Memento Tracer
An Innovative Approach Towards Balancing Web Archiving at Scale and Quality

Martin Klein
Los Alamos National Laboratory
@mart1nkle1n

Herbert Van de Sompel
DANS
@hvdsomp

http://tracer.mementoweb.org/

Web Archiving Issues

Scale

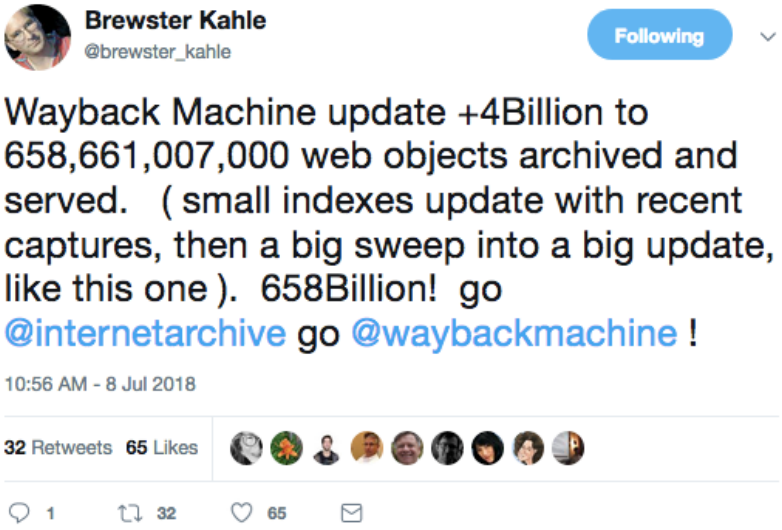
vs.

Quality

vs.

Resource Boundary

Scale!



https://twitter.com/brewster_kahle/status/1016003169589981184

Scale!!

https://twitter.com/brewster_kahle/status/1139700494748663809

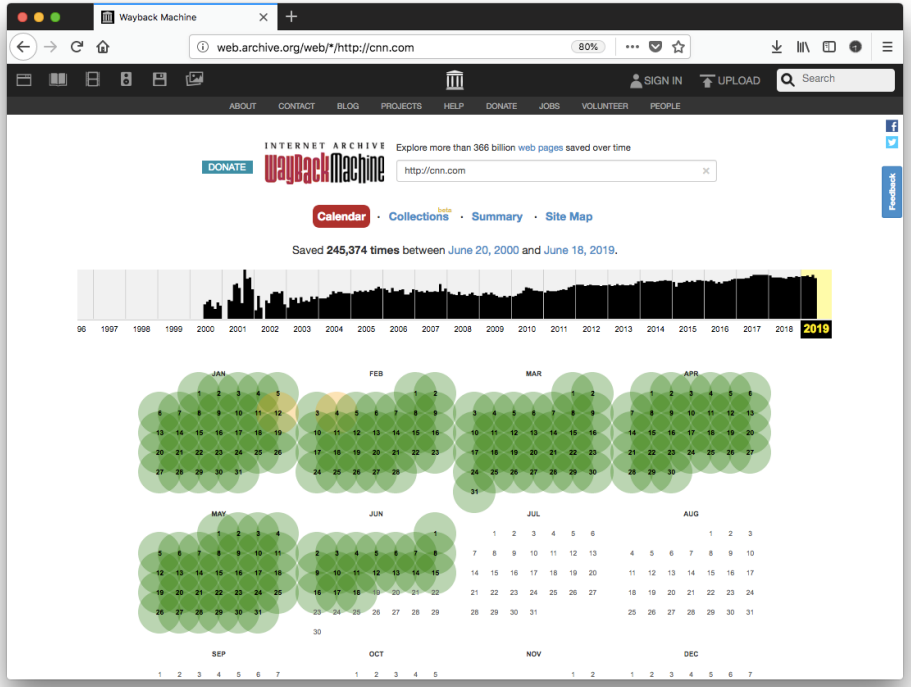
Scale!!!

https://twitter.com/brewster_kahle/status/1225167435399036939

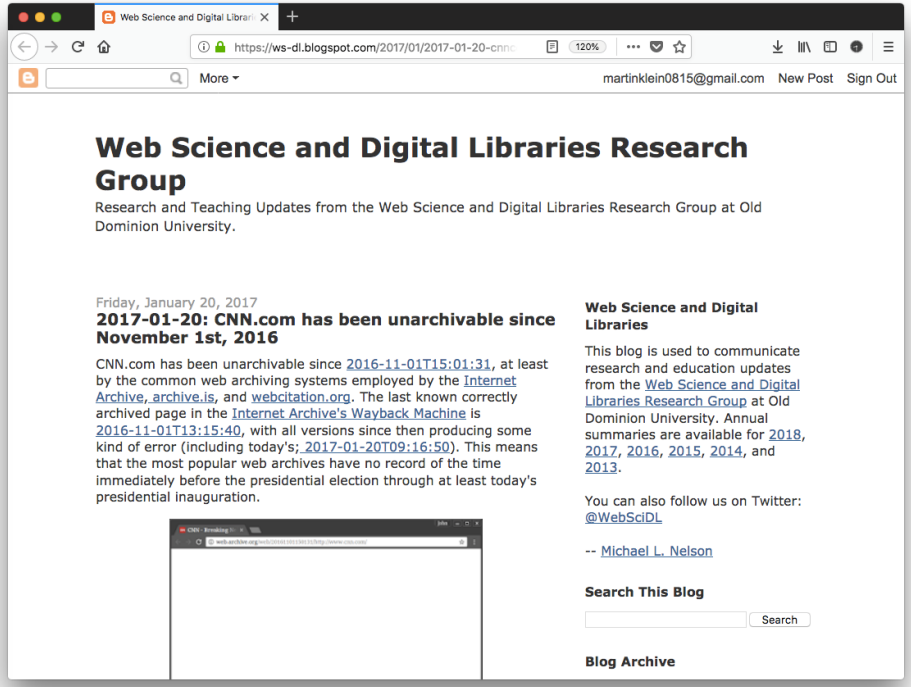
Memento Tracer
@mart1nkle1n @hvdsomp
IIPC WAC 2020 – 2021, June 15 2021

Data Archiving and Networked Services DANS
Los Alamos NATIONAL LABORATORY

Fidelity?



http://web.archive.org/web/*/http://cnn.com



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

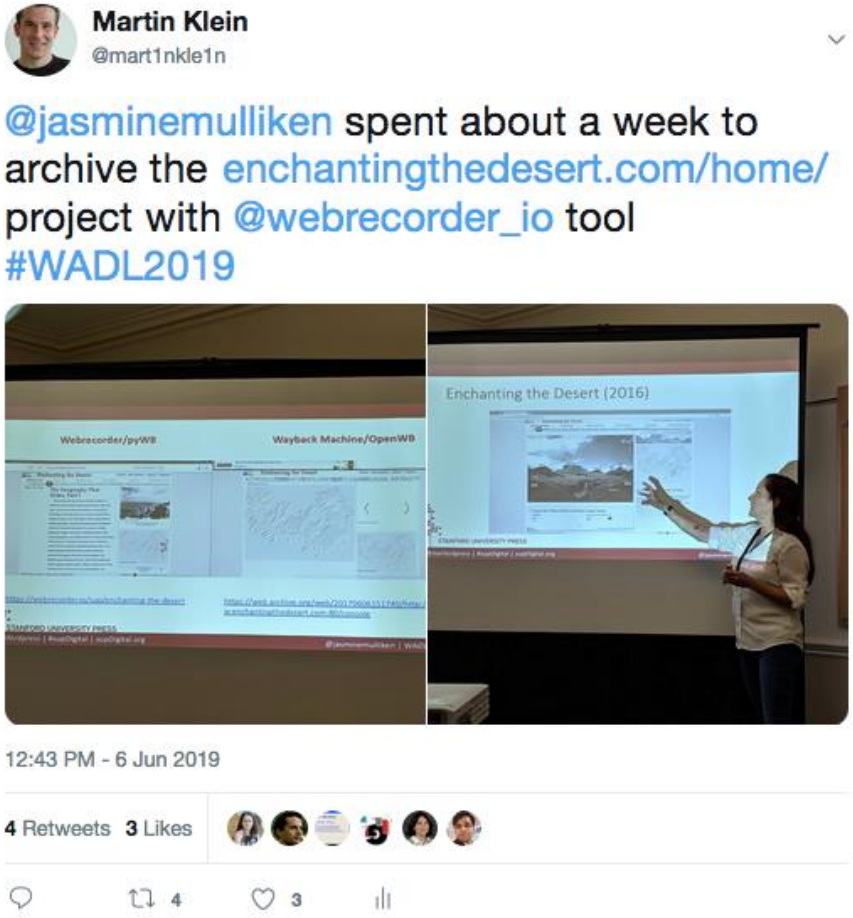
Memento Tracer
@mart1nkle1n @hvdsomp
IIPC WAC 2020 – 2021, June 15 2021

Data Archiving and Networked Services DANS
Los Alamos NATIONAL LABORATORY

Fidelity!!

https://twitter.com/MellonFdn/status/1138811967060267011 https://twitter.com/ianmilligan1/status/1136703505442324481

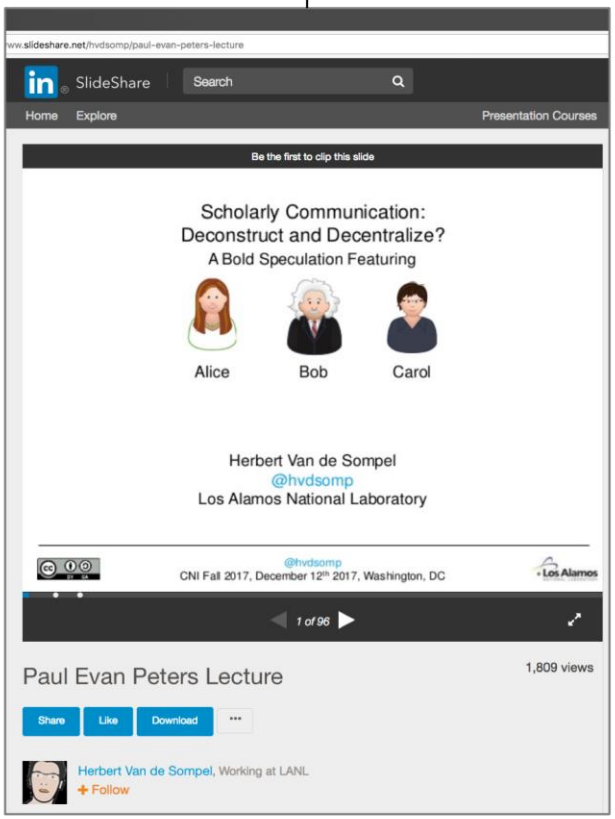
Scale?



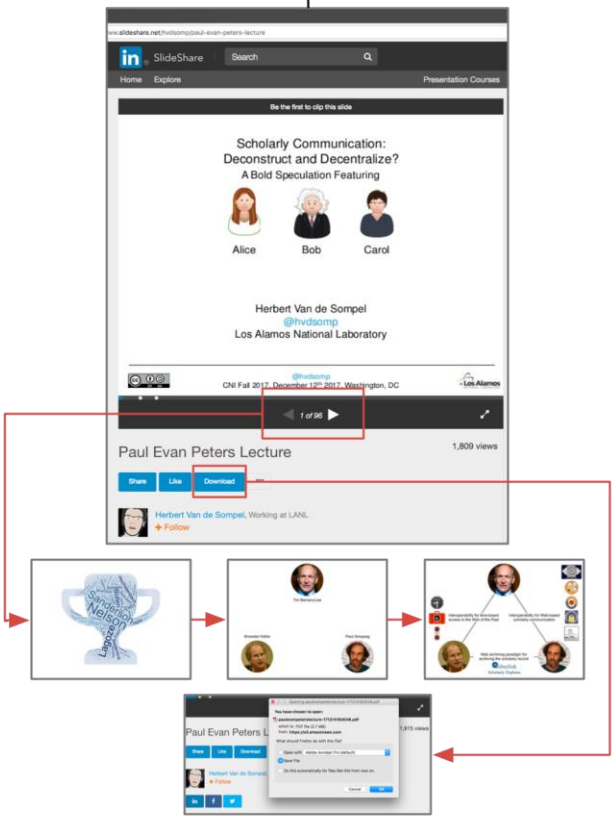
https://twitter.com/mart1nkle1n/status/1136705116738904067
http://blog.supdigital.org/completing-the-archives-or-how-were-extending-the-life-of-web-based-digital-scholarship/

Resource Boundary?

the page at URI-A is not the artifact

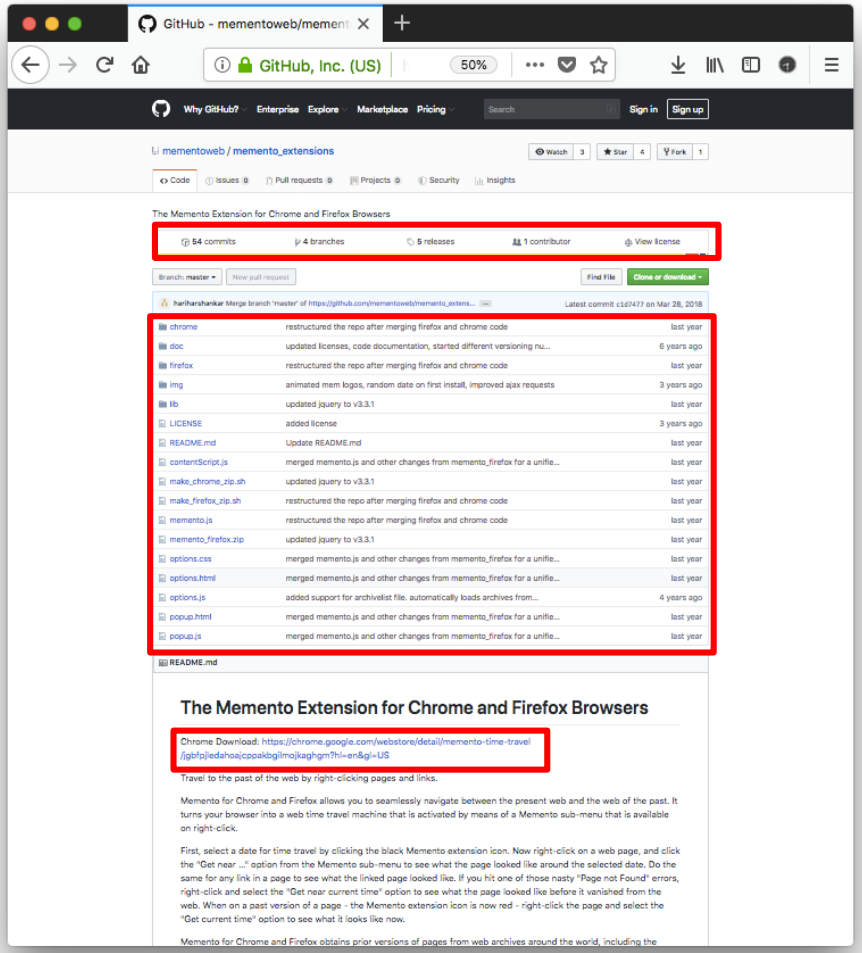


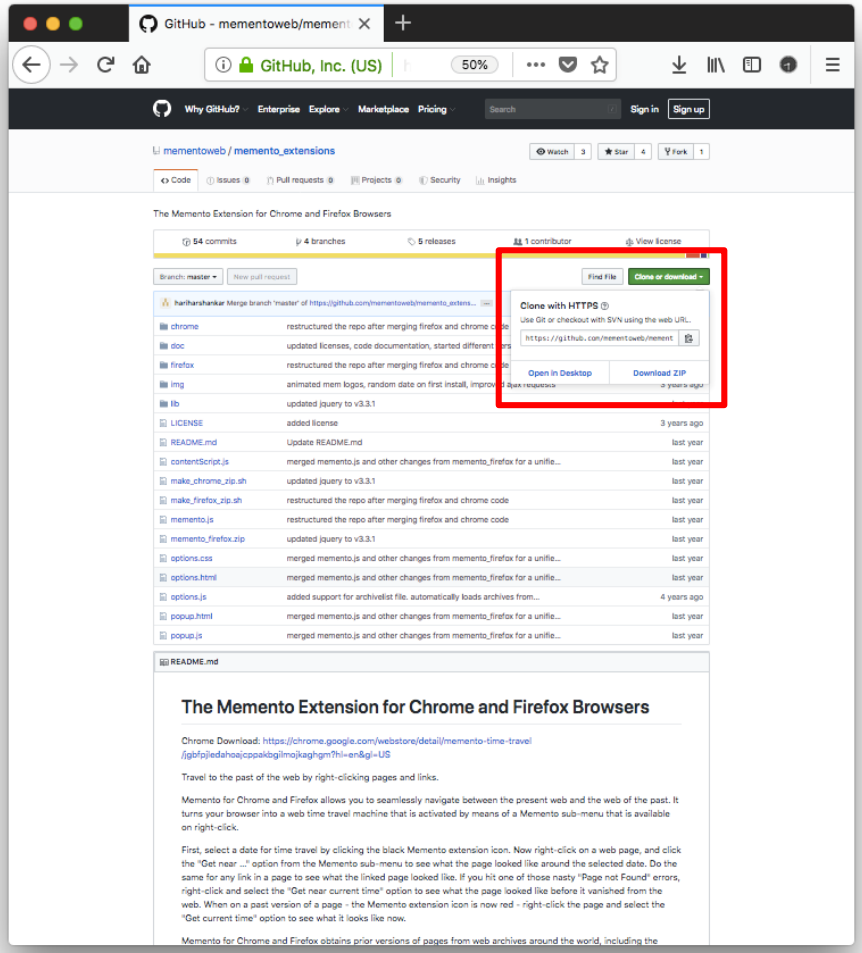
the artifact includes resources reachable via URI-A



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

Resource Boundary?





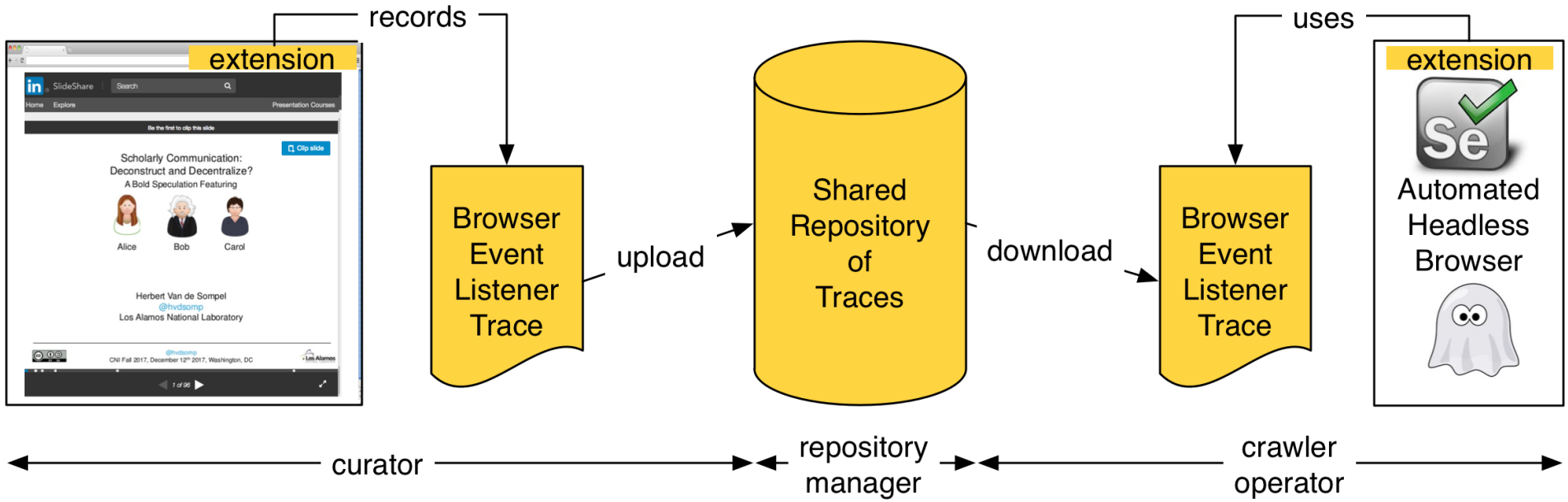
https://github.com/mementoweb/memento_extensions

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

Capture with a Trace

- Slideshare deck
  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

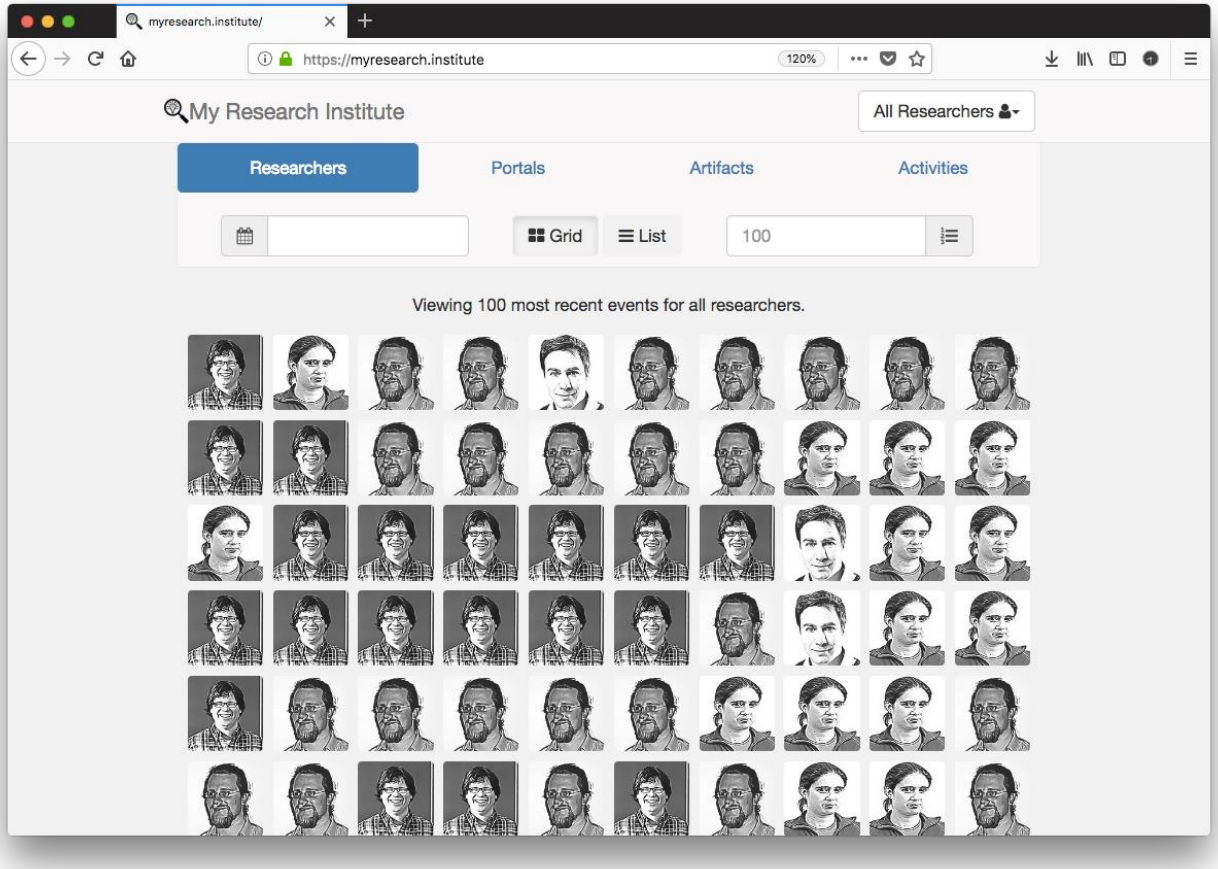
Memento Tracer Benefits

- 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 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

Herbert Van de Sompel
DANS
@hvdsomp

http://tracer.mementoweb.org/