Introducing the Memento Tracer Framework for Scalable High-Quality Web Archiving

Martin Klein
Los Alamos National Laboratory
USA
mklein@lanl.gov
https://orcid.org/0000-0003-0130-2097

Herbert Van de Sompel
DANS
The Netherlands
hvdsomp@gmail.com
https://orcid.org/0000-0002-0715-6126

Web archiving at scale and with high-quality is not a trivial endeavor. Given the dynamic nature of the web and the increasingly complex features incorporated into web pages, sophisticated capturing and archiving approaches are needed. This workshop introduces the novel Memento Tracer Framework and invites attendees to a hands-on experience with our framework that was designed to track, capture, and archive scholarly artifacts. Memento Tracer takes an institutional perspective and focuses on artifacts created by or relevant to individual communities. Workshop participants will learn about scalable and high-quality web archiving and leave with the satisfaction of having actively contributed to saving parts of the (scholarly) web.

Memento Tracer, Web Archiving at Scale, High-Quality Web Archiving
Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction and Background

Current web archiving approaches often either excel at capturing at scale or with high quality. Despite various attempts [1], approaches that combine scale and quality remain elusive. For example, the Internet Archive's crawler is optimized for scale and hence enables an archive of more than 703 billion web resources [2]. However, the quality of the captures varies is often hindered by dynamic elements and interactive features contained in the captured resources. For instance, the CNN.com homepage has not been properly archived (and hence can not be replayed correctly) in the Internet Archive since November 2016 [3].

An example on the other end of the spectrum is Webrecorder [4]. While browsing a web page, this tool archives the page and captures all the elements the user interacts with. With this approach, Webrecorder provides high-fidelity captures but lacks the ability to archive resources at web scale as only individual user interactions with the single web resource trigger the archiving process, similar to a screen recording session.

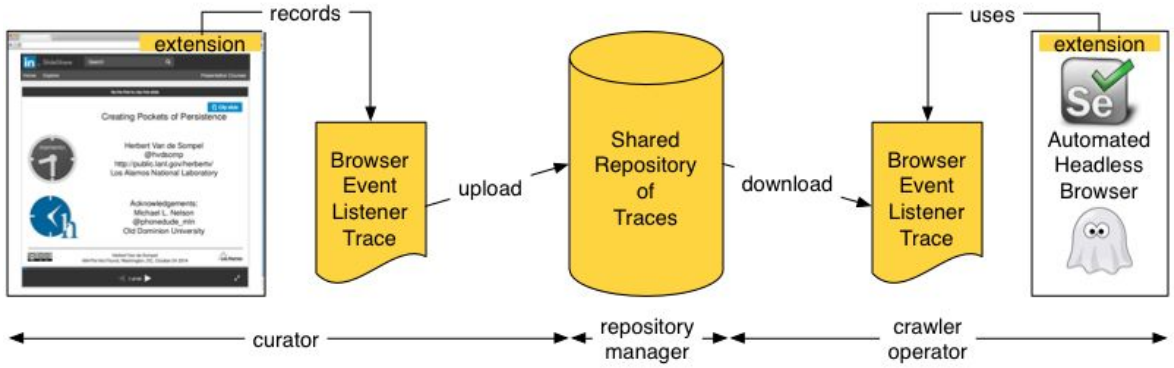
The Memento Tracer framework aims to find a balance between operating at web scale and providing high-quality archival records. The Tracer approach, visualized in the figure below, is based on a human curator interacting with a web resource to establish its essential components (boundary), and to record these interactions as a trace. This Trace will then be used to guide a browser-based capture process that automatically replays the recorded transactions and thereby delivers high-quality captures of web resources. A trace can be applied to all resources of the same class and hence enabling the framework to operate at scale. To enable transparency, foster collaboration, and avoid duplicate efforts, traces and versions of traces can be shared with a community of practice in a shared repository.

In the "Scholarly Orphans" project, we focus on archiving scholarly artifacts - web resources scholars across disciplines and throughout the research life cycle create in productivity portals such as GitHub, FigShare, Publons, and SlideShare to conduct aspects of their research and to communicate research outcomes. Hence, we designed Memento Tracer an institutional pipeline to track, capture, and archive these artifacts. The

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019



workshop will introduce the Memento Tracer framework to the participants and allow for hands-on exploration of the entire pipeline.

II. WORKSHOP TOPICS

The workshop will begin with a high-level overview of the Memento Tracer framework and outline its capabilities by means of intuitive and easy-to-follow examples. This introduction of the novel framework is necessary in order to provide attendees with the background and knowledge to fully take advantage of the second part of the workshop. Here we will collect web archiving use cases from participants and identify scenarios where and how Memento Tracer can be applied. The workshop participants are then encouraged to try out the framework themselves while the organizers will provide guidance and (technical) support. Participants will create traces for productivity portals such as Github or SlideShare, upload their traces to a shared repository, use their own or a publicly shared trace to capture and archive artifacts, replay the archived record, and provide feedback about its quality.

At the conclusion of the workshop, participants will have gained a deep level of understanding of various practical aspects of web archiving, identified archiving use cases relevant to their home organization, utilized the Memento Tracer framework for the creation of archival records related to their use case, and done a high-level quality analysis of the created archival record.

III. WORKSHOP LOGISTICS

This is a hands-on workshop where the attendees are encouraged to participate, ask questions, and provide feedback. We will keep the lecture portion to a minimum and allow significant time for exploration of the Memento Tracer framework and participants' use cases.

A. *Audience and Attendees*

This workshop aims to bring together librarians, archivists, and other professionals that are passionate about addressing challenges in modern web archiving.

Since this will be the first time we organize this workshop, we have no past data to rely on but we expect 20-30 people attending this workshop.

B. *Format and Duration*

We propose a half-day workshop, ideally in the morning, that does not exceed four hours. We plan on a short break around the halfway mark and will leave plenty of time for questions and feedback from the attendees.

C. *Special Requirements*

The workshop requires a space where attendees can sit down and use their laptops. Internet connectivity as well as power supply is essential. In addition, the organizers will need a projector in the room for slides and live demonstrations.

REFERENCES

[1] Squidwarc. https://github.com/N0taN3rd/Squidwarc
[2] https://twitter.com/brewster_kahle/status/1087515601717800960
[3] CNN.com has been unarchivable since November 1st 2016. https://ws-dl.blogspot.com/2017/01/2017-01-20-cnncom-has-been-unarchivable.html
[4] https://webrecorder.io/