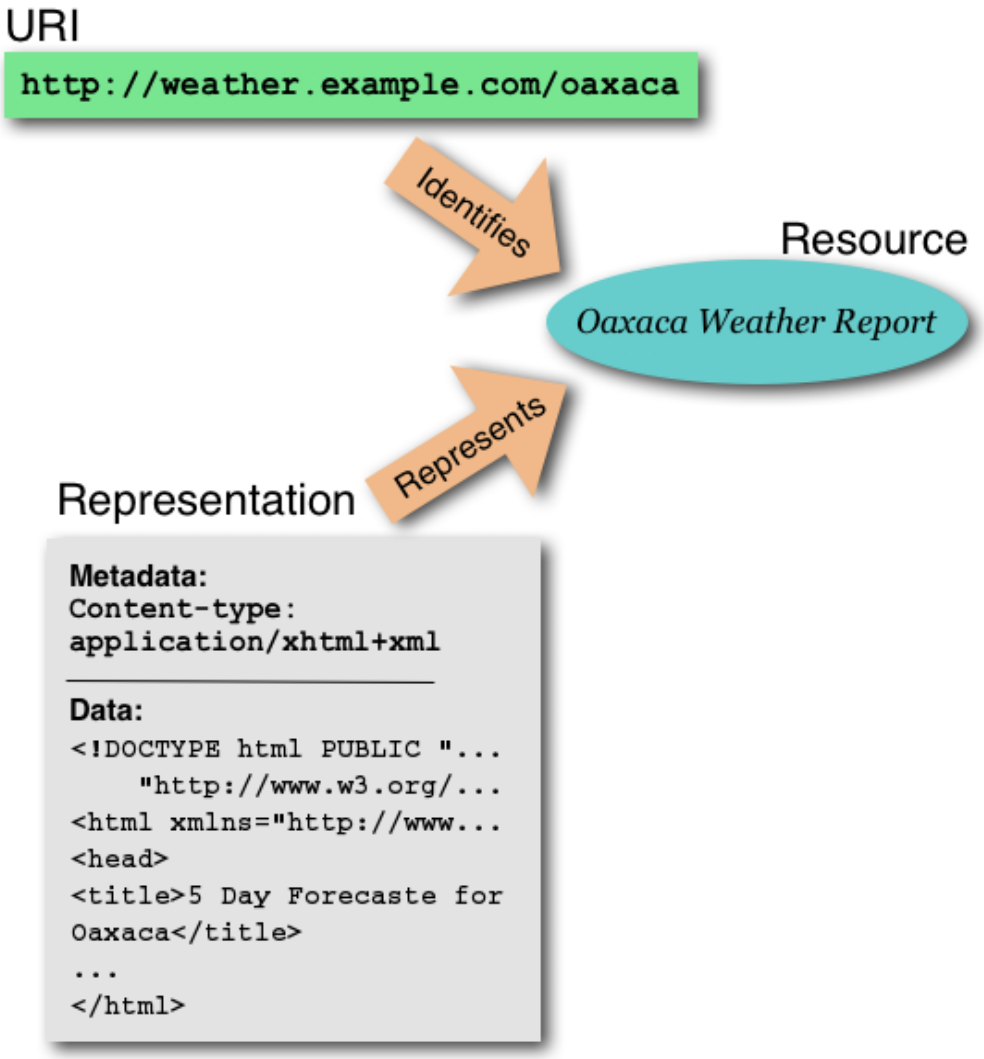
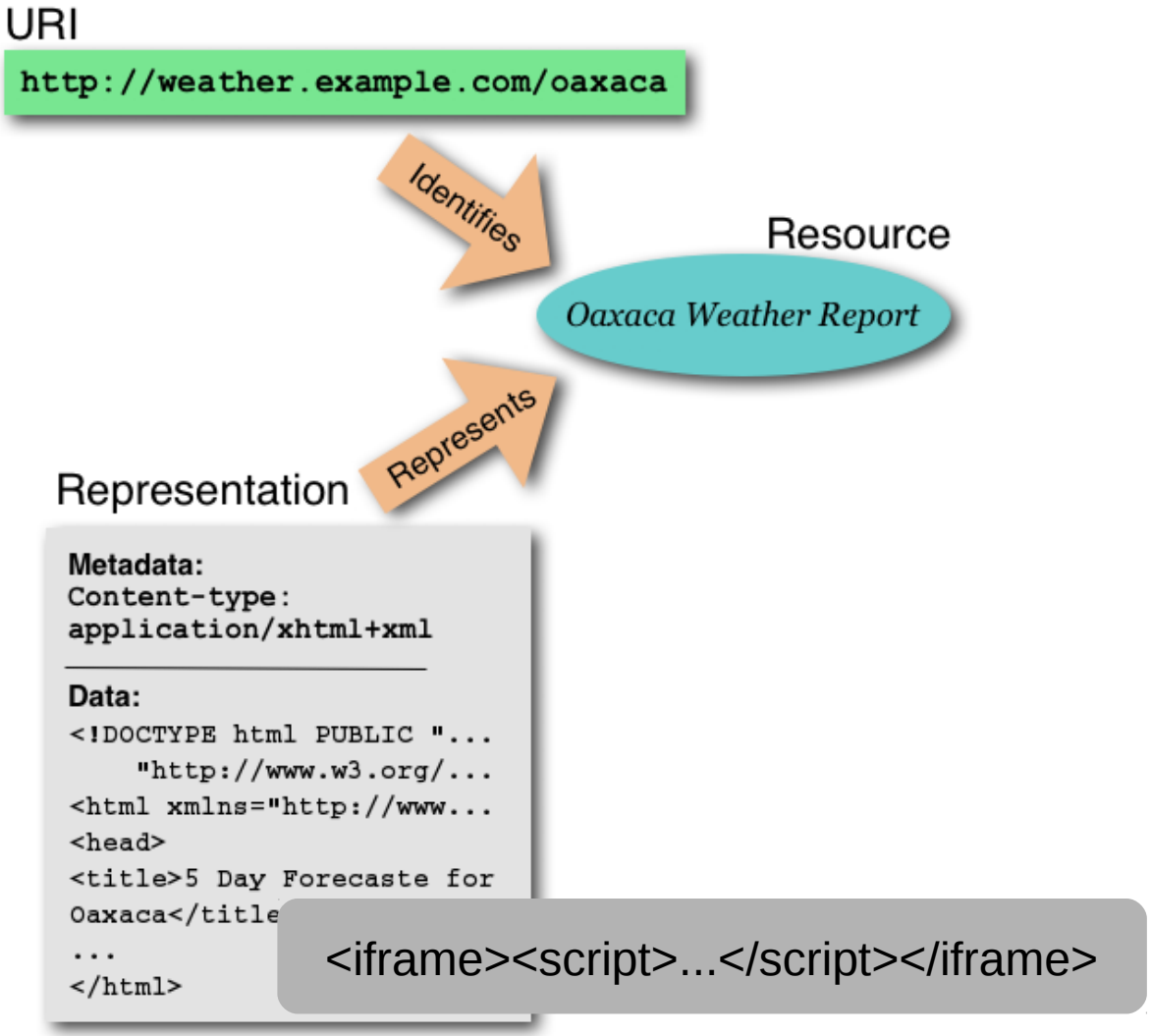
Archiving Deferred Representations Using a Two-Tiered Crawling Approach

Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson
Old Dominion University
iPRES2015
11/04/2015

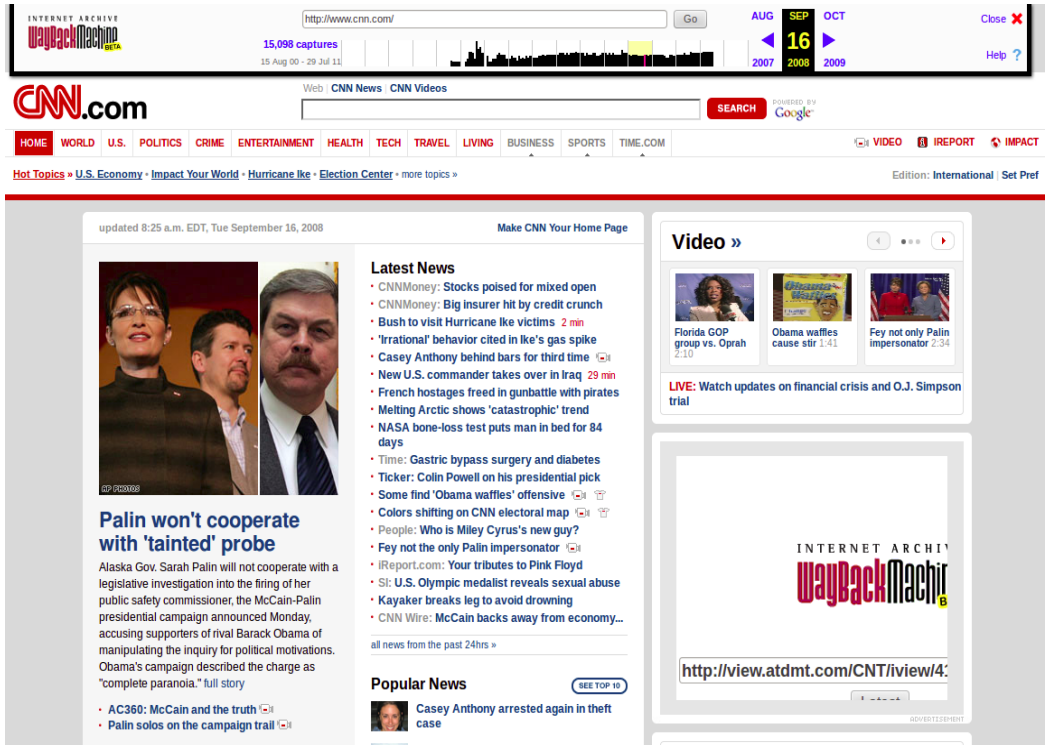
A simpler time...



Mass hysteria. Human sacrifices. Dogs and cats living together.

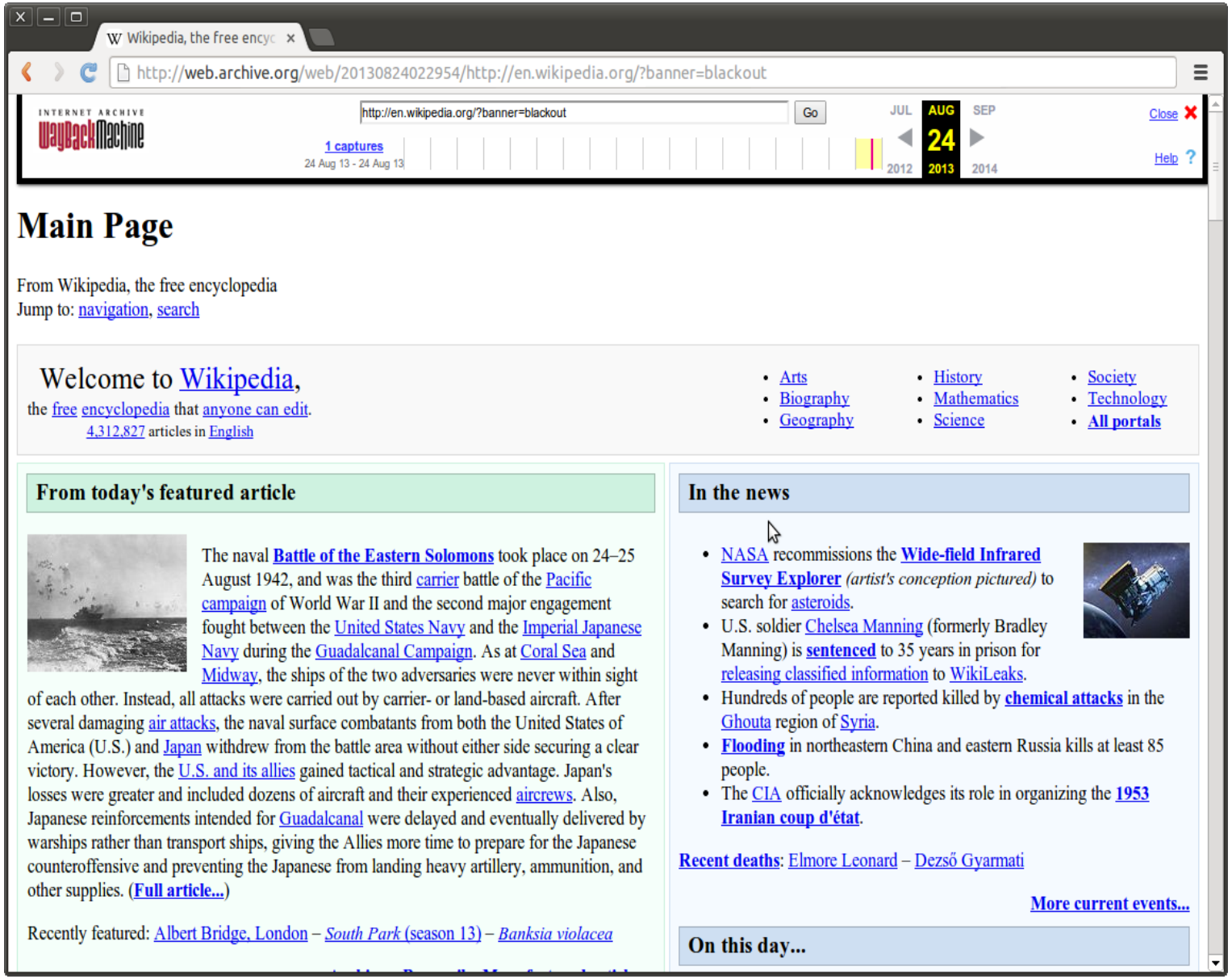


Missing resources (bad) and Temporal violations (worse)



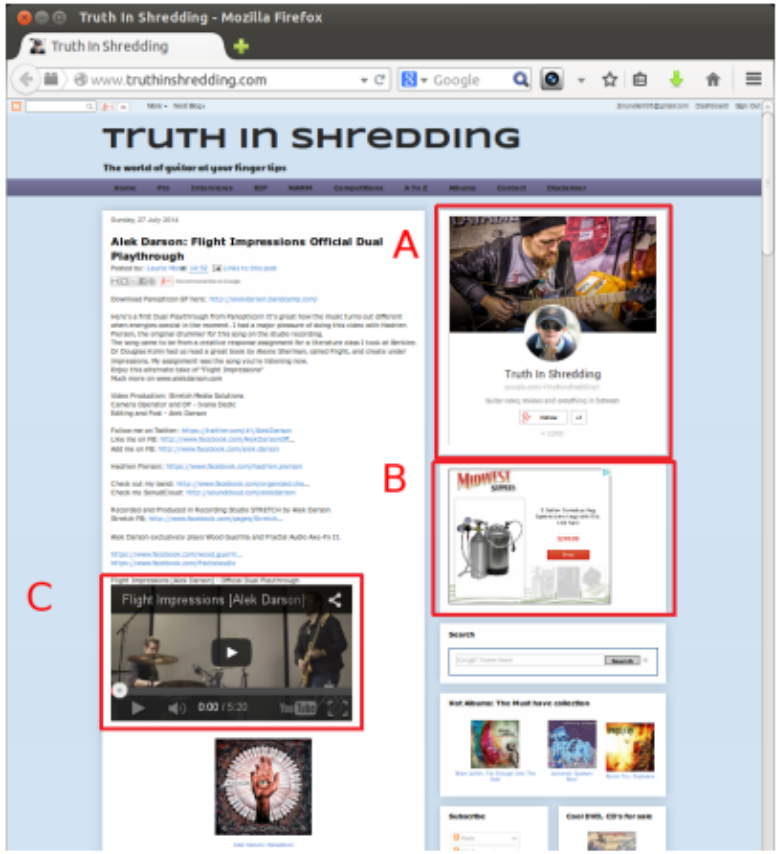
http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html

http://en.wikipedia.org/wiki/Main_Page January 18th, 2012

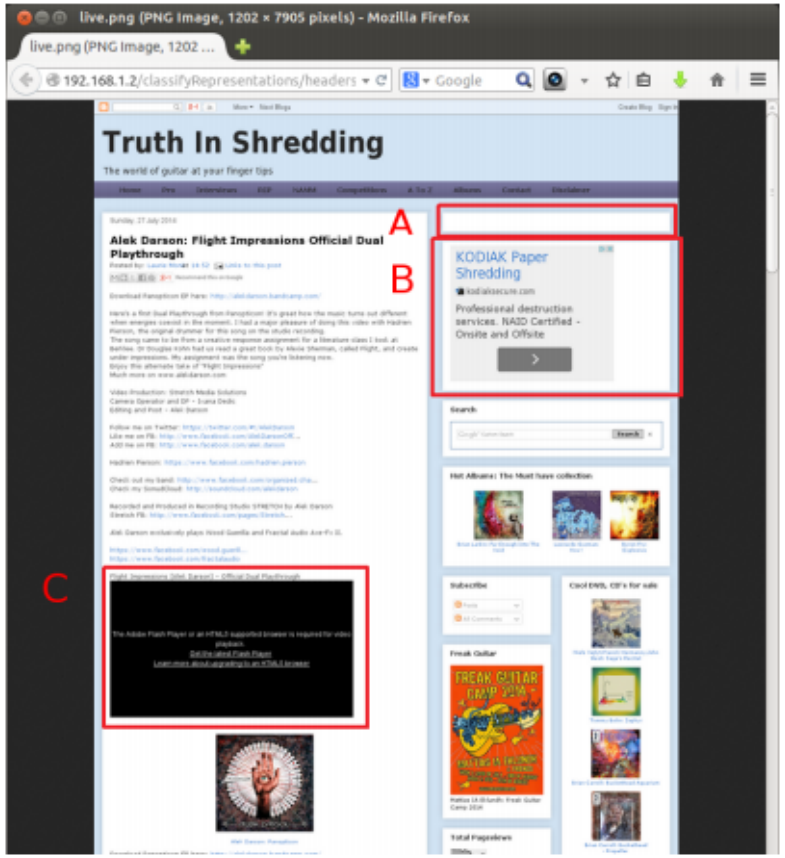


http://web.archive.org/web/20120118110520/http://en.wikipedia.org/wiki/Main_Page:
January 18th, 2012

Not all tools can crawl equally



Live Resource

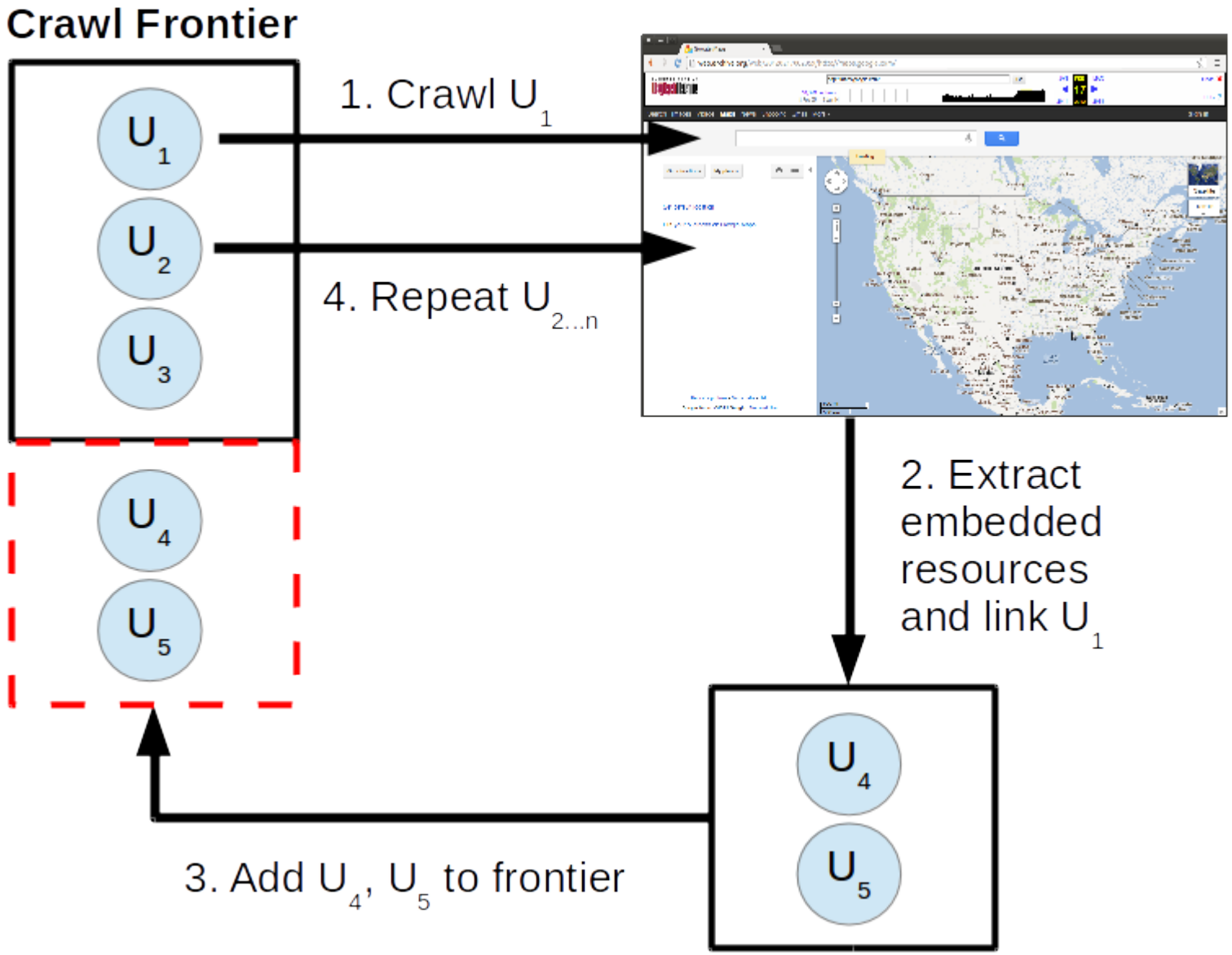


PhantomJS
Crawled

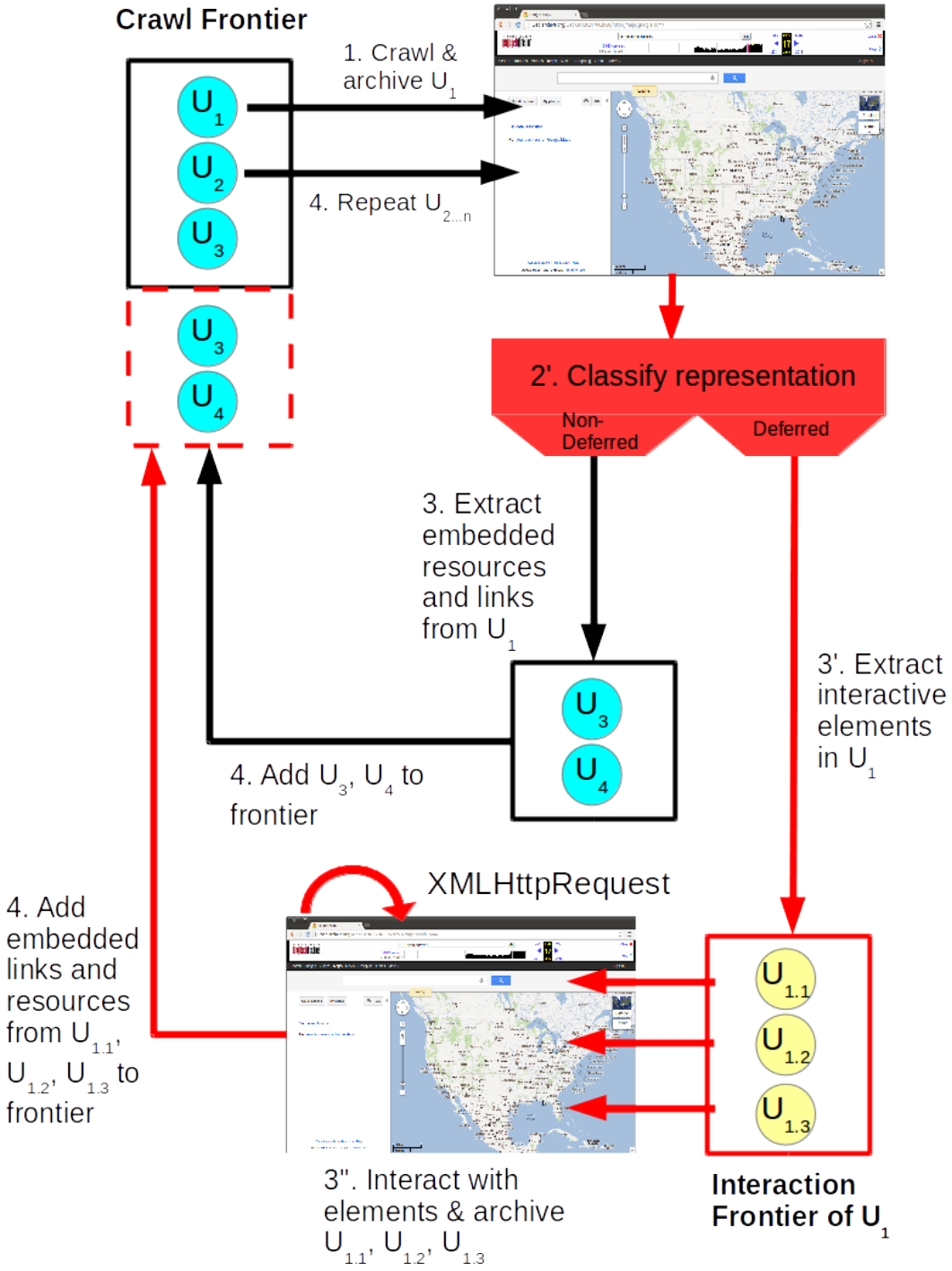
Heritrix Crawled,
Wayback replayed

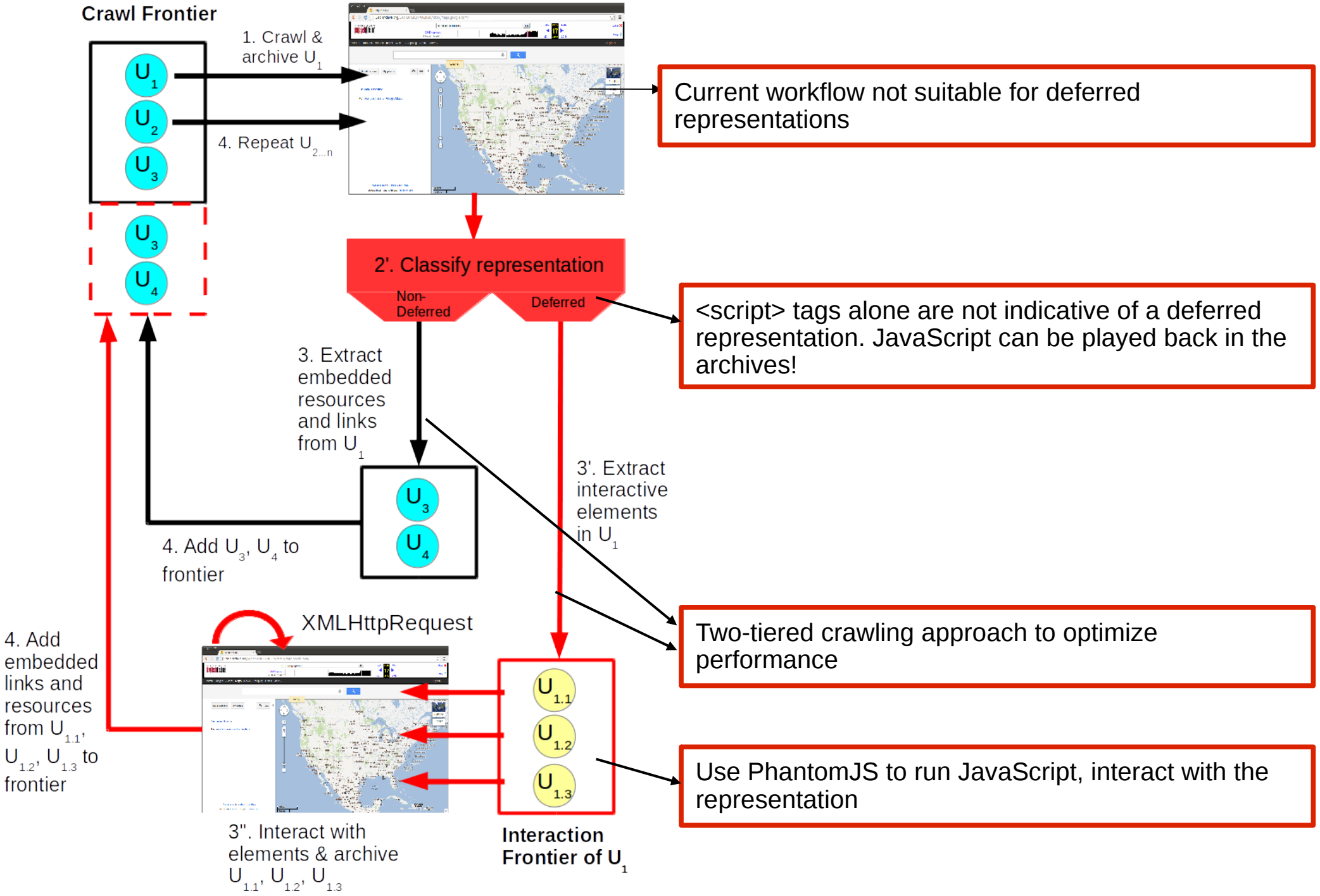
Current Workflow

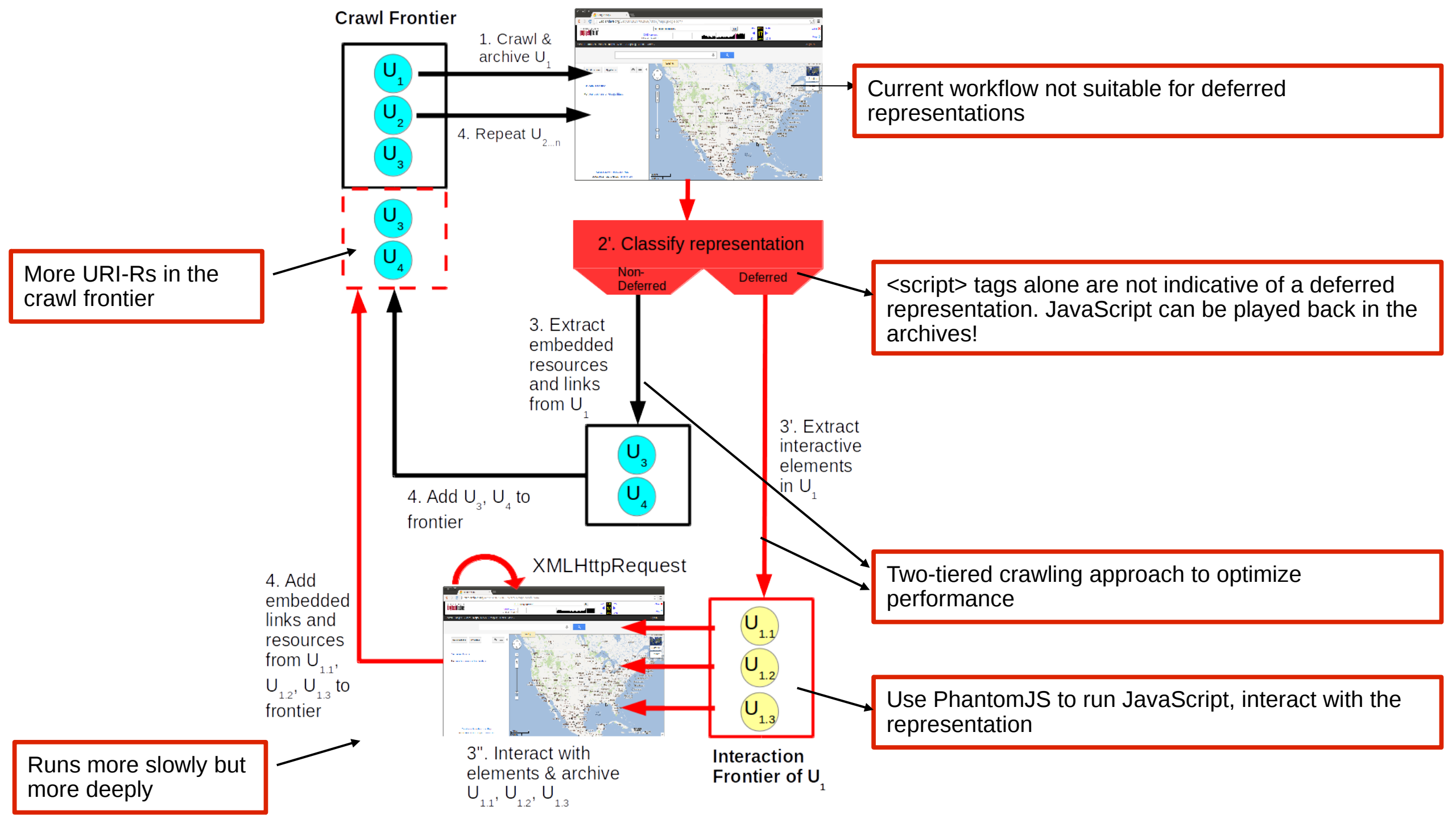
- Dereference URI-Rs
- Archive
- representation
- Extract embedded
- URI-Rs
- Repeat



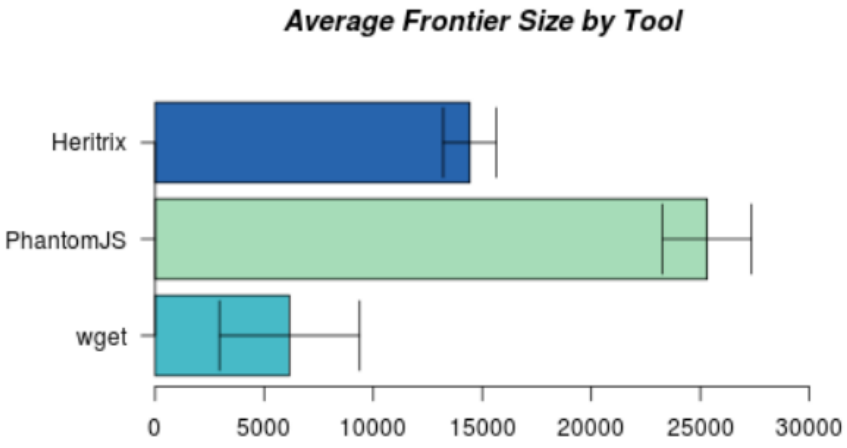
Proposed Workflow

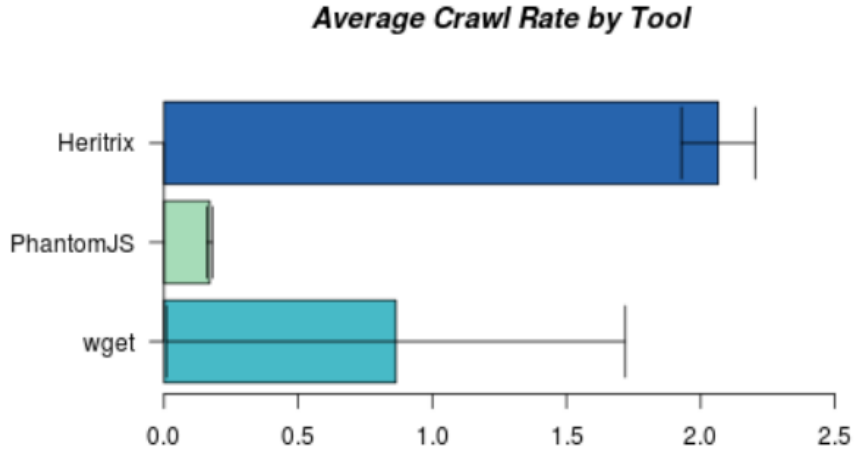


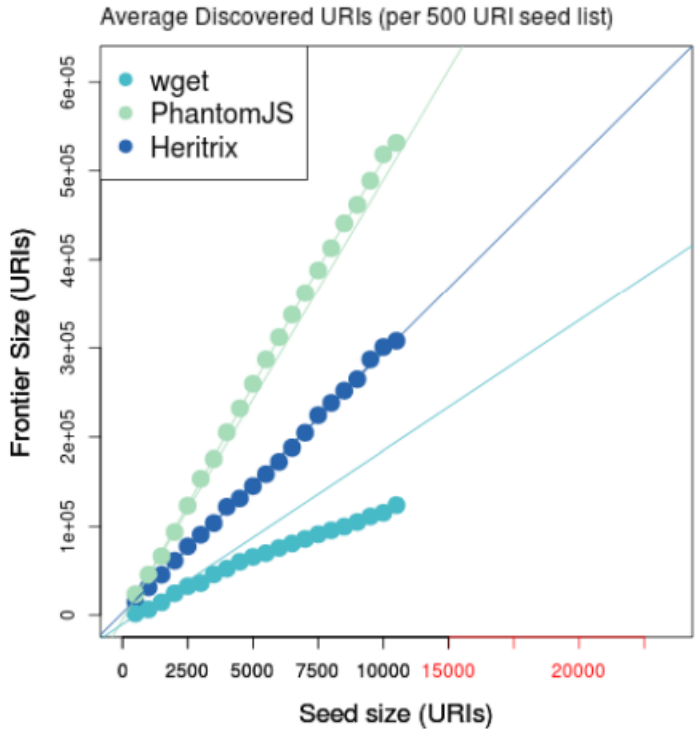

Crawl Frontier
U1
U2
U3
U3
U4
1. Crawl & archive U1
4. Repeat U2...n
2'. Classify representation
Non-Deferred
Deferred
3. Extract embedded resources and links from U1
U3
U4
4. Add U3, U4 to frontier
3'. Extract interactive elements in U1
XMLHttpRequest
4. Add embedded links and resources from U1.1, U1.2, U1.3 to frontier
3". Interact with elements & archive U1.1, U1.2, U1.3
U1.1
U1.2
U1.3
Interaction Frontier of U1
Current workflow not suitable for deferred representations
<script> tags alone are not indicative of a deferred representation. JavaScript can be played back in the archives!
Two-tiered crawling approach to optimize performance
Use PhantomJS to run JavaScript, interact with the representation

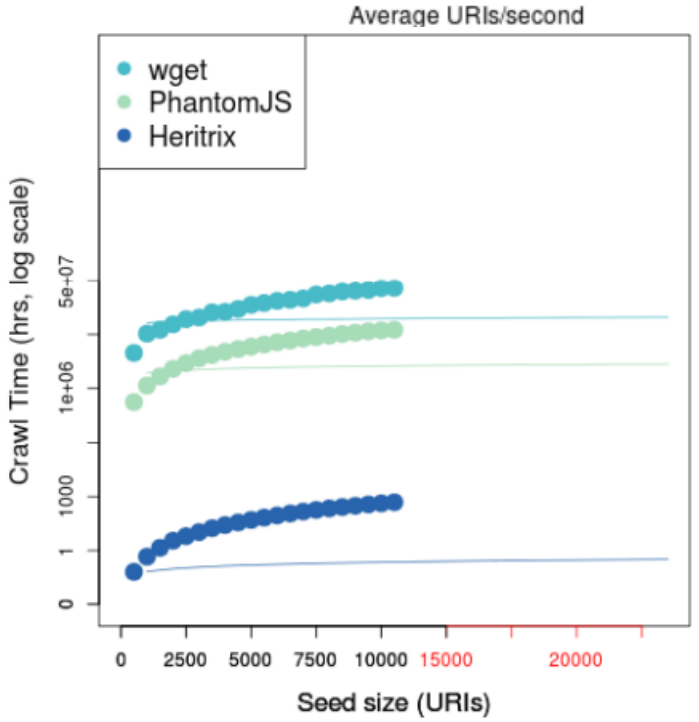

Crawl Frontier
U1
U2
U3
1. Crawl & archive U1
4. Repeat U2...n
U3
U4
2'. Classify representation
Non-Deferred
Deferred
3. Extract embedded resources and links from U1
U3
U4
4. Add U3, U4 to frontier
3'. Extract interactive elements in U1
XMLHttpRequest
4. Add embedded links and resources from U1.1, U1.2, U1.3 to frontier
U1.1
U1.2
U1.3
3". Interact with elements & archive U1.1, U1.2, U1.3
Interaction Frontier of U1
Current workflow not suitable for deferred representations
More URI-Rs in the crawl frontier
<script> tags alone are not indicative of a deferred representation. JavaScript can be played back in the archives!
Two-tiered crawling approach to optimize performance
Use PhantomJS to run JavaScript, interact with the representation
Runs more slowly but more deeply

Run-time & Frontier size PhantomJS vs. Heritrix

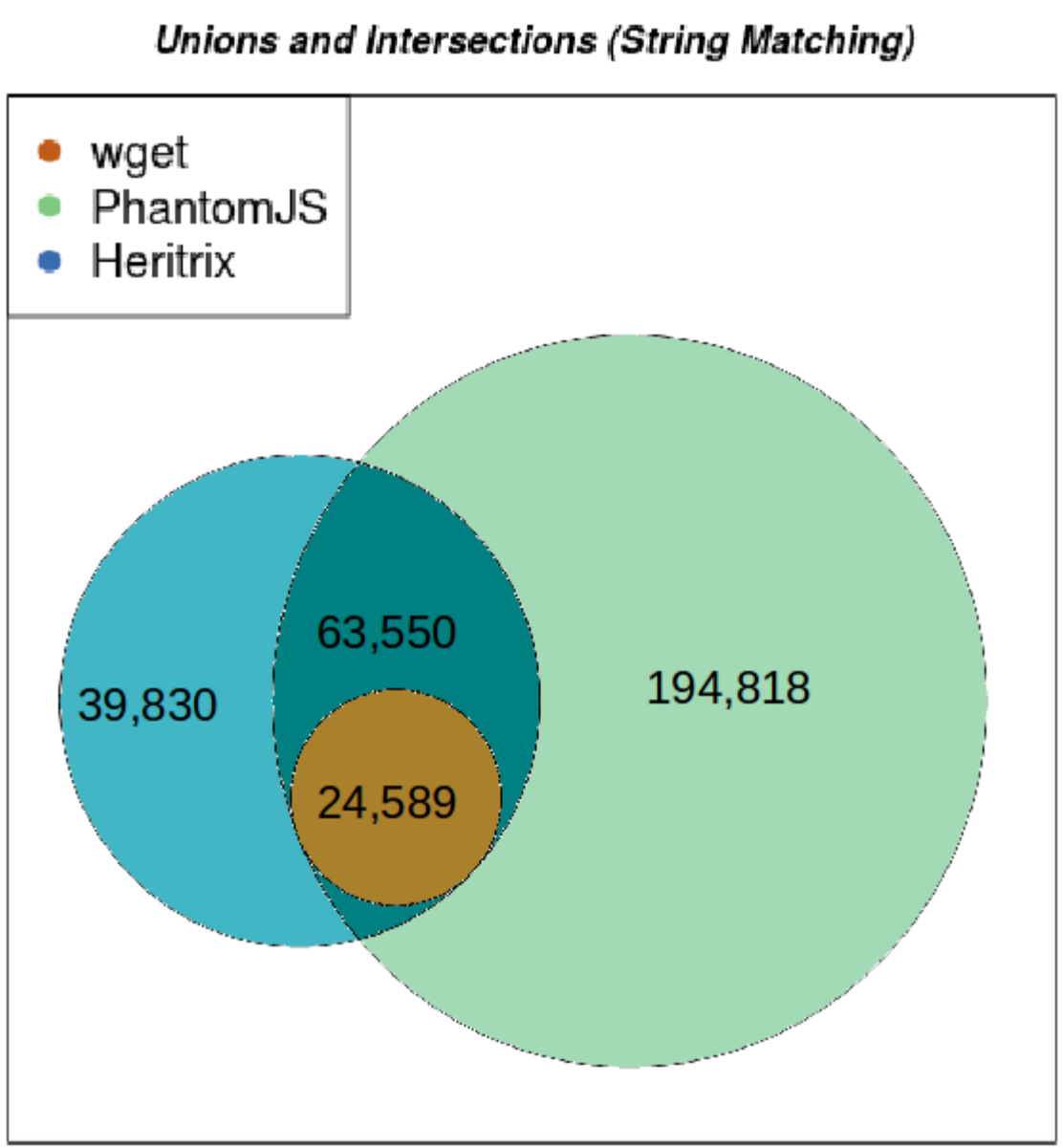


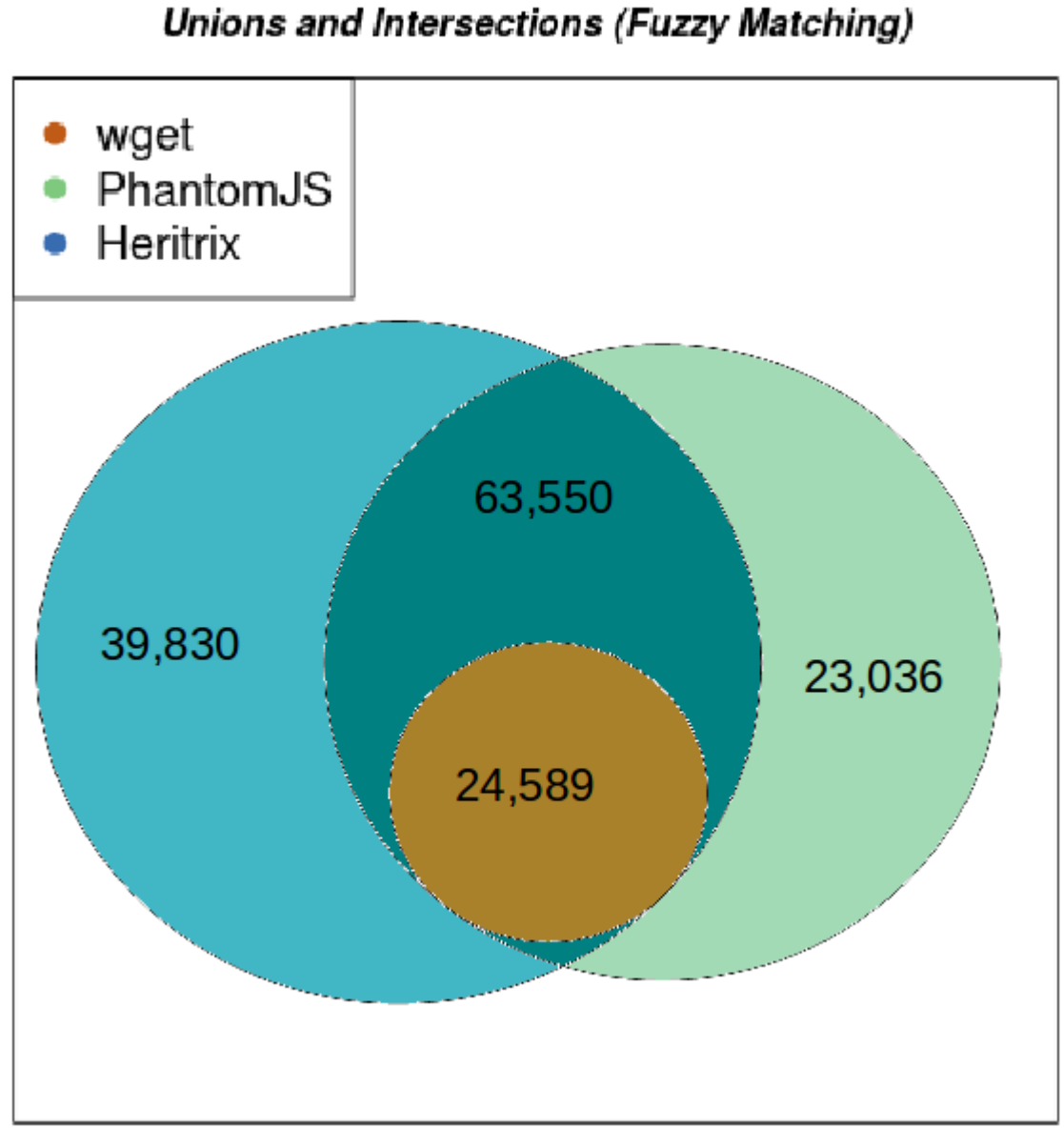






To appear: iPres2015


Unions and Intersections (String Matching)
wget
PhantomJS
Heritrix
39,830
63,550
24,589
194,818


Unions and Intersections (Fuzzy Matching)
wget
PhantomJS
Heritrix
39,830
63,550
24,589
23,036

Trim Policy	Original URI-R	Trimmed URI-R
No Trim	http://example.com/folder/index.html?param=value	http://example.com/folder/index.html?param=value
Origin Trim	http://example.com/folder/index.html?callback=cs.odu.edu	http://example.com/folder/index.html
Base Trim	http://example.com/folder/index.html?param=value	http://example.com/folder/index.html
Session Trim	http://example.com/folder/index.html?param=value&sessionid=12345	http://example.com/folder/index.html?param=value
HTTP Trim	http://example.com/folder/index.html?param=value&httpParam=http://www.test.com/	http://example.com/folder/index.html?param=value

Trim Type	URI Duplicates	URI and Entity Duplicates	Accuracy
No Trim	6,469	4,684	0.68
Origin Trim	7,078	4,749	0.68
Base Trim	10,359	5,191	0.56
Session Trim	8,159	4,921	0.64
HTTP Trim	7,315	4,868	0.67

Table 4: Detected duplicate URIs, entity bodies, and the overlap between the two using the five URI string trimming policies.

Constructed a classifier for Deferred Representations

Features	Classification	Accuracy	F-measure	Precision	Recall
DOM Features Only	Deferred	79%	79%	78%	81%
	Non-deferred			76%	80%
DOM & Resource Features	Deferred	81%	82%	79%	81%
	Non-deferred			90%	80%

Table 8: Classification success statistics for DOM-only and DOM and Resource feature sets.

Performance metrics of a two-tiered crawling approach

Crawl Strategy	Crawl Time (hrs)	Crawl Rate (t_{URI})	Frontier Size ($\|F\|$)
wget	416.16	0.864	129,443
Heritrix	407.53	2.065	302,961
PhantomJS	8,684.38	0.170	531,484
Heritrix + PhantomJS	9,100.54	0.152	537,609
Heritrix + PhantomJS with Classifier	6,495.23	0.196	458,815

Table 9: A summary of *extrapolated* performance (based on our calculations) of single- and two-tiered crawling approaches.

The classifier helps crawl deferred representations most efficiently

Crawler	URI-R Set	Seed Size	Frontier Size	Crawl Time (hrs)
P	Deferred	5,187	311,903	84.9
H	Non-deferred	4,813	124,728	23.6
H	Deferred	5,187	171,499	26.7
P	All URI-Rs	10,000	438,388	686
H	All URI-Rs	10,000	275,234	48.3
Two-tier	All URI-Rs	10,000	399,202	133

Table 10: A simulated two-tiered crawl showing that the frontier sizes can be optimized while mitigating the performance impact of PhantomJS's (P) crawl speed vs Heritrix's (H).

Current & Future Work

- Using PhantomJS to execute actions on the client
  - Pushing buttons
  - Selecting drop-downs
  - Archiving resulting representation changes
- Represent representation state in WARCs
  - Graph structure of embedded resources
  - Replay in the Wayback Machine

http://ws-dl.blogspot.com/2015/06/2015-06-26-phantomjsvisualevent-or.html

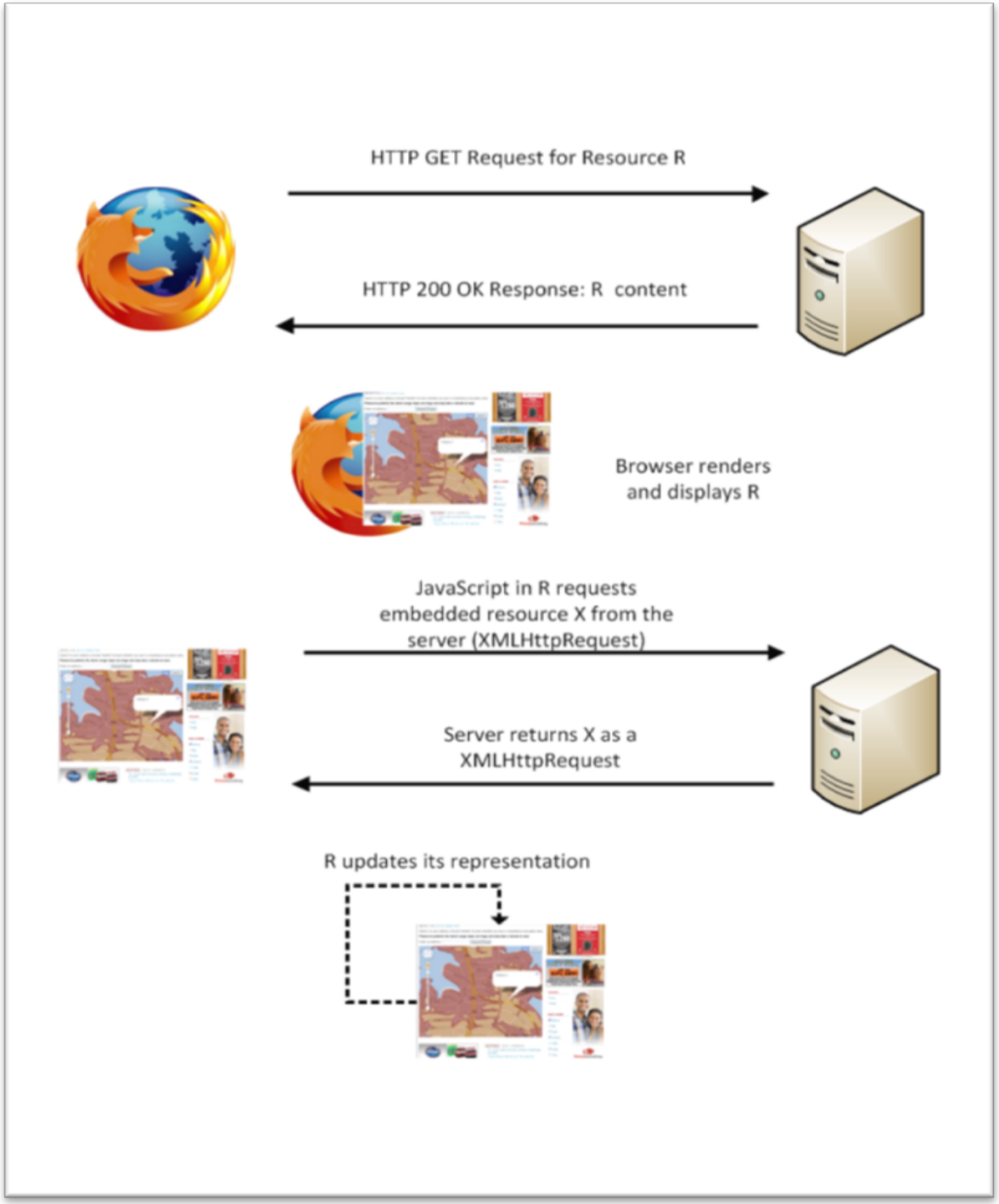
Backups

Trim Type	URI Duplicates	URI and Entity Duplicates	Accuracy
No Trim	6,469	4,684	0.68
Origin Trim	7,078	4,749	0.68
Base Trim	10,359	5,191	0.56
Session Trim	8,159	4,921	0.64
HTTP Trim	7,315	4,868	0.67

Table 4: Detected duplicate URIs, entity bodies, and the overlap between the two using the five URI string trimming policies.

Web Browsing Process

- User-controlled
- Interaction
- Environment variables



Web Browsing Process

At any given time, users get **“a”** representation.

There is no longer **“the”** representation that archives target.