Analysing and understanding news consumption patterns by tracking online user behaviour with a multimodal research design

Martijn Kleppe and Marco Otte
National Library of the Netherlands (KB), The Netherlands and Vrije Universiteit Amsterdam, The Netherlands

Abstract

Understanding people's online behaviour has traditionally been a field of interest of commercial research agencies. However, academic researchers in a variety of fields are interested in the same type of data to gain insights in the Web behaviour of users. Digital Humanities scholars interested in the use of digital collections are, e.g., interested in the navigation paths of users to these collections. In our case we wanted (1) to analyse the way news consumers visit news websites and (2) understand how these websites fit in their daily news consumption patterns. Until now most common applied scholarly research methods to analyse online user behaviour focus on analyses of log files provided by website owners or recalled user behaviour by survey, diary, or interview methods. Only recently scholars started to experiment with gathering real-world data of Web behaviour by monitoring a group of respondents. In this article we describe the set-up of 'The Newstracker', a tool that primarily allowed us to analyse online news consumption of a group of young Dutch news users on their desktop and laptop computers. We demonstrate the workflow of the Newstracker and how we designed the data collection and pre-processing phase. By reflecting on the technical, methodological, and analytical challenges we encountered, we illustrate the potential of online monitoring tools such as the Newstracker. We end our article with discussing its limitations by stressing the need for a multimethod study design when aiming not only to analyse but also to understand online user behaviour.

Correspondence: Martijn Kleppe, Prins Willem-Alexanderhof 5,
2595 BE The Hague
Netherlands.
E-mail:
martijn.kleppe@kb.nl

1 Introduction

Understanding users' online behaviour is of growing interest to academic researchers in a variety of fields. Traditionally, in the marketing domain, commercial research companies map consumer behaviour to understand when and where customers decide to buy products. For this purpose, Web metrics of individual websites serve as detailed source of information on when, how, and at which section a user enters a website. Recently this type of data is also being used by cultural heritage institutes to understand the interest of their visitors (De Haan and Adolfsen, 2008) to track where their digital content is being reused (Navarrete Hernández, 2014) or to understand the query's users perform in search

doi:10.1093/llc/fqx030 Advance Access published on 29 June 2017

systems by analysing the log files (Batista and Silva, 2002; Huurnink, 2010). In this type of research, the website is the central research object providing traces that Menchen-Trevino (2013) calls 'Horizontal Data sets'. These contain data that are 'organized around a specific type of trace, for example search terms, web browsing log files, tweets, hashtags, likes or friend and follower ties' (Menchen-Trevino, 2013, p. 331). An advantage of using this type of data is that they are not obtrusive to the respondents, since they are created automatically as users are surfing the Web. However, this also leads to an ethical disadvantage, since users are not aware that their online behaviour is being examined, nor could they give their consent to have their data being analysed. While Horizontal data sets are organized around one type of trace, Vertical data sets are organized around research participants that deliberately 'give permission for researchers to collect their digital traces' (Menchen-Trevino, 2013, p. 331).

Since mid-1990s of the previous century, commercial research agencies have started to collect these types of vertical data by building tools and panels of respondents whose online behaviour is monitored 24/7 to provide data on usage across media and purchase behaviour (Coffey, 2001; Napoli, 2010; Taneja and Mamoria, 2012). In the USA, companies such as comScore and Nielsen created 'Online Netview Panels',[1] while in The Netherlands, TNS Nipo and Wakoopa offer similar tools to create aggregated lists of the most visited websites on all platforms and devices.[2] Similar to television viewing rates, these lists are mainly created to gain more insight in the background of website visitors to provide potential advertisers with information on how to reach their online target audience in the best possible manner. Obviously these commercial research data contain very rich information, also for academics who are interested in collecting real-world Web use data. However, apart from lists of the most popular domains that are published as open data by companies such as Alexa and Similarweb,[3] data containing information about visits to each individual page and information about the background of the panel are not available. Main arguments of commercial agencies to not collaborate with scholars are to ensure the confidentiality of their respondents' identity and to prevent scholars to gain insight into the techniques applied by the companies.

Nevertheless, researchers in a variety of academic disciplines are interested in tracking online behaviour outside of a laboratory environment. Especially in the communications science realm, scholars experimented with several techniques of tracking people's online behaviour. To study the changing attitudes and opinions of Internet users, Ebersole (2000) used a relatively labour-intensive method by manually analysing the Web browser history of a group of students after each logged out of the computer. Tewksbury (2003) used public click data of a research company to compare its outcomes with survey results on the perceived consumption of public affairs news websites by the respondents. Findahl (2009) analysed the use of Internet in Sweden by using the commercial package PacketLogic[4] that can identify over 1,000 Internet application protocols (Findahl *et al.*, 2013). Munson *et al.* (2013) used a Browser widget to monitor visited websites to study selective exposure theory of political websites. Van Damme *et al.* (2015) focused on the multiplatform consumption of news by analysing mobile device logs, using monitoring software that collected the applications that had been activated, the location where the device was used (by using Global Positioning System co-ordinates) and which websites had been visited. While these studies hardly report on the privacy of their respondents, Menchen-Trevino and Karr (2012) gave her respondents an active voice in the logging process of their data which is in line with Lotz and Ross (2004) plea on finding a balance between online audience studies and the respondents privacy. To monitor the exposure to political communication during the November 2010 US general election campaign, Menchen-Trevino designed the software package Roxy[5] that collects Web-use data from participants with informed consent. During the logging process, respondents could log in the software server to go over the list of registered Uniform Resource Locators (URLs) and select websites he or she did not want to be logged for privacy reasons.

Striking about these pioneering monitoring studies is its multi-method approach. By default each does not only monitor Web use but also compares its outcome with either survey, diary, or interview data. By triangulating the results, these researchers try to overcome the critique on classic studies on media consumption that often deploy either surveys or diaries registering self-reported media behaviour (Reuters Institute for the Study of Journalism, 2015; Schrøder and Kobbernagel, 2010; Taneja *et al.*, 2012; Van Cauwenberge *et al.*, 2010). These methods strongly rely on the memory of the participants, while several scholars found respondents often overestimate their media use (Ebersole, 2000; Prior, 2009; Robinson, 1985). Furthermore, since filling in diaries and surveys on news consumption is very labour-intensive, its outcomes mainly focus on when media have been consumed or on which devices, while it remains unclear what has been consumed. One way to gain insight in the consumed news content is by focussing on metrics of individual news organizations (Batista and Silva, 2002; Boczkowski, Mitchelstein and Walter, 2011; Lee *et al.*, 2012; Usher, 2013) or the most clicked items (Boczkowski *et al.*, 2011; Karlsson and Clerwall, 2013; Lee *et al.*, 2012; Nederlandse Nieuwsmonitor, 2013). However, given the focus of these studies on individual websites or most-clicked articles, it remains unknown which genres of news websites constitute users' 24/7 news menu. Taneja *et al.* (2012) tried to overcome this problem by literally following 495 users throughout an entire day. However, even with this labour-intensive fieldwork, it proved not to be possible to incorporate the genres of consumed news items.

However, Web-monitoring tools, such as the above-mentioned examples, now offer a less labour intensive and more precise way of registering digitally consumed news items. By deploying these techniques, we could overcome the knowledge gap of the 24/7 news consumption menu. Therefore, we created the Newstracker. This is a monitoring tool we specially designed to (1) track the consumption of news 24/7 on laptop and desktop computers of a selected group of respondents and (2) to analyse the role of news in everyday browsing behaviour. In this article we do not aim to present substantial results but describe the set-up of 'The Newstracker' by illustrating the potential and limitations of the tool by its use in a multimethod study on the online news consumption of a group of young Dutch news users. When we describe results, they stem from another article that focuses explicitly on the research outcomes of the Newstracker and our study on understanding online news consumption plus their theoretical implications (Kleppe and Costera Meijer, under review). In this article, the described outcomes only serve to illustrate the possibilities and limitations of tracking tools like the Newstracker.

2 Designing the Newstracker

The Newstracker is a custom-built system that collects Web activities of specified and authenticated users, cleans the data by removing non-relevant data, extracts the associated content, and stores this as a new data set to be used for analysis. To create the Newstracker, we applied two out of several data mining techniques: Web usage mining and Web content mining (Liu, 2011; Srivastava *et al.*, 2000). Web usage mining aims to explore which user visits which Web pages on which moments through time. This gives an insight into what the user might be interested in, frequency of visits to particular Web pages and patterns in use. Web content mining allows us to actually see the content the user requested in the form of text and images that can be further analysed and linked to the usage data. To be able to analyse these users' online activities, they need to be captured and transformed into relevant and meaningful data. Mobasher (2007) identifies three phases in this process: data collection, pre-processing, and pattern discovery. In this section we will address how we implemented the first two. In Section 3, we will describe how we discovered the patterns in our data.

2.1 Data collection

Figure 1 shows the workflow of the Newstracker application and how we designed the data collection and pre-processing phase. To collect the data in the first phase, similar to Menchen-Trevino and Karr

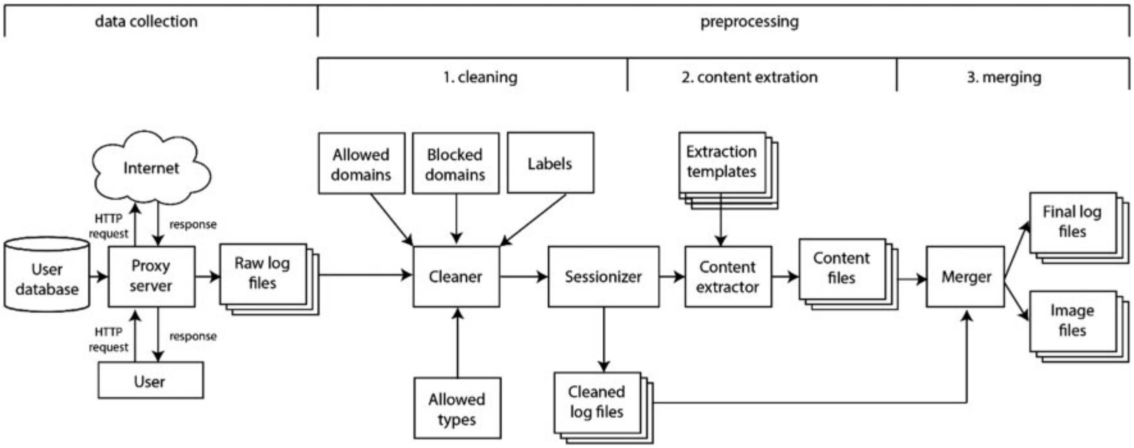

Fig. 1 Workflow of the Newstracker application, illustrating the two main phases: data collection and pre-processing. The latter consists of three stages: cleaning, content extraction, and merging

(2012), we employed a local proxy in the respondents' Internet browser. With this proxy, all Internet traffic was processed through a central proxy server, allowing us to register all visited URLs. In one of the first papers on commercial Internet Audience Measurement, Steve Coffey (2001), vice president of Media Metrix (now part of comScore), already described the use of a proxy as one out of three types of meters to measure online behaviour. Commercial agencies often use an Operating System meter that is in the respondents' deep layer of the system, allowing to keep track of all the activity on the PC. Though the data gathered are very rich, their developments are very costly and hardly feasible for academic researchers. The second type of meter is the Central Office Proxy that registers all file requests to the Internet through a central office. This type of proxy does not only record Web browsing behaviour but all programs running that use Internet such as Torrents, Instant Messaging Services, etc. Since we were only interested in the Internet browsing behaviour of our respondents, we used the third type of meter, namely, the Local Proxy. This sets up a virtual proxy server in the respondents' Internet Browser, records all file requests, and then passes them on to the Internet. In our case we used WinGate version 8.3.4[6] that allowed us to give each respondent personal log in credentials, so we knew for sure we were monitoring the respondents' browsing behaviour and not the clicks of someone else who was using the same computer. Each respondent received the user account and needed to configure his browser to use the proxy server. Since they did this themselves, they were also capable of shutting it down when they did not want the researchers to monitor their Web behaviour. To guarantee the respondents' privacy, we deliberately pointed them to this possibility. However, none of the respondents eventually switched off the proxy. Once the proxy was installed, all Internet traffic [over ports 80 HyperText Transfer Protocol (HTTP) and 443 HTTPs] was processed through the proxy server which was located at the premise of the Vrije Universiteit Amsterdam. Each so-called HTTP request is logged by the proxy server in a daily log file and contains sixteen fields: date of log entry, time of log entry, username, session ID, time taken for the request (ms), client IP address, client agent (browser type), server IP address, request method, URL (website address) request, server status, bytes sent by client, bytes received from server, protocol used, and referrer URL.

With this workflow, the proxy server created a log file in csv-format every day, consisting on average of over 150,000 entries. However, most of these entries were not relevant for our goals. In our case, less than 1% of the raw information contained relevant information about which Web pages the user was visiting. The rest consisted of references to scripts, external HTML code and images, advertisements add-ins, Cascading Style Sheet pages, etc. During the first stage of the pre-processing phase, the raw information thus needed to be cleaned, which is marked as the most time-consuming and computationally intensive step in Web usage mining (Mobasher, 2007).

2.2 Pre-processing

To make this cleaning phase as efficient as possible, we created a stand-alone C# program that consists of four modules: the cleaner, the sessionizer, the content extractor, and the merger.[7] The program was scheduled to run each night after the proxy server closed the previous day log file. First the cleaner module compared each log entry's URL request to a white list of permitted URLs. The reason to use such a white list is two-fold. First, given the near-infinite number of Web URLs, researchers who are monitoring Web users' behaviour for a specific research aim, must first set the universe of websites they are interested in by formulating a white list, containing these websites (Coffey, 2001). Otherwise they run the risk the data they are gathering is of such a size, they are incapable of making general conclusions. In our case, we were interested in the visited news websites and how these websites are part of the respondents everyday browsing behaviour. We therefor did not only put news websites on our white list, but we created a list of 3,564 websites that were most popular in The Netherlands according to several commercial research agencies (Kleppe and Bijleveld, 2017, DANS)[8]. This list thus contained news websites but also search engines, shopping websites, social media, etc. Since our research aim was to conclude what types of websites people visit 24/7, we added predefined labels to each URL, ranging from Shopping to Government and from Weather to News and Information. The latter category was further drilled down in subgenres such as General or Conventional news, Lifestyle, Sports, etc. Our second reason to use a white list is for privacy reasons of our respondents. Since we did not want to monitor their browsing behaviour to illegal or adult websites or their Web activities on websites they first had to log into, these were not included on the list. Furthermore, we offered our respondents the possibility to inspect the website domains that would be monitored. Contrary to Menchen-Trevino and Karr (2012) who allowed users to deselect registered websites afterwards, we deliberately offered our respondents to inspect the white list beforehand to give them the possibility to decide whether they would be willing to participate.

After the Cleaner Module determined if the URL was allowed according to the white list, the component type was checked. A URL can point to a large number of component or file types on the server, most common are the .html and .php types. In our case we allowed the component types: html, htm, dhtml, shtml, php, asp, and jsp. So if a logged URL passed the domain and component type check, it was written to a cleaned-up version of the log file, and other entries were ignored.

After this first cleaning step, the sessionizer module started. This module checked for duplicate entries within one session ID. Web activities are structured into sessions in which a user accesses a certain set of remote data. Items retrieved within the same session are seen as duplicate entries and need to be removed from the cleaned log file, which is called sessionization (Mobasher, 2007). We also determined that multiple entries of a URL within 5 min of the previous request without any other activity in that time would be marked as a duplicate entry. Only the first entry was kept in the final cleaned log file.

In Stage 2 of the pre-processing phase, the content of the URLs visited was extracted in the Content Extractor module. Specifically we wanted to extract the Title, Introduction text (if available), main text, and main image of the requested news item to experiment with automated content analyses of the news article in a later stage. However, these types of analyses were out of the scope of the research stage this article reports on. Extracting this type of information is not as straightforward as it sounds. The structure of a Web page is usually extremely complex, often generated by content management systems resulting in many, often very similar, levels of hierarchy of HTML tags. This complex structure makes it very hard to automatically find the relevant content to extract. Furthermore, each website has its own structure, increasing the complexity of content extraction over multiple websites. To solve this, we opted for using existing software for content extraction (Web Content Extractor version 7.2[9]). This application allows the construction of content templates per website to extract precisely the information needed. For a sub-set of predetermined news websites, we created those extraction templates.

The Content Extractor processed each entry in the cleaned proxy log file, and if a valid Extraction Template was available, it instructed the Web Extraction software to perform the content extraction. It then collected the extracted data, consisting of text and an image, and stored those at the location where the results of the Newstracker were collected. The extracted text included a reference to the image, which was stored in a domain-specific folder for later retrieval during future analysis.

Finally, in Stage 3 of pre-processing phase, the Merge Module combined the extracted content and the cleaned proxy log file into one final log file that holds all the relevant data for further analysis. The resultant log files and images were then ready for analysis.

3 Analysing Newstracker Data

We designed the Newstracker to study how the consumption of news websites fits in the daily surfing behaviour of university students. We tracked the Web behaviour of forty-two university students who used their laptop as their main informational device and agreed to have their browsing behaviour on their laptops being monitored in the period April–July 2015. We found the respondents through the personal network of our research assistants. Each assistant was the main contact person for around twelve respondents whom they did not know personally. They were in touch with them several times a week, creating mutual trust, guaranteeing their privacy, and preventing the respondents to quit their participation. Each respondent signed a consent form and filled in a survey about their personal background and current news use. At the end of the tracking period, they filled in an exit survey which results indicate the Internet speed was not affected by the implementation of the proxy in the respondents' browser. They did indicate they were aware of the tool running when starting up their browsers but soon forgot about it also because we had very limited breakdowns of the server which we were able to fix soon. To gain a better understanding of the registered browsing behaviour, we held in-depth interviews with twenty of our respondents.

4 Results

After the pre-processing phase of the gathered data, we had 16,162 relevant and labelled URLs, giving us insight in the daily surfing pattern of our respondents. Our results (Kleppe and Costera Meijer, under review) show that website selection depends on time of the day. Shopping websites, e.g., are increasingly being visited during the day, while News and Information websites are steadily being visited all day long. We also found that the subgenres of News and Information websites are a determining factor in everyday surfing behaviour. General and Conventional News websites are, e.g., visited during the whole day, while traffic to websites with Lifestyle news increases during the day.

In this section, we would like to elaborate on the results we found when we analysed how our respondents navigated to news websites. Since our Newstracker did not only register what sites people visited but also on which day and at what time, we manually analysed in which order each news website had been visited. In total we analysed 1,834 websites to determine if a news article had been read after the respondents first visited the homepage or whether they used a link to a specific news item referred by, e.g., a personal e-mail, newsletter, or post on social media like Facebook or Twitter. Analysing these types of navigation paths of Web users is especially relevant within the journalistic field, given its current debate on the 'death of the homepage'. This refers to the allegedly decreasing role of the homepage as means of

Table 1 A selection of the data we registered from respondent Liv (female, 19 years old)

Date	Time	URL
26 May 2015	16:23:09	http://www.rtlnieuws.nl/boulevard/entertainment/video-liefdesuiting-carlo-brengt-irene-tranen
26 May 2015	17:09:14	http://www.rtlnieuws.nl/nieuws/opmerkelijk/professor-cupmaat-vrouw-onbelangrijk-voor-man
26 May 2015	17:09:22	http://www.rtlnieuws.nl/nieuws/opmerkelijk/waarschuwing-zwangere-handen-van-masturbatie
26 May 2015	21:09:40	http://www.cosmopolitan.nl/lifestyle/news/a146866/struggles-die-alleen-sarcastische-personen-begrijpen/
29 May 2015	12:13:47	http://www.lindanieuws.nl/nieuws/interview/9-brutale-vragen-aan-tuinman-tom/
29 May 2015	12:30:25	http://www.upcoming.nl/spacyfiek/12400/met-deze-10-prijskaartjes-van-de-jumbo-is-iets-geks-aan-de-hand
29 May 2015	12:30:28	http://www.upcoming.nl/experimenter/view/11842

generating traffic to websites, compared to social media platforms that are expected to increasingly generate Web traffic (Barthel *et al.*, 2015; Kopf, 2015; Somaiya, 2014).

See Table 1 for a snippet of the tracking data of our respondent Liv (female, 19 years old). On 26 May, 16:23 o'clock, she went straight to a news item on the website of RTLnieuws.nl, a major Dutch news television show of a commercial broadcaster. Since she did not first go to the homepage of RTLnieuws.nl, we can assume she ended up at the website via a referrer. We observed the same type of navigation behaviour at 21:09 when she visited a news item at Cosmpolitan.nl, on 29 May at 12:13 o'clock when she visited Lindanieuws.nl (MK: A website of a Dutch celebrity who also published her own magazine and website) and at 12:30 when she visited an item at Upcoming.nl.

Unfortunately, we were not able to see via which referrer Liv came to visit the websites of RTLnieuws.nl, Cosmopolitan.nl, Lindanieuws.nl, or Upcoming.nl, since we did not register social media data. Social media like Facebook or Twitter are visited by the respondent after they automatically logged in. As mentioned before, for privacy reasons, we decided not to monitor the contents of websites after our respondents logged in. This decision is a clear privacy trade-off, since it limits our data set because most users log in automatically to social media. We therefor even do not know if they went to, e.g., Facebook.com, since they automatically bypass the homepage. However, our interviews do give an indication on the role of social media in the online consumption process of news. When we confronted Liv with our data and asked how she ends up at news websites, she was very clear: 'I think I mainly come on those websites through Facebook. (...) I liked pages of several news outlets but daily, I also see news items on Facebook that have been shared by my Facebook friends'.

Since we were able to analyse of each news item how the respondent navigated to it—either through the homepage or via a referrer—we found that 59% of the articles that were read on News and Information websites were clicked on after a respondent first visited the homepage. This could indicate the homepage is still the main driver of traffic to news websites. However, when we analysed the traffic to the subgenres of News and Information websites, we found that mainly General or Conventional News Websites received their traffic via the Homepage. Websites containing Lifestyle and Remarkable news are mainly visited via referrals (Kleppe and Costera Meijer, under review). These types of results could indicate visitors of General News Websites are, on the one hand, more deliberate news consumers, since they purposely go to the homepage. Visitors of Lifestyle and Remarkable news websites, on the other hand, might accidently visit these websites after someone reposted a news item on Twitter or liked it on Facebook as the above described quote of Liv possibly illustrates. Though it is tempting to make these conclusions, the type of data the Newstracker provides does not legitimize these speculative statements, showing an important limitation of Web-monitoring tools as research methods.

Again, our interviews allowed us to nuance this outcome. When confronting our respondents with their observed browsing pattern, we found they often visited the same types of websites but for different reasons and in different manners. Websites containing Lifestyle news, e.g., were visited for several reasons. One respondent stated in the interview that she sometimes visits a Lifestyle website because a friend of her shares a news item on Facebook and that is the only way for her to read articles on this website while taking a break. However, another respondent visited the same types of websites almost daily but with a different purpose and different surfing pattern. In her case, fashion was a big hobby, and she always wanted to stay in touch with the latest Lifestyle news (Kleppe and Costera Meijer, under review). These types of nuances in Web browsing behaviour cannot be observed when only the registered Web behaviour is being analysed. Monitoring data shows what people do. However, we realized these outcomes should not be taken for granted but should spark researchers' curiosity to try to understand why users visit those websites and navigate the way they do, showing the urgency of a multimethod approach when monitoring online Web behaviour. Not so much as Ebersole (2000) or Findahl (2013) do by making a methodological comparison of the results of tracking tools with surveys, but to supplement the results found in monitoring studies. In our case the interviews provided us with two fundamental insights. First, we were able to understand how people navigate to news items via referrals. For privacy reasons, we decided not to monitor those websites via the Newstracker, but the interviews gave us the right information to understand how people use referrers. Secondly, tools like the Newstracker mainly register what people do, in our case also in terms of genres of visited websites. However, our interviews show that a click on a website does not necessarily mean someone is interested in that particular website in the same way as somebody else is. Therefore, in our situation the interviews function as explanation tool for the registered behaviour in the Newstracker.

Setting up such a multimethod design can obviously only be done with a relatively small group of respondents, given the labour-intensive nature of conducting in-depth interviews. We are of course aware that especially commercial research agencies are mainly interested in the big data nature of large-scale monitoring studies to give detailed information of website visitors to the owners of the websites. However, as our research findings suggest, also those types of studies should be aware that a website click does not automatically reflect a uniform interest of the visitors.

5 Discussing the Newstracker

Our multimethod design enabled us to first register online browsing behaviour and then find explanations for the registered browsing patterns. Though this allowed us to understand more fully how the consumption of news websites fits in the daily surfing behaviour of university students, setting up this research design was not trivial. Therefor we finish by elaborating on the technical, methodological and future analytical challenges for researchers who also want to conduct online monitoring studies.

5.1 Technical challenges

While designing the Newstracker, we encountered several issues particularly during the pre-processing phase. First, caching is cited as an issue for reliably collecting user Internet use (Spiliopoulou, 2000). When using a cached Web page, a browser will not request the page or elements of the page and thus not produce the HTTP request needed for the proxy server to log activity. However, reopening or reloading a page in a browser does force the browser to look up and compare the current version in the local cache and the version online. This 'version check' will output an HTTP request that is logged by the proxy server, even though it might not result in actually downloading all the content of the page again if the cached version is still current. Since we only are interested in the requested Web page, the caching problem did not play any role in this research but might be taken into account for other researchers who consider using monitoring tools. There is however one situation where caching does come in to play and that is when a user hits the 'back' key in the browser (Chitraa and Thanamani, 2011). Usually the browser will not reload the page, but retrieve it from its local cache without any online activity. Those re-visits are lost when only using a proxy server to collect the data. This issue could be solved by adding local tracking software on the user's computer that tracks mouse and keyboard input, but this would vastly increase the complexity of the analysis and create an extra privacy issue.

In the cleaning stage we encountered a second issue, since the Newstracker did not deliver a perfectly clean output. As a URL ending with a forward slash (/) is possible, and this will load the index.html/php page of the folder, we decided to include URLs that ended in just a forward slash. These URLs however do not always point to any user-relevant components, but the URL is still kept in the cleaned log file. It is only at the content extraction stage when no relevant content can be retrieved that it becomes clear the URL is not valid for our data set. This is now solved at the final stage of the Newstracker by generating a final log file that holds only those entries that do have content.

A third issue was encountered during the extraction stage, since the content of any given website is fraud with problems. Not only is the actual page structure sometimes so complex that even advanced extraction software has problems reliably accessing the correct component, but website publishers might make changes to the template at any moment, rendering the extraction template invalid. An invalid template will result in a 'null' content extraction, even though content might be present. The only way to avoid such issues would be to regularly check the websites for which content extraction is needed and adapt extraction templates as needed. In our case the automated content extraction did not function as we envisioned, forcing us to do manual scraping of the contents after the fieldwork had finished, bringing in a methodological problem, since the content of news items could have been changed since the respondent visited that particular news item.

5.2 Methodological challenges

The basis of the Newstracker was a local proxy. Even though applying a proxy to monitor Web behaviour is a relatively easy and a low-cost solution, it does ask some technical knowledge of the respondent. They need to configure it, possibly lowering the willingness of using the system. But they also need to log on to it, running the risk of the user to simply forget to turn on the proxy settings, resulting in a possible loss of data. Furthermore, because of the differences between browsers and operating systems, there is no one way to help users configure and easily log on to a proxy server.

A second methodological challenge is setting up the white list of websites that need to be monitored. In our case, we created a list of 3,564 of the most popular Dutch websites (Kleppe and Bijleveld, 2017). However, since our fieldwork took place during 3 months, it is likely a new popular website was launched during this period that should have been included. Especially researchers who want to monitor Web behaviour over a longer period of time should be aware and capable of constantly updating their white list.

A third and possibly most important methodological challenge is to find respondents who are willing to have their online behaviour being monitored. Inspired by Menchen-Trevino and Karr's (2012) emphasis on the privacy of the respondents, we built in several limitations of our research design to guarantee the respondents' privacy as much as possible and feasible. However, we truly needed to convince our respondents to participate and mainly found users willing to participate within the personal networks of our student assistants by building mutual trust and stay in close contact with them. Not only is this a very labour-intensive set-up for the researchers but also in the current age of growing fear of monitored personal big data after the National Security Administration (NSA) scandal (Davis *et al.*, 2013), and growing awareness of pre-installed tracking software in websites and banners (Martijn, 2013), we expect it to become increasingly difficult to find respondents willing to participate in these kinds of monitoring studies.

5.3 Analytical challenges

In the data analyses phase, we expect the future analyses of the scraped content will be a challenge. Especially researchers who are monitoring the

Web behaviour of large amounts of respondents during longer period of time will gather large amounts of scraped content. Since manual analyses of the textual and/or visual content would then not be feasible anymore, we envision the deployment of automated content analyses techniques to detect the topics that are being discussed in the news items (Atteveldt, 2008; Bhulai *et al.*, 2012) or the use of machine learning techniques to classify the political background of articles (Dehghani *et al.*, 2014; Munson *et al.*, 2013; Oh *et al.*, 2009) in the next planned iteration of the Newstracker. This would enable us to calculate the topical online news consumption during the day. To automate the analyses of the textual and visual content, we aim to incorporate a third phase in the set-up of the Newstracker (see Fig. 1 for the current set-up) after the data collection and pre-processing phase. At this moment, the final output is a csv-file containing all tracked information and scraped content. However, we envision the output of this third phase would be an online dashboard comparable with similar like dashboards of television viewing rates.[10] It would not only show the most visited websites, their genres, and visits during the day but also the contents of the visited websites based on the automated content analyses. This might open the possibility to not only know that an article about politics has been read but it would also give us information which political issue was addressed, which politician or which political party was discussed or quoted, and even which political under stream can be read in the article.

Finally, in our use case on measuring the consumption of online news, the Newstracker should include monitoring online behaviour on mobile devices. Recent studies show that in particular young news consumers practically always have their smartphones with them, making it their most important device for consuming news (Bank of America, 2015; Reuters Institute for the Study of Journalism, 2015). Future research on the role of news consumption that aims to monitor actual news consumption should therefor take into account this practice of cross-device news consumption. However, using proxy servers on mobile devices is far from easy and straightforward. Newer Android systems (from version 4.1) offer a so-called global proxy setting that allows all Internet connections to be routed through the proxy server. Nevertheless, since most Android users do not use that or newer versions of the Android operating system, this solution does not include enough Android users. Internetwork Operating System (iOS) systems do not allow for any proxy settings unless the system is under policy restrictions usually only used in a corporate environment where access to the smartphone is limited for the user. One of the solutions to include most, if not all mobile systems, could be using a Virtual Private Network connection or VPN. Normally a VPN server allows for a secure, encrypted, and anonymized Internet connection, but in combination with a proxy server would allow mobile devices to access the Internet through a proxy server via the VPN server. Both the Android and iOS operating systems allow for VPN connections and thus could be connected to the proxy server to log the users' Internet activity. However, keeping those connections active is a serious limitation on mobile devices due to battery saving settings. This might be avoided by using special applications that restore the VPN connection as soon as the phone becomes active again. Up until now, only few academics have experimented with tracking news consumption on smartphones and tablets (Damme *et al.*, 2015; Hasu, 2012), and also professional research companies aim to gain more insight into multiplatform news consumption (Bilton, 2015; Hafkamp, 2014) to tailor the publication of advertisement to the users online behaviour. Although none have found the silver bullet yet, it is clear that the increasing cross-device consumption of news asks for new approaches of monitoring news consumption on all possible platforms.

Funding

This article is part of the research project 'The New News Consumer: User-based Innovation to Meet Paradigmatic Change in News Use and Media Habits' (www.news-use.com). This work was supported by The Netherlands Organisation for Scientific Research (NWO) and ten Dutch

journalism organizations. The Newstracker hardware and software was supplied and created by the Tech Labs of the Network Institute of the Vrije Universiteit Amsterdam. Additional funding was received by the Network Institute Academy Assistant programme.

Acknowledgements

This article could not have been written without the research activities of Leonie Durlinger, Stefan Heijdra, and Hildebrand Bijleveld and the help of the project team members Prof. Dr Irene Costera Meijer, Prof. Dr Marcel Broersma, Dr Chris Peeters, Dr Anna van Cauwenberge, Joelle Swart, MA, and Tim Groot Kormelink, MA.

References

Atteveldt, W. van (2008). *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston: BookSurge.

Bank of America (2015). *Bank of America Trends in Consumer Mobility Report*. http://newsroom.bankofamerica.com/files/doc_library/additional/2015_BAC_Trends_in_Consumer_Mobility_Report.pdf.

Barthel, M., Shearer, E., Gottfried, J., and Mitchell, A. (2015). *The Evolving Role of News on Twitter and Facebook*. Pew Research Center's Journalism Project, 14 July. http://www.journalism.org/2015/07/14/the-evolving-role-of-news-on-twitter-and-facebook/ (accessed 28 August 2015).

Batista, P. and Silva, M. (2002). Mining Web Access Logs of an On-line Newspape. In *Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*. http://xldb.di.fc.ul.pt/xldb/publications/rpec02.pdf.

Bhulai, S., Kampstra, P., Kooiman, L., Koole, G., and Kok, B. (2012). Trend visualization on Twitter: What's hot and what's not?. In *IARIA. Data Analytics 2012: The First International Conference on Data Analytics*, Barcelona, Data Analytics, pp. 43–8.

Bilton, R. (2015). Cross-device tracking, explained. *Digiday*, 21 August. http://digiday.com/publishers/deterministic-vs-probabilistic-cross-device-tracking-explained-normals/ (accessed 2 September 2015).

Boczkowski, P. J., Mitchelstein, E., and Walter, M. (2011). 'Convergence across divergence: understanding the gap in the online news choices of journalists and consumers in Western Europe and Latin America. *Communication Research*, **38**(3): 376–96. doi: 10.1177/0093650210384989.

Chitraa, V. and Thanamani, A. S. (2011). A novel technique for sessions identification in web usage mining preprocessing. *International Journal of Computer Applications*, **34**:23–7.

Coffey, S. (2001). Internet audience measurement: a practicioner's view. *Journal of Interactive Advertising*, **1**(2): 10–17.

Damme, K. V., Courtois, C., Verbrugge, K., and Marez, L. D. (2015). What's APPening to news? A mixed-method audience-centred study on mobile news consumption. *Mobile Media & Communication*, **3**(2): 196–213. doi: 10.1177/2050157914557691.

Davis, K., Popovich, N., Powell, K., MacAskill, E., Spencer, R., Van Gelder, L., and Sacha, B. (2013). NSA Files: Decoded. What the Revelations Mean for You. *The Guardian*, 1 November. http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1.

De Haan, J. and Adolfsen, A. (2008). *De virtuele cultuurbezoeker*. Den Haag: Sociaal en Cultureel Planbureau. http://www.scp.nl/dsresource?objectid=19697&.

Dehghani, M., Sagae, K., Sachdeva, S., and Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". *Journal of Information Technology & Politics*, **11**(1): 1–14. doi: 10.1080/19331681.2013.826613.

Ebersole, S. (2000). Uses and gratifications of the web among students. *Journal of Computer-Mediated Communication*, **6**(1). doi: 10.1111/j.1083-6101.2000.tb00111.x.

Findahl, O. (2009). *The Swedes and the Internet 2009*. Gävle: World Internet Institute.

Findahl, O., Lagerstedt, C., and Aurelius, A. (2013). Triangulation as a way to validate and deepen the knowledge about user behavior. A comparison between questionnaires, diaries and traffic measurement. In Patriarche, G., Bilandzic, H., Jensen, J. L., and Jurišši, J. (eds), *Audience Research Methodologies: Between Innovation and Consolidation*. New York, NY: Routledge, pp. 54–69.

Hafkamp, M. (2014). Van Thillo: 'Snel beter onderzoek combinatie print en digitaal. *Adformatie*. http://www.

adformatie.nl/nieuws/van-thillo-%E2%80%98snel-beter-onderzoek-combinatie-print-en-digitaal%E2%80%99 (accessed 1 December 2014).

Hasu, T. (2012). ContextLogger2'. http://contextlogger.org/.

Huurnink, B. (2010). *Search in Audiovisual Boradcast Archives*. University of Amsterdam. http://dare.uva.nl/document/2/83234.

Karlsson, M. and Clerwall, C. (2013). Negotiating professional news judgment and "clicks". *Nordicom Review*, **34**(2): 65–76. doi: 10.2478/nor-2013-0054.

Kleppe, M. and Bijleveld, H. (2017). Most popular websites in the Netherlands 2015. *DANS*. https://doi.org/10.17026/dans-x6h-6qqt

Kopf, D. (2015). What New York Times Content Is Popular on Facebook? *Priceonomics*. http://priceonomics.com/what-new-york-times-content-is-popular-on-facebook/ (accessed 24 August 2015).

Lee, A. M., Lewis, S. C., and Powers, M. (2012). Audience clicks and news placement: a study of time-lagged influence in online journalism. *Communication Research*, **41**: 505–30. doi: 10.1177/0093650212467031.

Liu, B. (2011). *Web Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-19460-3 (accessed 26 August 2015).

Lotz, A. and Ross, S. M. (2004). Toward ethical cyberspace audience research: strategies for using the internet for television audience studies. *Journal of Broadcasting & Electronic Media*, **48**(3): 501–12.

Martijn, M. (2013). Big Business is watching you. *De Correspondent*, 9 October. https://decorrespondent.nl/66.

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy & Internet*, **5**(3): 328–39. doi: 10.1002/1944-2866.POI336.

Menchen-Trevino, E. and 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. doi: 10.1080/19331681.2012.664966.

Mobasher, B. (2007). Web usage mining. In *Web Data Mining*. Springer Berlin Heidelberg (Data-Centric Systems and Applications), pp. 449–83. http://link.springer.com/chapter/10.1007/978-3-540-37882-2_12 (accessed 26 August 2015).

Munson, S. A., Lee, S. Y., and Resnick, P. (2013). Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. In *Seventh International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6119 (accessed 25 September 2015).

Napoli, P. M. (2010). *Audience Evolution: New Technologies and the Transformation of Media Audiences*. New York, NY: Columbia University Press. http://cup.columbia.edu/book/audience-evolution/9780231150347.

Navarrete Hernández, T. (2014). *A History of Digitization: Dutch Museums*. University of Amsterdam. http://dare.uva.nl/record/1/433221 (accessed 25 September 2015).

Nederlandse Nieuwsmonitor (2013). *Seksmoord op horrorvakantie: de invloed van bezoekersgedrag op krantenwebsites op de nieuwsselectie van dagbladen en hun websites*. Amsterdam: Nederlandse Nieuwsmonitor. http://www.nieuwsmonitor.net/d/244/Seksmoord_op_Horrorvakantie_pdf.

Oh, A. H., Lee, H. -J., and Kim, Y.-M. (2009). User Evaluation of a System for Classifying and Displaying Political Viewpoints of Weblogs. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA*, May 17-20, 2009. The AAAI Press. http://aaai.org/ocs/index.php/ICWSM/09/paper/view/238 (accessed 30 September 2015).

Prior, M. (2009). The immensely inflated news audience: assessing bias in self-reported news exposure. *Public Opinion Quarterly*, **73**(1): 130–43. doi: 10.1093/poq/nfp002.

Reuters Institute for the Study of Journalism (2015). *Digital News Report 2015*. Oxford. http://www.digitalnewsreport.org/.

Robinson, J. P. (1985). The validity and reliability of diaries versus alternative time use measures. In Time, Goods, and Well-Being. Ann Arbor: University of Michigan.

Schrøder, K. C. and Kobbernagel, C. (2010). Towards a typology of cross-media news consumption: a qualitative-quantitative synthesis. *Northern Lights: Film and Media Studies Yearbook*, **8**(1): 115–37. doi: 10.1386/nl.8.115_1.

Somaiya, R. (2014). How Facebook Is Changing the Way Its Users Consume Journalism. *The New York Times*, 26 October. http://www.nytimes.com/2014/10/27/business/media/how-facebook-is-changing-the-way-its-users-consume-journalism.html (accessed 24 August 2015).

Spiliopoulou, M. (2000). Web usage mining for web site evaluation. *Communications of the ACM*, **43**(8): 127–34. doi: 10.1145/345124.345167.

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N. (2000). Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations Newsletter*, **1**(2): 12–23. doi: 10.1145/846183.846188.

Taneja, H. and Mamoria, U. (2012). Measuring media use across platforms: Evolving audience information systems. *International Journal on Media Management*, **14**(2): 121–40. doi: 10.1080/14241277.2011.648468.

Taneja, H., Webster, J. G., Malthouse, E. C., and Ksiazek, T. B. (2012). Media consumption across platforms: Identifying user-defined repertoires. *New Media & Society*, **14**(6): 951–68. doi: 10.1177/1461444811436146.

Tewksbury, D. (2003). What do Americans really want to know? Tracking the behavior of news readers on the internet *Journal of Communication*, **53**(4): 694–710. doi: 10.1111/j.1460-2466.2003.tb02918.x.

Usher, N. (2013). Al Jazeera english online. Understanding Web metrics and news production when a quantified audience is not a commodified audience. *Digital Journalism*, **1**(3): 335–51. doi: 10.1080/21670811.2013.801690.

Van Cauwenberge, A., d'Haenens, L. S. J., and Beentjes, J. W. J. (2010). Emerging consumption patterns among young people of traditional and internet news platforms in the Low Countries. *Observatorio*, **4**(3): 335–52.

Notes

1 http://en-us.nielsen.com/sitelets/cls/digital/online-netview.html

2 http://www.tns-nipo.com/ons-aanbod/marktonderzoek/digital/digital-analytics/, http://wakoopa.com/, and http://www.vinex.nl/ddmm/

3 http://www.alexa.com/topsites and http://www.similarweb.com/global

4 http://www.proceranetworks.com/products/platforms

5 http://www.roxyproxy.org/

6 http://www.wingate.com

7 We made all our code available for reuse as open-source code in the Network Institute repository at Github https://github.com/NITechLabs/NewsTracker.

8 http://www.vinex.nl/resultaten/archief/, http://www.alexa.com/topsites/countries/NL, http://www.similarweb.com/country/netherlands, and http://www.dmoz.org/World/Nederlands/

9 http://www.newprosoft.com/web-content-extractor.htm

10 https://kijkonderzoek.nl/