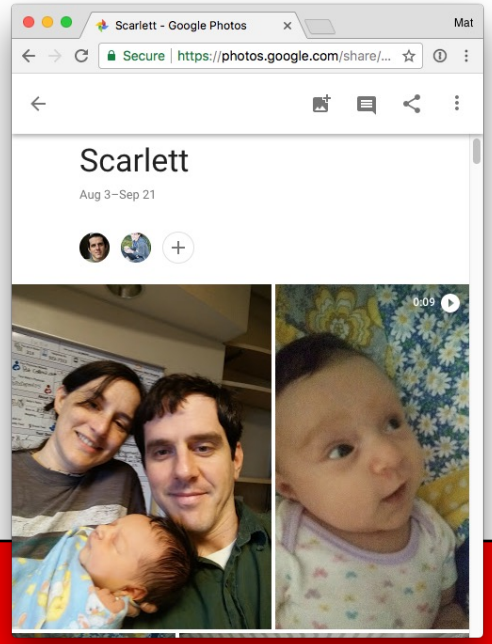
- Our history on the web
- The source(s) of archives dictate the narrative
- More sources → More comprehensive narrative

The Webs as History

- Long tail of web archives
- Not all {public,private} content is fit for {public,private} consumption
- The “key” of a URI may need expansion for additional contexts







Time on the Web(s)

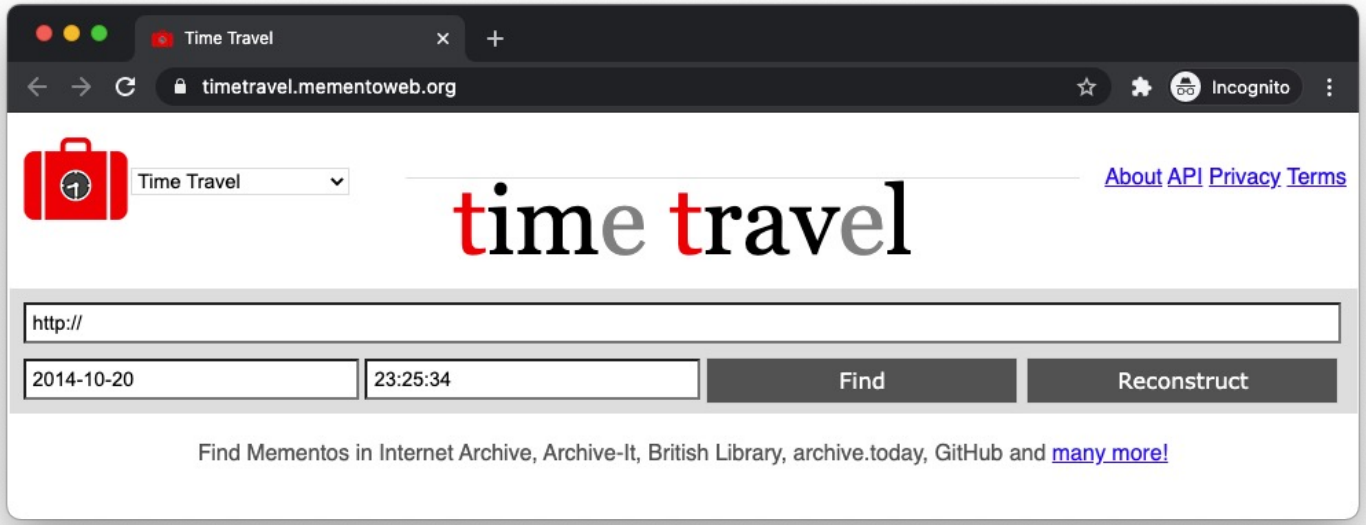
- Memento provides temporal negotiation, syntax, and semantics
- Memento aggregation allow for efficiently querying collective sources
- One endpoint solely for aggregation has additional functional potential

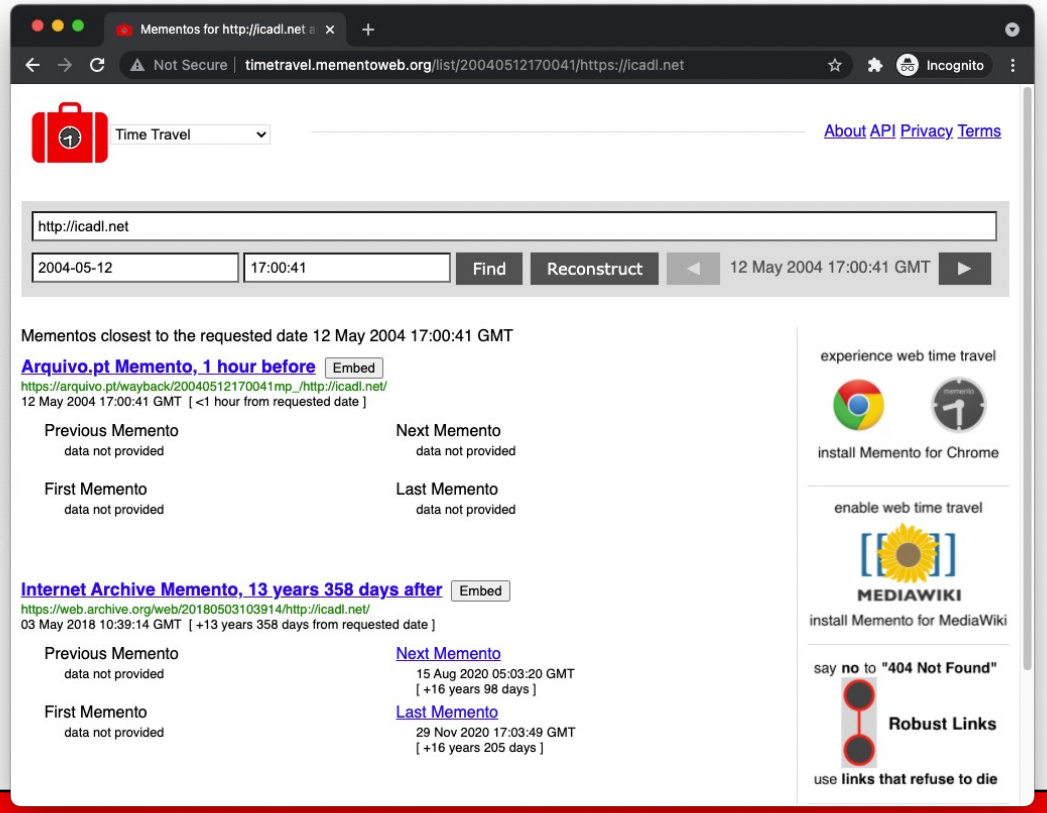
Questions to Consider

- RQ1: In what was can we leverage Memento aggregators for more sophisticated querying of web archives?
- RQ2: With newfound hierarchical use cases, what problems should we anticipate and how can they be mitigated?

Related Work – Time Travel

- De facto official Memento aggregation





https://mementoweb.org/

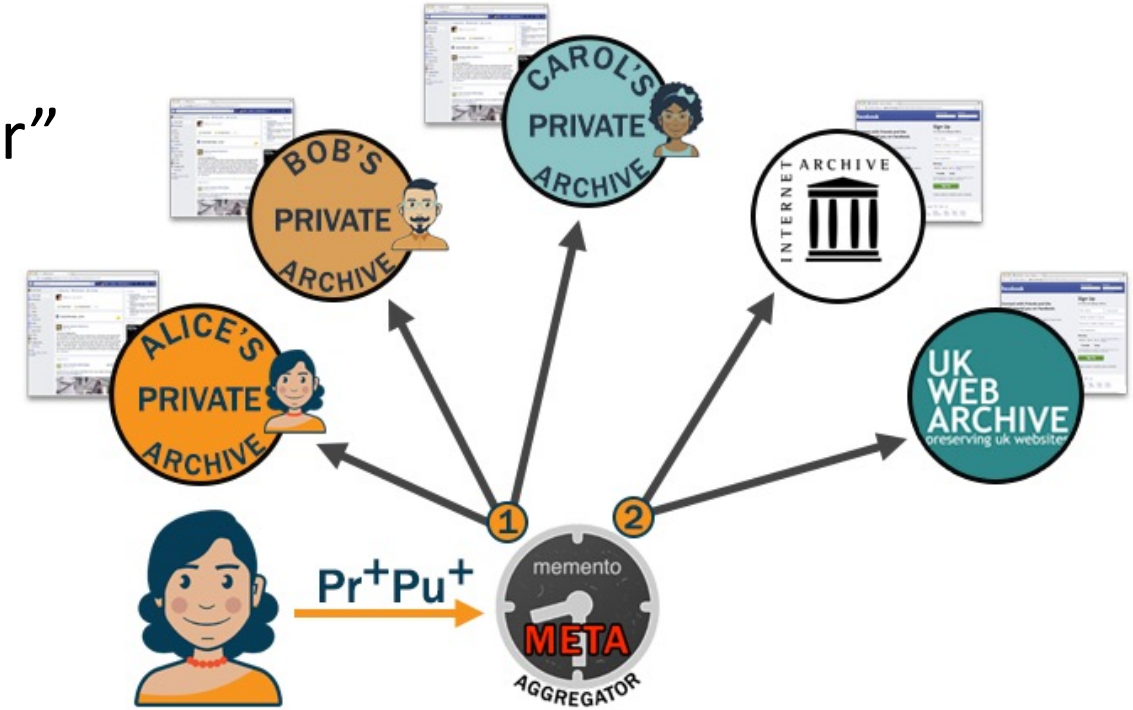
Related Work – MemGator

- Open source Memento aggregator
- Ripe for:
  - Distribution, Decentralization
  - Personal/Private use
  - Modification & Adaptation
- Used extensively in web archive research, less so an introspecting the aggregation process and component

Alam & Nelson. MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go. *JCDL 2016*, pp. 243-244

Related Work – Mementity Framework

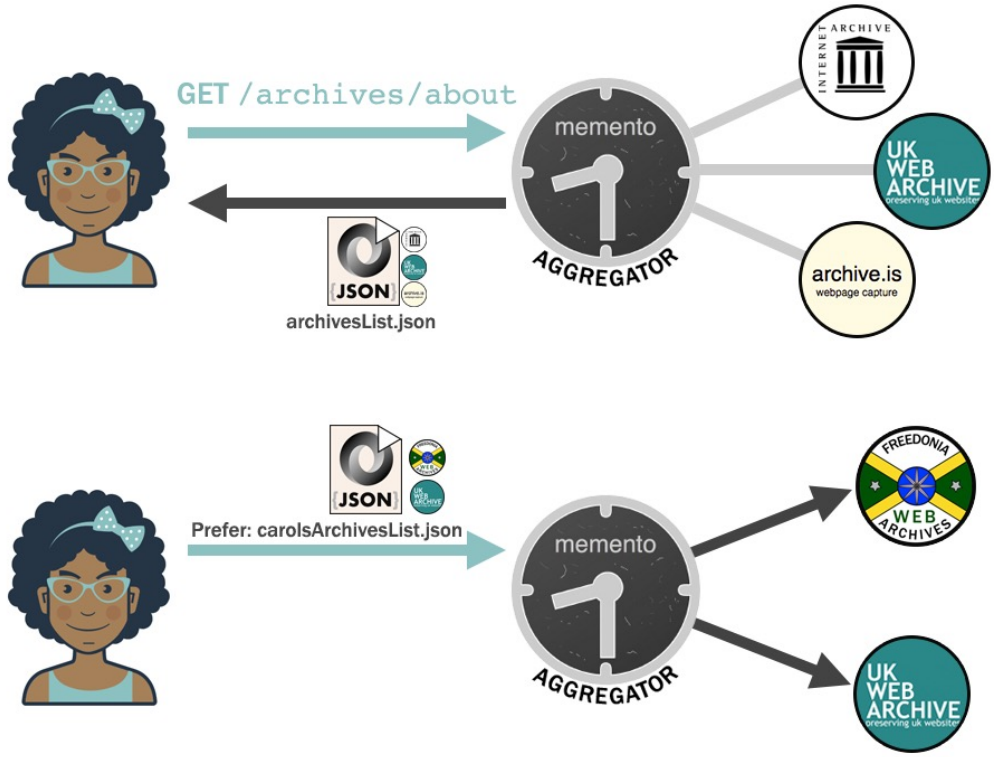
- Aggregators as means of combining captures from both private & public web archives
- Introduced “Memento Meta Aggregator”
- Outstanding issues on MMA



Kelly, Nelson, and Weigle, “A Framework for Aggregating Private and Public Web Archives,” In Proc. of JCDL 2018, pp. 273-282.

Related Work – Prefer

- HTTP Prefer for expression & consumption of clients' preferences
- Prior work re: web archiving deals w/ custom set of archives to-be-aggregated



Snell. 2014. Prefer Header for HTTP. IETF RFC 7240. (June 2014).

Kelly, Alam, Nelson, and Weigle, "Client-Assisted Memento Aggregation Using the Prefer Header," Presented at the Web Archiving and Digital Libraries Workshop (WADL 2018)

URI Templating

```
t0: {scheme & hostname}/{resource type}/{format}/{URI-R}
t1: https://myarchive.org/timemap/link/http://example.com
m0: {scheme & hostname}/{datetime}/{URI-R}
m1: http://archive.md/20210619183508/https://icadl.net/icadl2021/
m2: https://archive.ph/eoQRZ
```

- t_0: Templated URL of a capture of the past web (memento)
  - Replicated with t_1
- m_0: A URI-M template
  - m_1: datetime and URI-R apparent
  - m_2: URI-R need not be apparent, available in HTTP Link header (RFC)

Base Querying Models

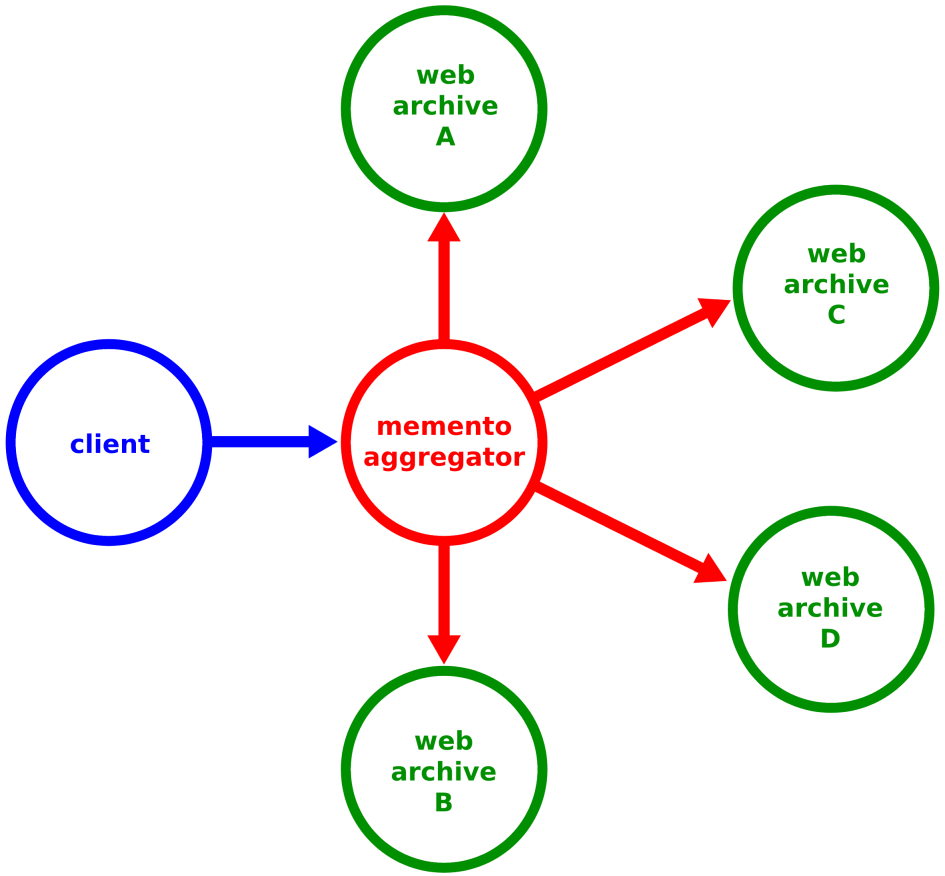
- Proxy-style querying
- Conventional querying
- Aggregator Chaining

- Client relays request to a single archive
  - Base case for client-aggregator interaction

Base Querying Models

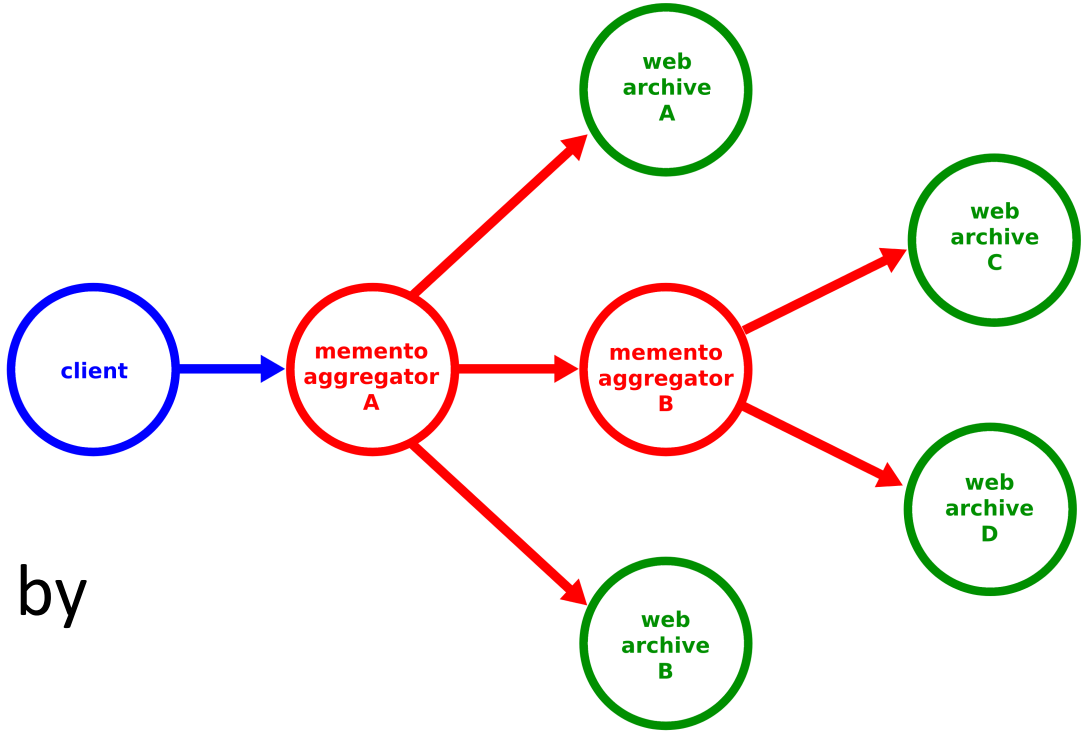
- Proxy-style querying
- Conventional querying
- Aggregator Chaining

- Client requests TimeMap containing all URI-Ms for a provided URI-R



Base Querying Models

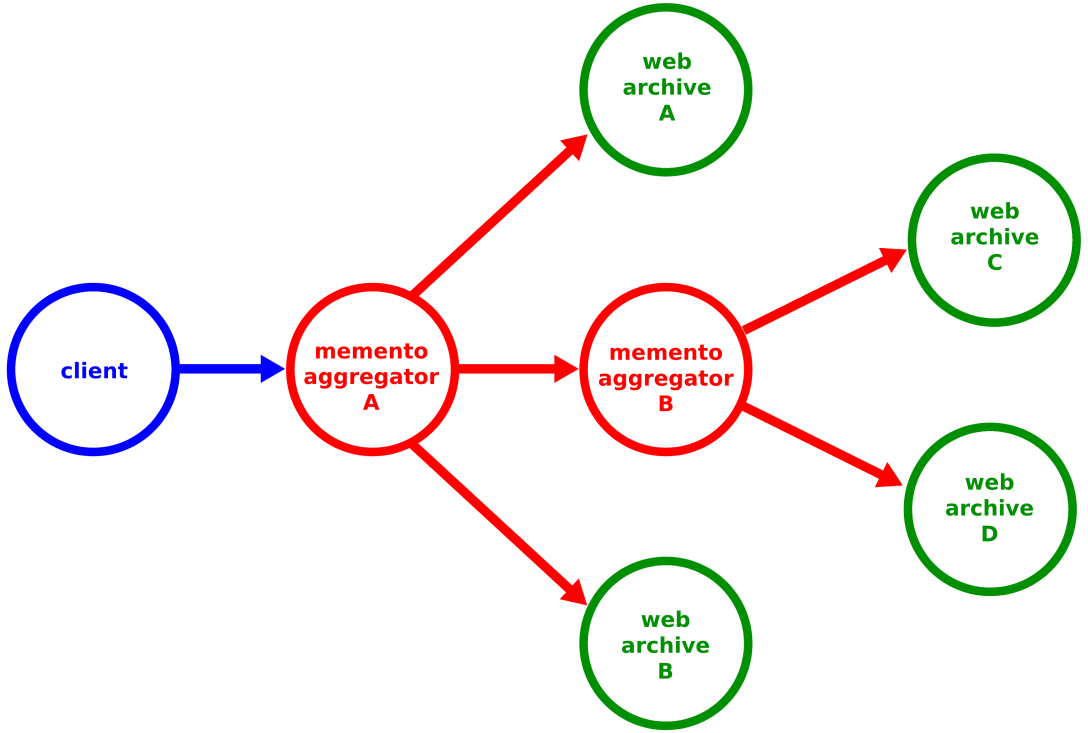
- Proxy-style querying
- Conventional querying
- Aggregator Chaining



- Allows for scoped archival knowledge by each aggregator
- Decentralized, facilitates inclusion of more sources w/o human intervention
- Useful, but potentially problematic…

Problematic Scenarios

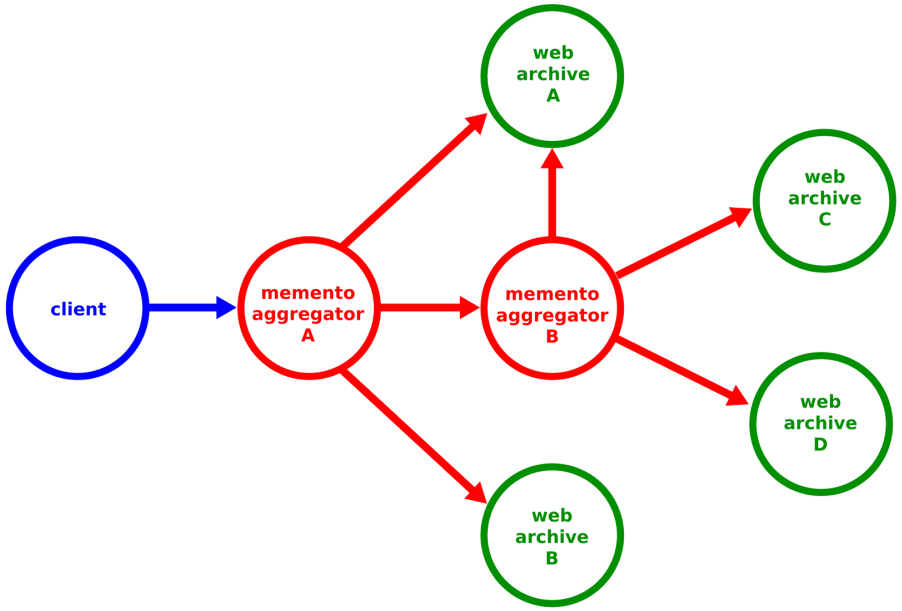
- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



FIX IMAGE

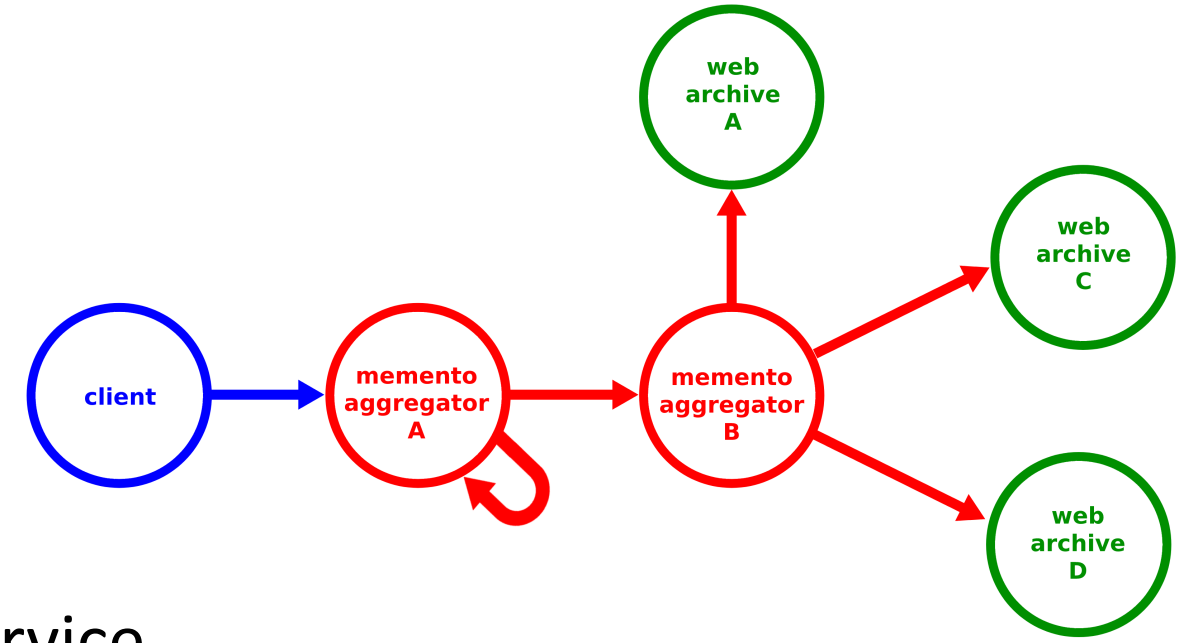
Problematic Scenarios

- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



Problematic Scenarios

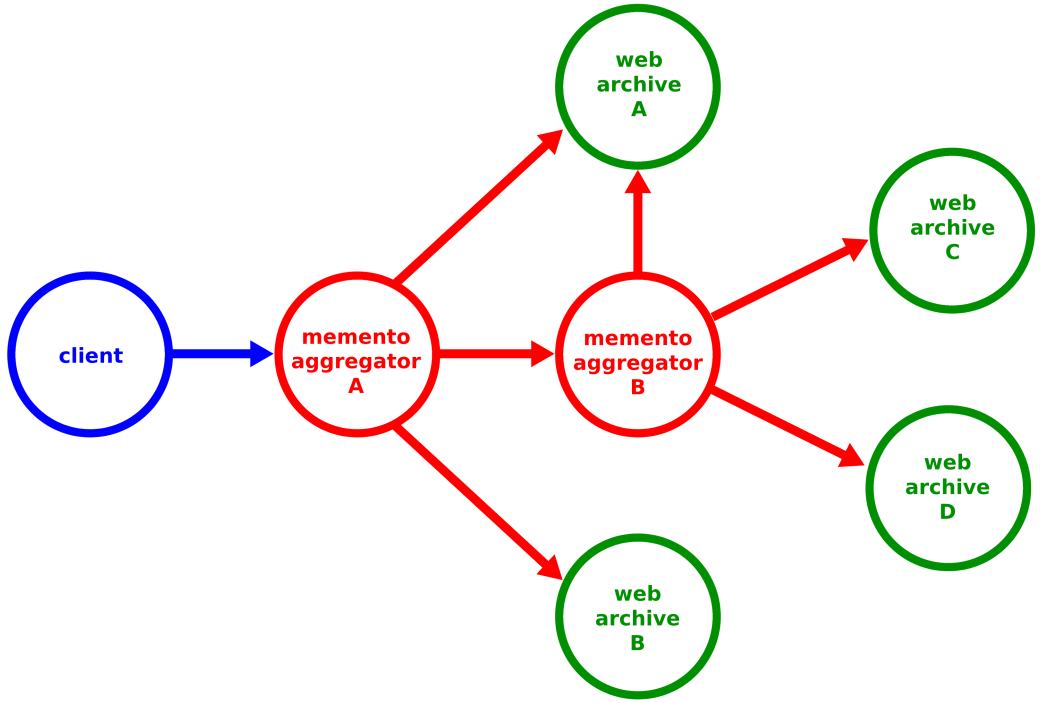
- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



- Reverse proxies may not expose service's own access points to service

Problematic Scenarios

- When a Tree Becomes a Graph
- Self-reference
- Duplication of Sources



Newfound Capabilities

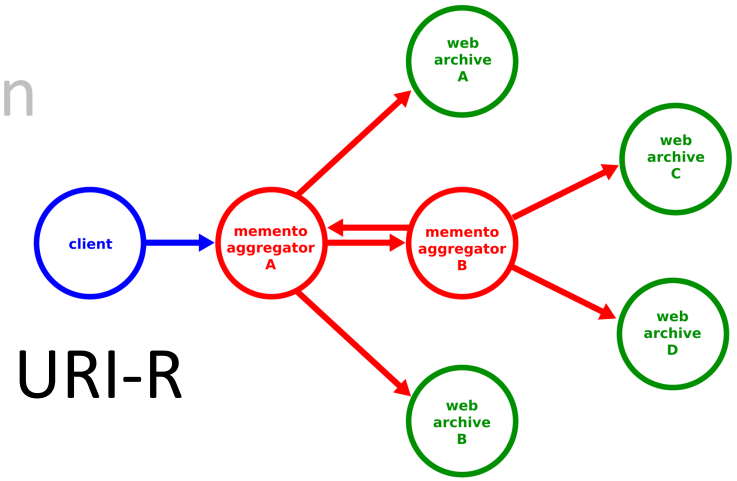
- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming

Newfound Capabilities

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming

Newfound Capabilities

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming



Nonce passed as query parameter independent of URI-R

- Allows short-circuiting
- Provide requestor indication that requestee was requestor in chain
- Potentially violates HTTP's statelessness.

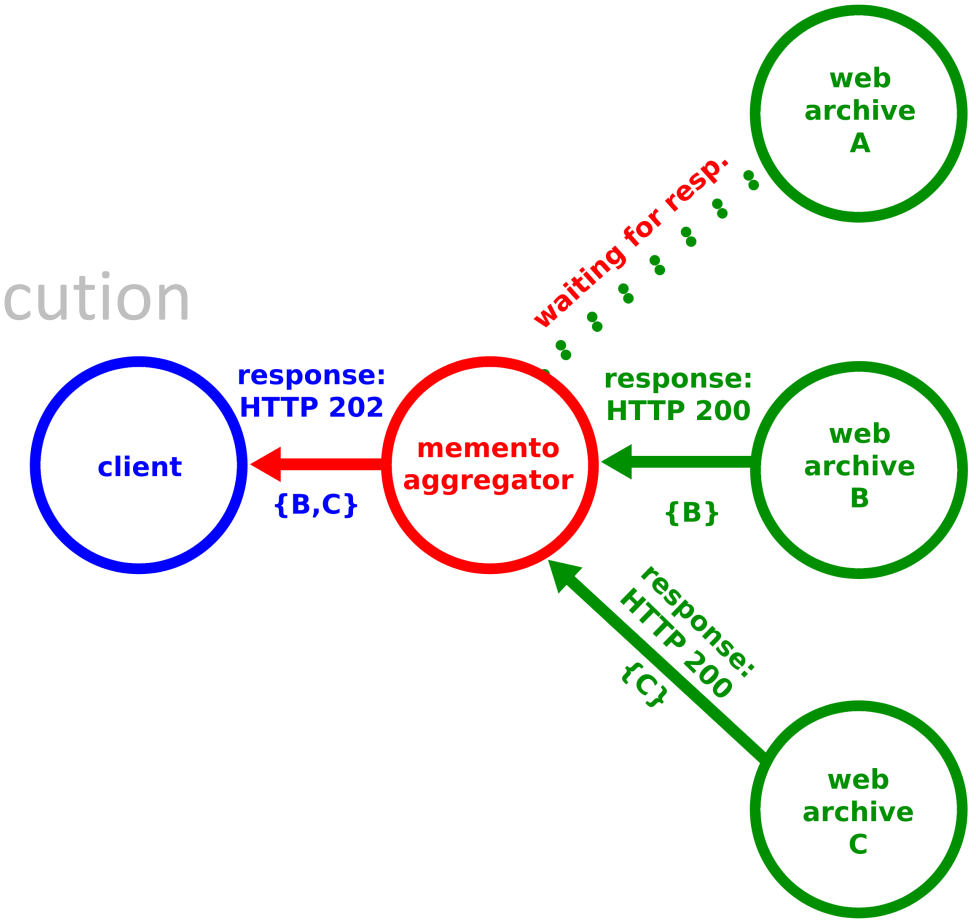
Newfound Capabilities

- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming

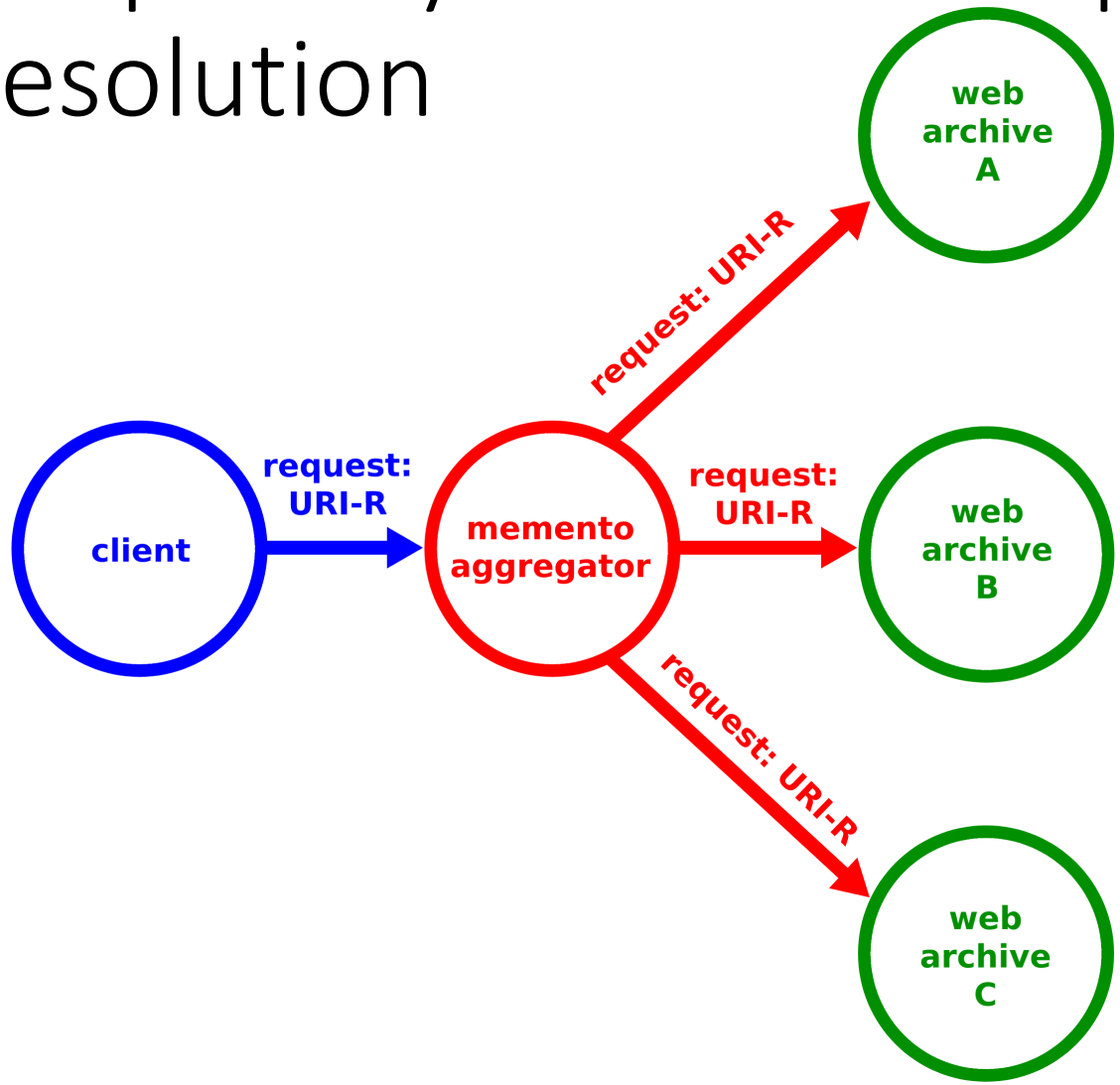
See `github.com/machawk1/mink`

Newfound Capabilities

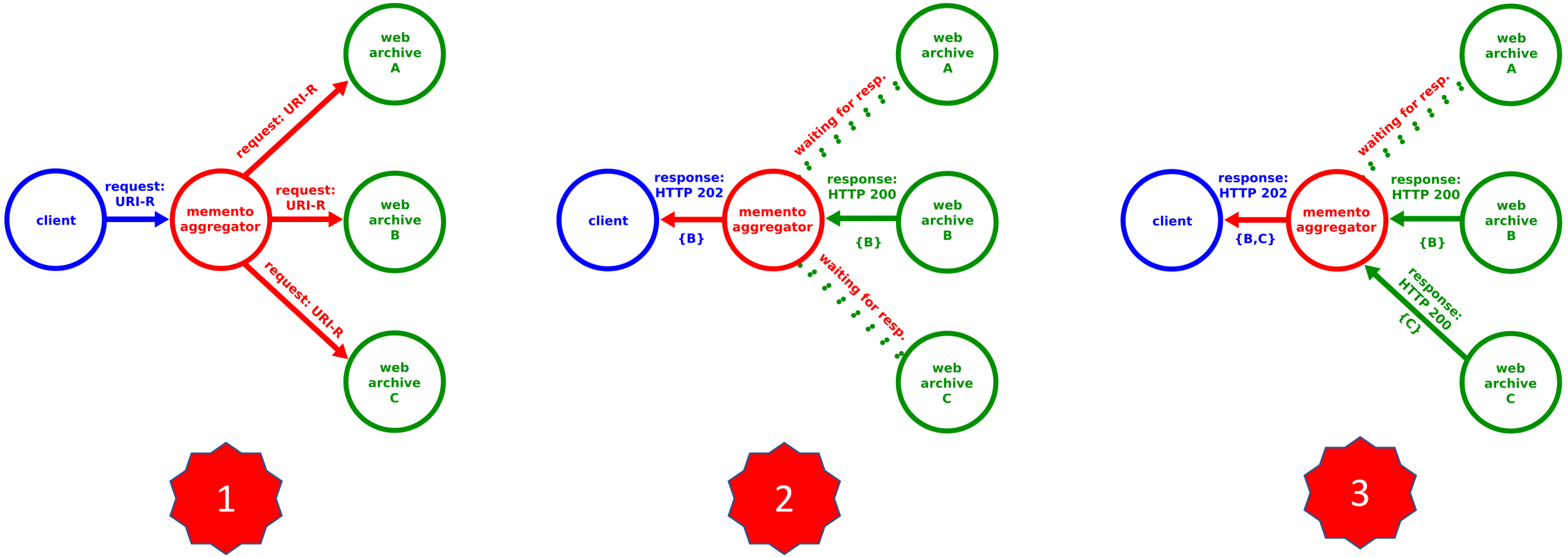
- User-defined set of archives
- Cycle detection
- Rescoping the aggregator for client-side execution
- Preliminary results streaming



Extended Capability Allows for Sophisticated Request Resolution



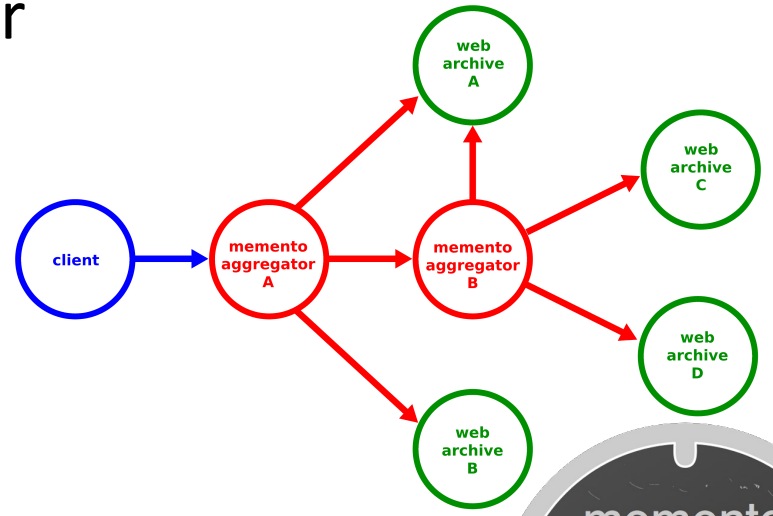
Asynchronous Processing and Streaming Response



IN CLOSING…

More Capable Aggregator→More Usable Web Archives

- Formally defined client-aggregator querying models
- Aggregator chaining facilitates archival aggregation but is problematic
- Much functional potential remains untapped for open source Memento Aggregators



Potential next steps:

- Client-side aggregators
- Re-consider functional cohesion