Exploration of novel Web archiving strategies

Robert Sanderson, Lyudmila Balakireva, Harihar Shankar, Herbert Van de Sompel
(Los Alamos National Laboratory)

As part of the Memento (Time travel for the Web) project, we have started exploring alternative Web archiving strategies. Specifically, our focus is on archival approaches that do not rely on Web crawling but rather are based on a transactional paradigm: Web content is archived at the very moment it is being used. This presentation will report on the status of two ongoing activities in this realm: a client-side and a server-side transactional archiving approach.

In the client-side approach, a browser plug-in captures browser activity as it occurs and writes it to disk. The captured representations of visited resources are recurrently transferred to a Web-based archive. Both the transfer mechanism and the Web archive are based on a widely deployed tool, making the approach lightweight to adopt. The Web-based archive can remain private, allowing a user to revisit her precise browsing history using Memento's time travel approach. But a user can also opt to make her archive public, thereby extending the Web's archival collection. Research is being conducted to explore smart mechanisms to allow selective sharing of the Web resources visited by a user.

In the server-side approach, representations of Web resources that are served to a client are simultaneously pushed into an archive associated with the server. This approach, inspired by the PageVault and Vignette WebCapture tools, yields a fine-grained archive of content served over time. The archive is Memento-enabled and hence a server can redirect a Memento client to an accurate representation for a resource as it was served at a specified moment in time. If required, the content of a server's transactional archive can recurrently be offloaded to a regular Web archive.

It is our intention to publicly release both the client-side and server-side archiving tools. Further information about Memento is available at http://mementoweb.org.

memento

Memento is funded by the Library of Congress