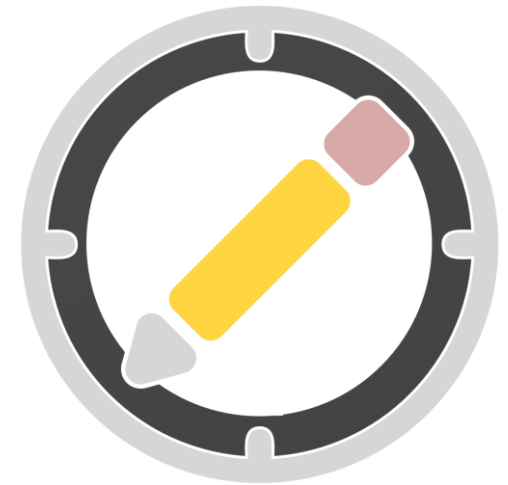


# Memento Tracer

## An Innovative Approach Towards Balancing Web Archiving at Scale and Quality

Martin Klein  
Los Alamos National Laboratory  
[@mart1nkle1n](https://twitter.com/mart1nkle1n)

Herbert Van de Sompel  
DANS  
[@hvdsomp](https://twitter.com/hvdsomp)



<http://tracer.mementoweb.org/>



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021

# Web Archiving Issues

**Scale**

VS.


**Quality**

VS.

**Resource Boundary**



# Scale!

 **Brewster Kahle**  
@brewster\_kahle Following

Wayback Machine update +4Billion to 658,661,007,000 web objects archived and served. ( small indexes update with recent captures, then a big sweep into a big update, like this one ). 658Billion! go @internetarchive go @waybackmachine !

10:56 AM - 8 Jul 2018

32 Retweets 65 Likes

1 32 65

[https://twitter.com/brewster\\_kahle/status/1016003169589981184](https://twitter.com/brewster_kahle/status/1016003169589981184)



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



# Scale!!

 **Brewster Kahle**  
@brewster\_kahle

Following

Wayback Machine update +4Billion to  
658,661,007,000 web objects  
served. (small indexes upda  
captures, then a big sweep ir  
like this one ). 658Billion! go  
[@internetarchive](#) go [@wayba](#)

10:56 AM - 8 Jul 2018

32 Retweets 65 Likes



1 32 65

 **Brewster Kahle**  
@brewster\_kahle

Following

Now 750 Billion web objects now in the  
Wayback Machine. (that be Billions and  
Billions). Love the sharing going on on the  
web. go [@waybackmachine](#) go  
[@internetarchive](#) !

INTERNET ARCHIVE

**WayBackMachine**

7:05 PM - 14 Jun 2019

54 Retweets 141 Likes



2 54 141

[https://twitter.com/brewster\\_kahle/status/1139700494748663809](https://twitter.com/brewster_kahle/status/1139700494748663809)



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



# Scale!!!



**Brewster Kahle**  
@brewster\_kahle

Following

Wayback Machine update +4Billion to  
658,661,007,000 web objects  
served. (small indexes update  
captures, then a big sweep in  
like this one). 658Billion! go  
[@internetarchive](#) go [@wayback](#)

10:56 AM - 8 Jul 2018

32 Retweets 65 Likes



1 32 65



**Brewster Kahle**  
@brewster\_kahle

Following

Now 750 Billion web objects now in the  
Wayback Machine. (that be Billions and  
Billions). Love the sharing going on on the  
web. go [@waybackma](#)  
[@internetarchive](#) !



**Brewster Kahle**  
@brewster\_kahle



Wayback Machine now has 898,570,440,000 URL's --  
serving millions of users as a foundation for what has  
been said on a constantly shifting WWW. (up 17 billion  
URLs in less than 1 month ago [twitter.com](#)  
[/brewster\\_kahle...](#) ) Go [@internetarchive](#) !

7:05 PM - 14 Jun 2019

54 Retweets 141 Likes



2 54 141

**Brewster Kahle** @brewster\_kahle · Jan 10, 2020

Wayback Machine just grew to 881,352,519,000 web URL's. That is 881  
Billion. For every one that becomes important in the news or in someones  
personal world, we crawl and store millions of others just-in-case. go  
[@internetarchive](#)

2:20 PM · Feb 5, 2020 · Twitter Web App

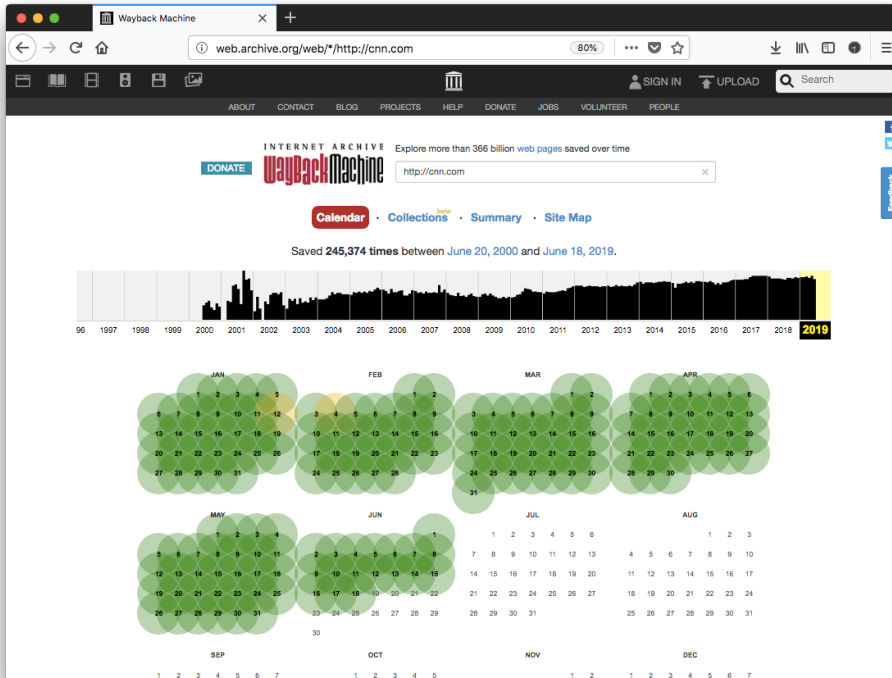
[https://twitter.com/brewster\\_kahle/status/1225167435399036939](https://twitter.com/brewster_kahle/status/1225167435399036939)



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



# Fidelity?



[http://web.archive.org/web/\\*/http://cnn.com](http://web.archive.org/web/*/http://cnn.com)

**Web Science and Digital Libraries Research Group**  
Research and Teaching Updates from the Web Science and Digital Libraries Research Group at Old Dominion University.

Friday, January 20, 2017  
**2017-01-20: CNN.com has been unarchivable since November 1st, 2016**

CNN.com has been unarchivable since 2016-11-01T15:01:31, at least by the common web archiving systems employed by the [Internet Archive](#), [archive.is](#), and [webcitation.org](#). The last known correctly archived page in the [Internet Archive's Wayback Machine](#) is 2016-11-01T13:15:40, with all versions since then producing some kind of error (including today's; 2017-01-20T09:16:50). This means that the most popular web archives have no record of the time immediately before the presidential election through at least today's presidential inauguration.

**Web Science and Digital Libraries**  
This blog is used to communicate research and education updates from the [Web Science and Digital Libraries Research Group](#) at Old Dominion University. Annual summaries are available for 2018, 2017, 2016, 2015, 2014, and 2013.

You can also follow us on Twitter: [@WebSciDL](#)

-- [Michael L. Nelson](#)

**Search This Blog**

**Blog Archive**

<https://ws-dl.blogspot.com/2017/01/2017-01-20-cnncom-has-been-unarchivable.html>



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



# Fidelity!!

**M** Mellon Foundation  
@MellonFdn

Following

.@Rhizome's Webrecorder equips digital archivists with tools to preserve the dynamic content that comprises today's web—including social media feeds about major events, videos, and new media art—for future use and scholarship: [ow.ly/Ramf30oegAT](https://ow.ly/Ramf30oegAT) #DH #Archives



**Preserving Social Media and Internet Art, Before Platforms ...**  
Rhizome's Webrecorder software is used to archive complex, interactive websites and new media art for future use and scholarship.  
[mellon.org](https://mellon.org)

8:15 AM - 12 Jun 2019

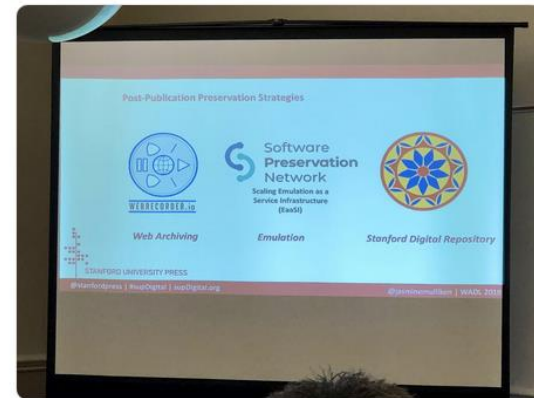
1 Retweet 3 Likes



**Ian Milligan**  
@ianmilligan1

Following

.@jasminemulliken: With all of these innovative born-digital manuscripts that @stanfordpress is now publishing... the importance of #webarchiving to support post-publication preservation strategies. #WADL2019



12:36 PM - 6 Jun 2019

2 Retweets 5 Likes



<https://twitter.com/MellonFdn/status/1138811967060267011>

<https://twitter.com/ianmilligan1/status/1136703505442324481>



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



# Scale?



**Martin Klein**

@mart1nkle1n

@jasminemulliken spent about a week to archive the [enchantingthedesert.com/home/](http://enchantingthedesert.com/home/) project with @webrecorder\_io tool #WADL2019



12:43 PM - 6 Jun 2019

4 Retweets 3 Likes



<https://twitter.com/mart1nkle1n/status/1136705116738904067>

<http://blog.supdigital.org/completing-the-archives-or-how-were-extending-the-life-of-web-based-digital-scholarship/>



**Memento Tracer**

@mart1nkle1n @hvdsomp

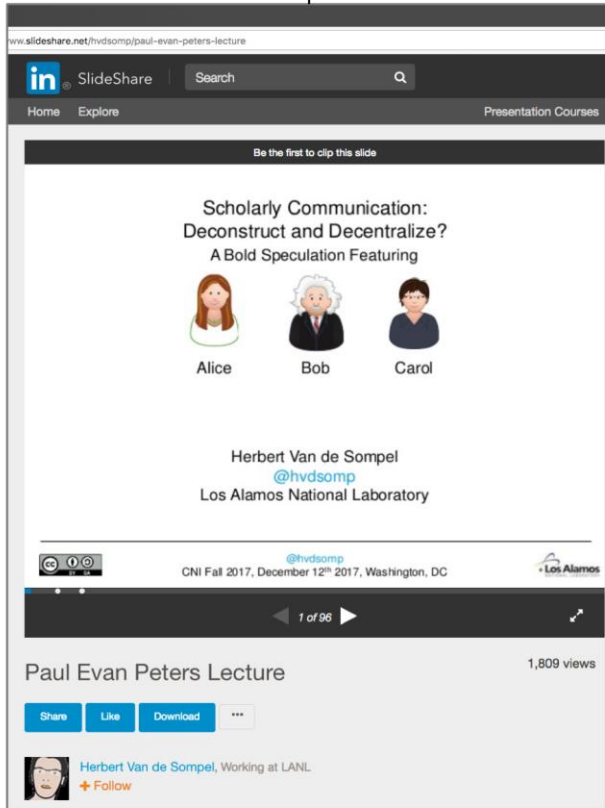
IIPC WAC 2020 – 2021, June 15 2021



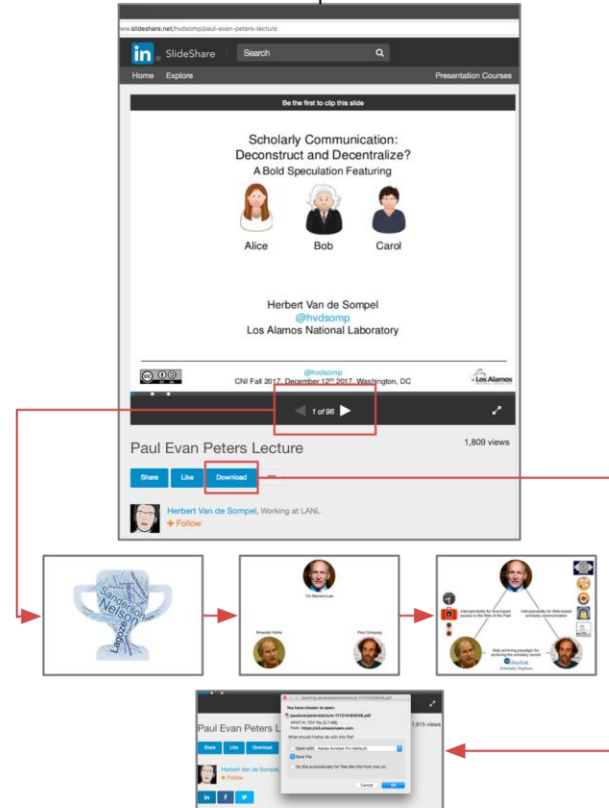


# Resource Boundary?

the page at URI-A  
is not the artifact



the artifact includes resources  
reachable via URI-A



<https://www.slideshare.net/hvdsomp/paul-evan-peters-lecture>



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



# Resource Boundary?

GitHub - mementoweb/memento\_extensions

54 commits | 4 branches | 5 releases | 1 contributor | View license

chrome restructured the repo after merging firefox and chrome code last year

doc updated licenses, code documentation, started different versioning nu... 6 years ago

firefox restructured the repo after merging firefox and chrome code last year

img animated mem logos, random date on first install, improved ajax requests 3 years ago

lib updated jquery to v3.3.1 last year

LICENSE added license 3 years ago

README.md Update README.md last year

contentScript.js merged memento.js and other changes from memento\_firefox for a unifil... last year

make\_chrome\_zip.sh updated jquery to v3.3.1 last year

make\_firefox\_zip.sh restructured the repo after merging firefox and chrome code last year

memento.js restructured the repo after merging firefox and chrome code last year

memento\_firefox.zip updated jquery to v3.3.1 last year

options.css merged memento.js and other changes from memento\_firefox for a unifil... last year

options.html merged memento.js and other changes from memento\_firefox for a unifil... last year

options.js added support for archivelist file, automatically loads archives from... 4 years ago

popup.html merged memento.js and other changes from memento\_firefox for a unifil... last year

popup.js merged memento.js and other changes from memento\_firefox for a unifil... last year

### The Memento Extension for Chrome and Firefox Browsers

Chrome Download: <https://chrome.google.com/webstore/detail/memento-time-travel/fgbfjedahaajcpaakbgilmojkaqghm7hi-en&gl=US>

Travel to the past of the web by right-clicking pages and links.

Memento for Chrome and Firefox allows you to seamlessly navigate between the present web and the web of the past. It turns your browser into a web time travel machine that is activated by means of a Memento sub-menu that is available on right-click.

First, select a date for time travel by clicking the black Memento extension icon. Now right-click on a web page, and click the "Get near ..." option from the Memento sub-menu to see what the page looked like around the selected date. Do the same for any link in a page to see what the linked page looked like. If you hit one of those nasty "Page not Found" errors, right-click and select the "Get near current time" option to see what the page looked like before it vanished from the web. When on a past version of a page - the Memento extension icon is now red - right-click the page and select the "Get current time" option to see what it looks like now.

Memento for Chrome and Firefox obtains prior versions of pages from web archives around the world, including the

GitHub - mementoweb/memento\_extensions

54 commits | 4 branches | 5 releases | 1 contributor | View license

Clone or download

chrome restructured the repo after merging firefox and chrome code last year

doc updated licenses, code documentation, started different versioning nu... 6 years ago

firefox restructured the repo after merging firefox and chrome code last year

img animated mem logos, random date on first install, improved ajax requests 3 years ago

lib updated jquery to v3.3.1 last year

LICENSE added license 3 years ago

README.md Update README.md last year

contentScript.js merged memento.js and other changes from memento\_firefox for a unifil... last year

make\_chrome\_zip.sh updated jquery to v3.3.1 last year

make\_firefox\_zip.sh restructured the repo after merging firefox and chrome code last year

memento.js restructured the repo after merging firefox and chrome code last year

memento\_firefox.zip updated jquery to v3.3.1 last year

options.css merged memento.js and other changes from memento\_firefox for a unifil... last year

options.html merged memento.js and other changes from memento\_firefox for a unifil... last year

options.js added support for archivelist file, automatically loads archives from... 4 years ago

popup.html merged memento.js and other changes from memento\_firefox for a unifil... last year

popup.js merged memento.js and other changes from memento\_firefox for a unifil... last year

### The Memento Extension for Chrome and Firefox Browsers

Chrome Download: <https://chrome.google.com/webstore/detail/memento-time-travel/fgbfjedahaajcpaakbgilmojkaqghm7hi-en&gl=US>

Travel to the past of the web by right-clicking pages and links.

Memento for Chrome and Firefox allows you to seamlessly navigate between the present web and the web of the past. It turns your browser into a web time travel machine that is activated by means of a Memento sub-menu that is available on right-click.

First, select a date for time travel by clicking the black Memento extension icon. Now right-click on a web page, and click the "Get near ..." option from the Memento sub-menu to see what the page looked like around the selected date. Do the same for any link in a page to see what the linked page looked like. If you hit one of those nasty "Page not Found" errors, right-click and select the "Get near current time" option to see what the page looked like before it vanished from the web. When on a past version of a page - the Memento extension icon is now red - right-click the page and select the "Get current time" option to see what it looks like now.

Memento for Chrome and Firefox obtains prior versions of pages from web archives around the world, including the

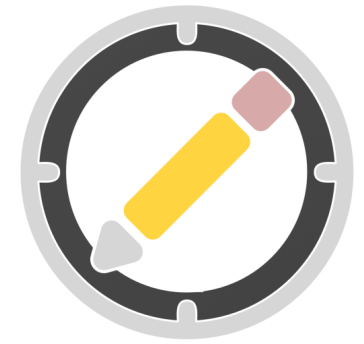
[https://github.com/mementoweb/memento\\_extensions](https://github.com/mementoweb/memento_extensions)



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021



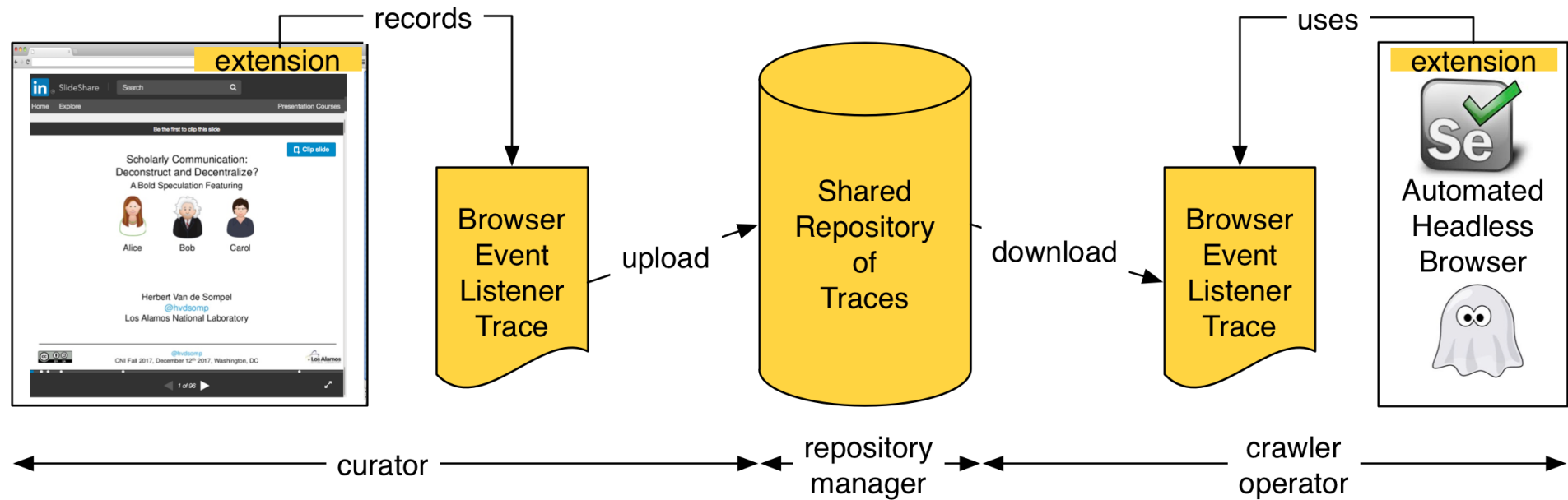
# Memento Tracer Framework



Inspired by:

- LOCKSS
  - Same automated approach for resources of a class
- Webrecorder
  - Manual recording of web resources
- Various attempts aimed at automating interactions/behaviors
  - E.g., Brozzler, Browsertrix

# Memento Tracer Framework





# Memento Tracer DEMO



## Creating Traces

- for GitHub repos  
<https://www.dropbox.com/s/7nw21zxlbvhs41e/github.mov?dl=0>
- for Slideshare slide decks  
<https://www.dropbox.com/s/aq6yqt5ccgwdpe2/slideshare.mov?dl=0>
- GitHub repos and Slideshare slide decks nested  
[https://www.dropbox.com/s/gtr5ms5bnc5vhds/github\\_slideshare\\_nested.mov?dl=0](https://www.dropbox.com/s/gtr5ms5bnc5vhds/github_slideshare_nested.mov?dl=0)

## Capture with a Trace

- Slideshare deck  
[https://www.dropbox.com/s/jiwpjh4db5n3919/capture\\_slideshare.mov?dl=0](https://www.dropbox.com/s/jiwpjh4db5n3919/capture_slideshare.mov?dl=0)

# Current Memento Tracer Capabilities

- Single clicks/links
- All links in an area
- Repeated click on links, with stop condition
  - Slides
  - Pagination
- Scroll
  - Defined number of viewports
  - Until "the end"
- Hover
  - E.g., navigational elements
- Nested traces i.e., "trace in a trace"
  - Trace for portal A → follow link to portal B → execute trace for portal B



**Memento Tracer**

@mart1nkle1n @hvdsomp

IIPC WAC 2020 – 2021, June 15 2021

# Memento Tracer Benefits

- Scalability
  - Trace created once is applicable to all web resources of the same class
  - Traces shared via repository (edits, versioning)
- Quality
  - Trace used as set of instructions for browser-based capture framework
  - Resource boundary explicit
- Tradeoff
  - Quality vs performance analysis:  
<https://arxiv.org/abs/1909.04404>



# Memento Tracer Challenges

- Language used to express Traces (interoperability)
- Organization of the shared repository for Traces
- Limitations of the browser event listener approach for recording Traces
- Selection of a Trace for capturing a web publication by other means than URI pattern



**Memento Tracer**

@mart1nkle1n @hvdsomp

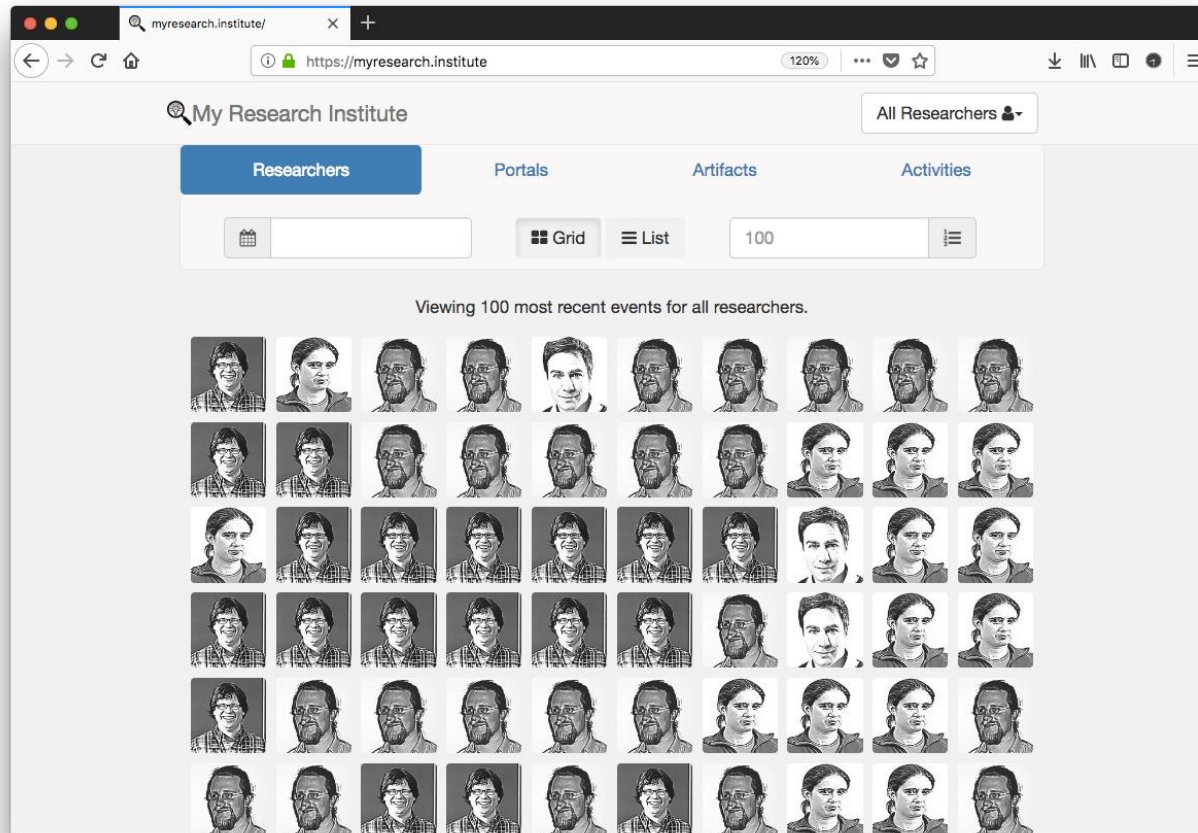
IIPC WAC 2020 – 2021, June 15 2021





# Memento Tracer in Action

<https://myresearch.institute/>



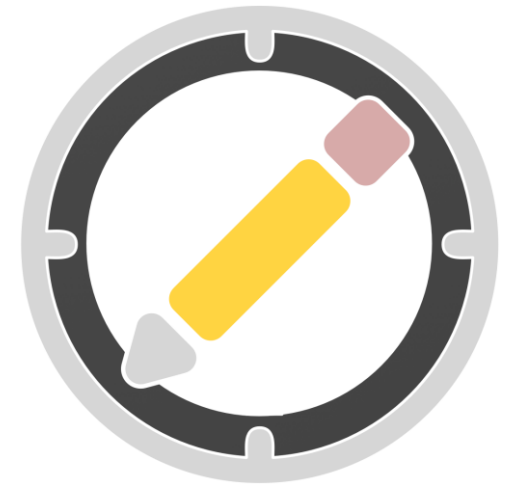
For more details and statistics, see our 2019 CNI Spring meeting slides:  
<https://www.slideshare.net/martinklein0815/an-institutional-perspective-to-rescue-scholarly-orphans>

# Memento Tracer

## An Innovative Approach Towards Balancing Web Archiving at Scale and Quality

Martin Klein  
Los Alamos National Laboratory  
[@mart1nkle1n](https://twitter.com/mart1nkle1n)

Herbert Van de Sompel  
DANS  
[@hvdsomp](https://twitter.com/hvdsomp)



<http://tracer.mementoweb.org/>



**Memento Tracer**  
@mart1nkle1n @hvdsomp  
IIPC WAC 2020 – 2021, June 15 2021