# Web Historian: Enabling Multi-method and Independent Research with Real-world Web Browsing History Data

Ericka Menchen-Trevino[1]
[1]American University

**Abstract**
Research analyzing real-world web browsing data has generally been collected from digital service providers or the online panelists of corporate research panels. These approaches limit the replicability and the kind of work that can be done. Web Historian's first tool, a Chrome browser extension addresses this problem by enabling researchers to securely collect web browsing history data from participants with a robust informed consent process, and direct benefits to participants. Data visualizations of web browsing history inform participants of what they are submitting to the research project and help them gain further knowledge of their own browsing habits. Web Historian uses data that are already on the user's computer. Participants can submit up to 90 days of browsing history within just a few minutes. Since researchers can recruit participants themselves web browsing history data can be added to other forms of data collection, qualitative or quantitative. The visualizations can also be used for educational purposes.

**Keywords:** Methodology; digital trace data; mixed-methods
**doi**: 10.9776/16611

**Contact**: menchent@american.edu

## 1 Introduction

Digital technologies create records of their use as part of their normal functioning. These records are generally held by the companies that create the digital tools and services such as Facebook, Google, and many smaller providers. Some commercial research firms such as Knowledge Networks, comScore, and Nielsen collect digital traces from panelists. Academic researchers have generally relied on these sources to access and analyze traces of digital behavior. These are necessary and important partners, but it is critical to research replicability and academic freedom for researchers to have the ability to collect digital trace data themselves.

Paul Lazarsfeld, one of the most influential media researchers of the 20th century forged many partnerships with media industries. He said at a conference for media practitioners:

> [W]e academic people always have a certain sense of tightrope walking: at what point will the commercial partners find some necessary conclusion too hard to take and at what point will they shut us off from the indispensable sources of funds and data? (Lazarsfeld, 1941, p. 10-13)

Within the constraints he faced Lazarsfeld was able to do pioneering work in political communication and several other fields, as researchers who partner with corporations that provide digital tools and services do today. However, it is important to develop other approaches to safeguard academic freedom. Furthermore, digital service providers are not required to use an informed consent process. They are governed by terms of service agreements and privacy policies that are seldom understood by users, occasionally leading to outrage from technology users (Albergotti, 2014).

Partnering with commercial research firms who recruit and administer panels of participants who agree to take surveys and have their web browsing logged in return for incentives is one way to obtain data about digital behavior that does not come with the same risks to academic freedom. This approach does, however, involve a lack of transparency that cuts off the ability of others to independently replicate the research. The source code of the programs used to collect the digital traces is, to my knowledge, never provided by the companies and is protected as a trade secret. In most cases the details of the recruitment process, incentives provided, and any data cleaning or aggregation algorithms are also kept secret. Seemingly small details in how data are collected or cleaned can have a major impact on the analysis, so transparent alternatives are needed. Because digital trace data often includes potentially identifying information it is unlikely that the data itself can be released into open repositories, however, if the software used to collect it is openly available a study can be replicated.

Web Historian is an open-source web browser extension and planned mobile app that allows researchers to securely collect web browsing history data from participants with a robust informed consent process, and direct benefits to participants. Web Historian was designed to overcome the problems inherent to obtaining traces of web behavior via service providers and commercial research companies. Web Historian also facilitates integrating web behavior logs into multi-method research, including both qualitative and quantitative techniques. Web browsing observations by themselves can be quite difficult to interpret without further context, which could be provided by collecting survey data, experimental data, field observations or in-depth interviews. The Google Chrome extension described in this poster is the first browser from the Web Historian project, which will be adding browsers once the Chrome extension is tested further. Google Chrome is the most popular browser in most countries around the globe (StatCounter, 2015), so it was chosen as the project's starting point.

## 2 Digital Trace Data Collection in Web Use Research

Some researchers have collected real-world web browsing directly from participants previously using a proxy-based system called Roxy (Menchen-Trevino & Karr, 2012). Proxy systems do not work well on today's web since major websites such as Facebook and Google now use HTTPS security, which does not allow for logging via proxy.

## 3 System Design

Web Historian's Chrome browser extension was designed to quickly inform participants of what their browsing data contains using visualizations (see Figure 1-4). The visualizations serve two related purposes, informed consent and participant knowledge gain. Unlike studies of opinion or future behavior, the patterns of the past habitual behavior of web browsing may not be known by the participants themselves.

The visualizations are created using client-side JavaScript only and do not send any history data to the research project unless the participant opts in to the study. Those who do opt-in are then asked to take a study-related survey. Participants have the option to delete whatever data they choose before participating. Web Historian keeps track of how much information the user deleted, but not any details about the content.

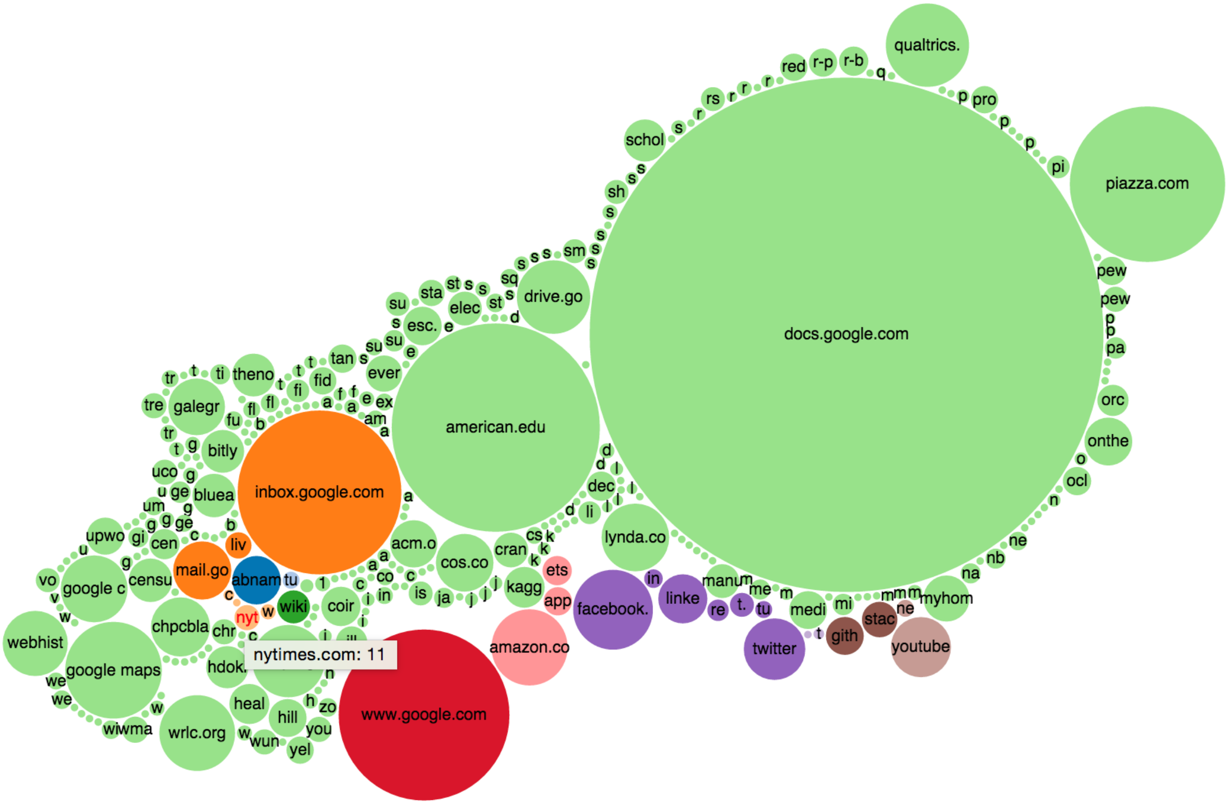
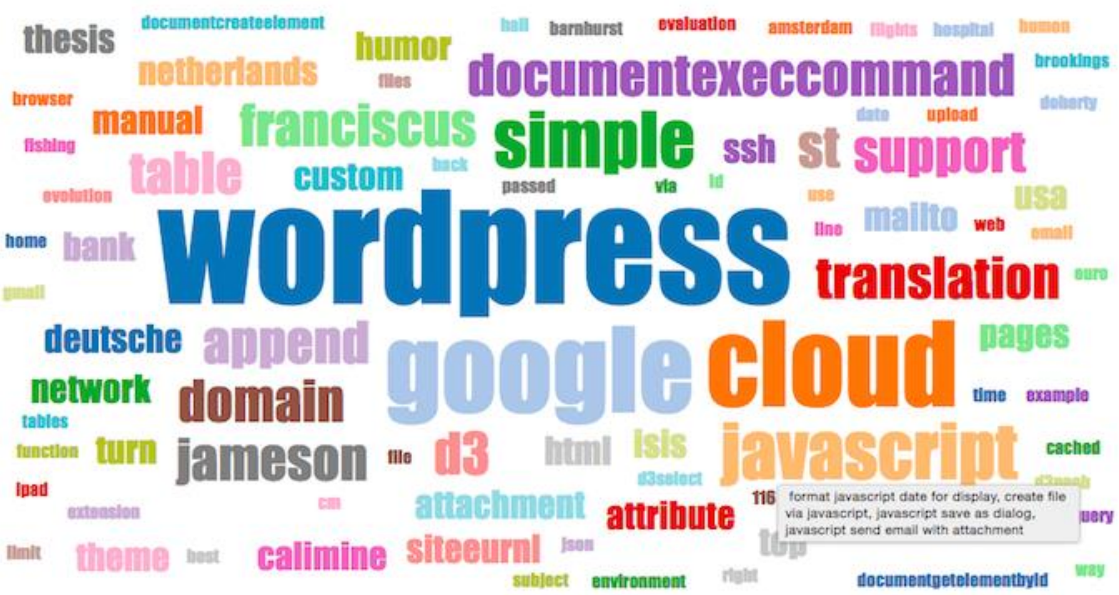


Figure 1. Websites Visited, 30 days: Circle size represents the number of visits to a domain. Tooltip identifies the domain name and number of visits.



format javascript date for display, create file via javascript, javascript save as dialog, javascript send email with attachment

Figure 2. Search Words, seven days: Word size increases when it is used in different search terms. Tooltip shows the search terms in which the word appears. Word color is arbitrarily selected from a color palate.

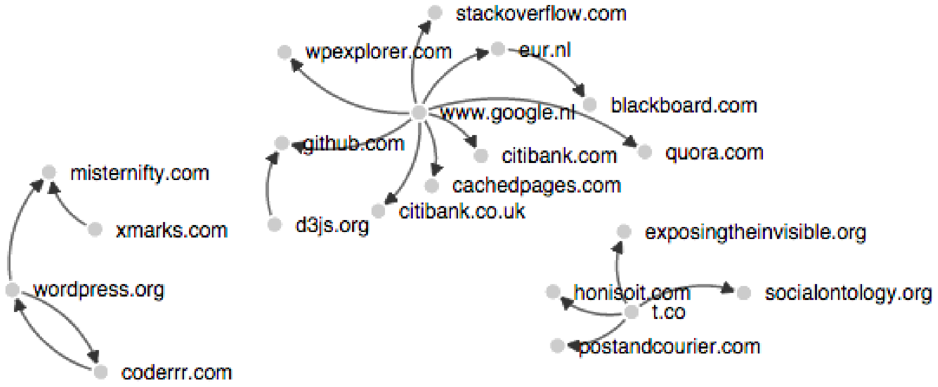
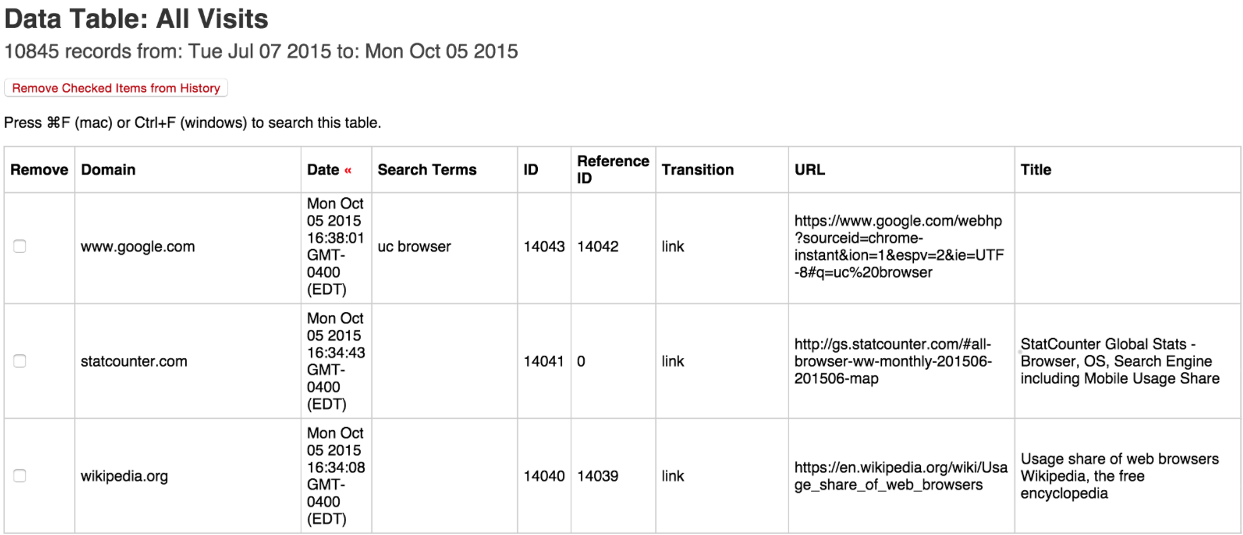


Figure 3. Network, three days: Arrows connect domains where the user browsed from the origin to the destination domain. Hovering over the node or label allows the user to reposition the node manually.



**Data Table: All Visits**

10845 records from: Tue Jul 07 2015 to: Mon Oct 05 2015

Press ⌘F (mac) or Ctrl+F (windows) to search this table.

| Remove | Domain | Date « | Search Terms | ID | Reference ID | Transition | URL | Title |
|---|---|---|---|---|---|---|---|---|
| □ | www.google.com | Mon Oct 05 2015 16:38:01 GMT-0400 (EDT) | uc browser | 14043 | 14042 | link | https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=uc%20browser | |
| □ | statcounter.com | Mon Oct 05 2015 16:34:43 GMT-0400 (EDT) | | 14041 | 0 | link | http://gs.statcounter.com/#all-browser-ww-monthly-201506-201506-map | StatCounter Global Stats - Browser, OS, Search Engine including Mobile Usage Share |
| □ | wikipedia.org | Mon Oct 05 2015 16:34:08 GMT-0400 (EDT) | | 14040 | 14039 | link | https://en.wikipedia.org/wiki/Usage_share_of_web_browsers | Usage share of web browsers Wikipedia, the free encyclopedia |

Figure 4. Data Table, three visit records: Checking the box in the remove field allows users to remove records by pressing the “Remove Checked Items from History” button. Sorted column is indicated by a red arrow.

## 4 Implementation

A small-scale test of Web Historian has been performed in the Netherlands in the spring of 2015. A convenience sample of 11 Dutch citizens age 18-29 were recruited and were paid five euros for submitting their data. Most of the participants (82%) submitted the maximum possible history length of 90 days. Google Chrome keeps 90 days of browsing history by default and previous research suggests that default settings are quite powerful influences on behavior (Shah & Sandvig, 2008), but I am not aware of published research on how often web users clear their browsing history.

The number of visits to websites ranged from 20,983 to 2,079 with an average of 11,454 visits logged for each participant with a total of 125,994 visits total. Of the eight participants for whom deletion logging was enabled three choose to delete some of their history. The highest percentage of deleted URLs for an individual user was 0.8%. A larger and more representative sample is need to estimate average browsing history length and rates of record deletion.

Nine of the 11 participants subsequently completed an in-depth interview about their political information consumption, including a review of their Web Historian visualizations. Although most found their Websites Visited and Search Words visualizations interesting and quite informative about their browsing habits in general, typically only a small amount of this information was relevant to political information consumption. Rapid filtering and categorization of potentially relevant sites are planned features to address these issues. Some were confused by the network visualization and what it was meant to represent. The visual clutter of the node labels has been improved, and new visualization ideas for web browsing paths are in development.

## 5 Conclusion

The Web Historian Chrome extension does address important problems researchers face in studying real-world web behavior by providing a mechanism for collecting the web browsing history data that is already stored in participants’ web browsers. The informed consent process is truly informative such that participants can directly benefit from learning more about their own web browsing routines. While this short paper has focused on social scientific uses of web histories, it could also be part of a historical digital humanities project if the data were available.

Since the browsing data submitted through Web Historian is created prior to participating in the research project this minimizes observation effects.

While the Google Chrome browser is the most popular in the world (StatCounter, 2015), individuals, particularly youth, tend to use multiple devices and multiple browsers (Millennial Media, 2014). Further tools must be added to the Web Historian project to collect and integrate web browsing history data from more sources. While additional tools will increase coverage, they may increase participant burden as well. Web browsing logs for large and diverse populations will likely be partial for the foreseeable future. When research questions concern relative uses and propensities rather than the absolute frequency of a behavior this may not be particularly problematic. Current information about the Web Historian project is available at http://webhistorian.org/, and the source code is available at https://github.com/erickaakcire/webhistorian.

## 6 References

Albergotti, R. (2014, July 1). Furor erupts over Facebook's experiment on users. *Wall Street Journal*. Retrieved from http://www.wsj.com/articles/furor-erupts-over-facebook-experiment-on-users-1404085840

Lazarsfeld, P. F. (1941). Some notes on the relationships between radio and the press. *Journalism Quarterly*, *18*, 10–13.

Millennial Media. (2014). *Cross-screen consumer behavior: Decoded*. Baltimore, MD. Retrieved from http://www.millennialmedia.com/download/13

Menchen-Trevino, E., & Karr, C. (2012). Researching real-world Web use with Roxy: Collecting observational Web data with informed consent. *Journal of Information Technology & Politics*, *9*(3), 254–68. http://doi.org/10.1080/19331681.2012.664966

Shah, R. C., & Sandvig, C. (2008). Software defaults as defacto regulation - The case of the wireless internet. *Information, Communication & Society*, *11*(1), 25–25. http://doi.org/10.1080/13691180701858836

StatCounter. (2015). StatCounter global stats: Top browsers per country, June 2015. Retrieved November 30, 2015, from http://gs.statcounter.com/#desktop-browser-ww-monthly-201410-201510