Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.

WYKORZYSTANIE INTERNETU W BADANIACH MARKETINGOWYCH

Praca zbiorowa pod redakcją

Ewy Zeman-Miszewskiej

Katowice 2005

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

ISBN 83-7246-824-9

WYDAWNICTWO AKADEMII EKONOMICZNEJ
IM. KAROLA ADAMIECKIEGO W KATOWICACH
ul. 1 Maja 50, 40-287 Katowice, tel. (032) 257-76-35, fax (032) 257-76-43
www.ae.katowice.pl e-mail:wydawucz@sulu.ae.katowice.pl

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

Krzysztof Kapera
Mariusz Kuziak

METODY POMIARÓW I ŹRÓDŁA BŁĘDÓW W BADANIACH OGLĄDALNOŚCI WITRYN INTERNETOWYCH

Media telematyczne, do których według T. Gobana-Klasa zalicza się także Internet, cechują między innymi interakcyjny, nieliniowy sposób odbioru, głębia informacyjna, elastyczność w formie treści i wykorzystaniu oraz relatywnie duża przepustowość kanału[1]. Te charakterystyki, wraz z relatywnie dużą grupą odbiorców, której wielkość w pierwszym kwartale 2005 roku wyniosła 28,4% Polaków[2], sprawiają, że Internet coraz rzadziej pomijany jest w działaniach komunikacyjnych firm, instytucji i organizacji. Dążenie do efektywnego wykorzystywania tego kanału skutkuje tym, że coraz większe znaczenie ma rzetelny pomiar oddziaływania poprzez podstawowe narzędzie działań marketingowych i komunikacyjnych, jakim jest witryna www.

Rozwój technologii internetowych na przestrzeni ostatnich kilkunastu lat, zwłaszcza w obszarze World Wide Web, doprowadził do wykształcenia się zbioru metod i narzędzi badania oglądalności witryn internetowych. Badania te na bardzo ogólnym poziomie podzielić można na takie, w których informacje pozy-

[1] T. Goban-Klas: Media i komunikowanie masowe. Teorie i analizy prasy, radia, telewizji i Internetu. Wydawnictwo Naukowe PWN, Warszawa-Kraków 2001, s. 25.
[2] Net Track, Millward Brown SMG/KRC, styczeń-marzec 2005 rok, reprezentatywna próba Polaków pomiędzy 15 a 75 rokiem życia (n=2565).

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

skiwane są od strony użytkownika (*user-centric*) i te, w których dane pochodzą bezpośrednio od mierzonego systemu, czyli witryny www (*site-centric*)[3]. Metody te w pewnym sensie się uzupełniają i w związku z tym do pewnego momentu uprawnione było twierdzenie, iż wykorzystywanie tylko badań *site-centric* nie zapewnia żadnej wiedzy o użytkownikach odwiedzających witrynę, a ograniczenie się wyłącznie do projektów *user-centric* pozostawia bez odpowiedzi pytanie o liczbę odwiedzających strony www[4].

Czterema podstawowymi metodami badania oglądalności witryn internetowych są kwestionariuszowe badania użytkowników Internetu, analiza logów (plików rejestrów dostępu), audyt oglądalności za pomocą systemów trackingowych oraz panel użytkowników sieci Internet. Każda z tych metod ma swoją specyfikę, możliwości pomiaru (np. zestaw dostępnych wskaźników charakteryzujących zachowania użytkowników) i ograniczenia (reprezentatywność, różna szczegółowość pomiaru).

Badania kwestionariuszowe użytkowników sieci Internet

Kwestionariuszowe badania dotyczące użytkowników sieci Internet można podzielić na dwa rodzaje: badania prowadzone off-line (poza siecią) oraz badania prowadzone on-line (poprzez Internet). W pierwszym przypadku mamy do czynienia z projektami wykorzystującymi dobrze znane i ugruntowane metodologicznie metody tradycyjne, zapewniające reprezentatywny dobór próby w obrębie populacji generalnej i wyniki obarczone relatywnie łatwym do określenia błędem pomiaru. Badania takie mogą być realizowane jako projekty *ad hoc*, pomiary okresowe lub ciągłe. Za każdym razem jednak badanie natrafia na problem zdefiniowania podmiotu badania, czyli użytkownika Internetu. Możliwymi zawężeniami definicyjnymi (choć nie jedynymi) mogą być np. rozróżnienie na korzystanie aktywne (samodzielne) i pasywne (bierne obserwowanie korzystania przez osobę trzecią), częstotliwość korzystania z Internetu (jednorazowy kontakt, sporadyczny dostęp, regularne wykorzystanie), intensywność korzystania (średni czas poświęcany na korzystanie z medium podczas jednego

[3] W literaturze przedmiotu sporadycznie spotyka się określanie mianem *user-centric* badań zorientowanych na użytkownika i mianem *site-centric* badań zorientowanych na witrynę. Por. M. Sobocińska: Badania marketingowe przez Internet – zalety i ograniczenia oraz zastosowania. W: Marketing – handel – konsument w globalnym społeczeństwie informacyjnym. T. I. Red. B. Gregor. Łódź 2004, s. 467.

[4] A. Tarkowski: Badacze na tropie internautów. „Brief" 2001, nr 2(18), s. 66-67.

Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.

dostępu lub łączny czas korzystania podczas jakiegoś okresu), wykorzystywane narzędzia (aplikacje) internetowe (poczta elektroniczna, www, *instant messaging* itd.), miejsce dostępu do Internetu (lokalizacja, z której następuje połączenie z Internetem), wiek itp. W przypadku kwestionariuszowych badań użytkowników Internetu, problemem są też koszty pozyskiwania danych, przekładające się na wielkość uzyskanej próby (w tym subgrupy użytkowników Internetu) i tym samym na szczegółowość (rozdzielczość) wyników.

Jednym z podstawowych problemów pojawiających się w kwestionariuszowych badaniach oglądalności witryn internetowych (i generalnie w badaniach kwestionariuszowych) jest bazowanie na pamięci respondenta. Przy olbrzymiej ilości faktów, informacji i zdarzeń absorbujących umysły nas wszystkich, mało kto jest w stanie zanotować świadomie wystąpienie określonego zdarzenia, które następnie – w badaniu, którego się nie spodziewa i do którego nie może się mentalnie przygotować – miałby odtworzyć. W większym stopniu spodziewać można się uwzględnienia w odpowiedziach respondenta wydarzeń znaczących dla niego, stanowiących zaskakujące odstępstwo od „normy" oraz tych, które są powtarzane z określoną regularnością. W pełni znajdują w tym przypadku zastosowanie reguły psychologii uczenia się i zapamiętywania. O ile więc w badaniach kwestionariuszowych stosunkowo dobrze można zbadać charakterystyki respondentów lub określony stan (jestem mężczyzną, pracuję, mam dostęp do Internetu, znam portal Onet.pl, posiadam prywatne konto poczty elektronicznej itp.), o tyle zachowania (odwiedzałem witrynę on-line „Rzeczpospolitej", wysłałem 25-30 wiadomości poczty elektronicznej w ciągu tygodnia, 3 razy byłem na czacie itp.) poddają się takiemu pomiarowi w dużo mniejszym stopniu. W badaniach kwestionariuszowych mierzony jest zatem w mniejszym stopniu fakt rzeczywistego wystąpienia określonego zachowania, w większym zaś – deklaracja jego wystąpienia.

Kwestionariuszowe badania oglądalności lepiej w związku z tym sprawdzają się w odniesieniu do krótkich lub bardzo długich okresów, gdyż z większym prawdopodobieństwem jesteśmy w stanie trafnie określić, czy korzystaliśmy z Internetu wczoraj lub czy kiedykolwiek odwiedzaliśmy np. witrynę internetową tygodnika „Wprost", niż odpowiedzieć, czy na ową witrynę zajrzeliśmy w miesiącu poprzedzającym miesiąc bieżący. Pomiar oglądalności dokonywany tradycyjnymi metodami kwestionariuszowymi realizowanymi off-line jest przy tym w stanie w miarę precyzyjnie odzwierciedlać popularność największych na rynku i najbardziej rozpoznawanych witryn. Z niewielkim raczej prawdopodobieństwem respondent będzie pamiętał, że ze stron wyszukiwarki trafił na którąś

Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.

z małych witryn tematycznych lub też, że w wyniku kliknięcia, został przeniesiony na prywatną stronę www jednego z użytkowników bezpłatnego hostingu.

Badania deklaratywne oglądalności witryn internetowych obciążone są nie tylko błędami wynikającymi z ułomności pamięci respondentów, ale także obarczone błędami świadomego „fałszowania" wyników badania przez samych respondentów. Ankietowany może bowiem nie chcieć np. przyznać, że odwiedza witryny zawierające treści o charakterze erotycznym. Ze względów emocjonalnych może też nie chcieć powiedzieć, że korzysta z określonej witryny, choć w rzeczywistości odwiedza ją, gdyż oferuje dostęp do nieosiągalnych w innych miejscach zasobów lub usług, np. wierni słuchacze Radia Zet – przez niechęć do RMF FM – mogą pomijać w swoich wskazaniach, należący do RMF FM, portal INTERIA.PL.

Z kolei badania kwestionariuszowe prowadzone on-line, oprócz powyższych problemów, napotykają także na bariery związane z doborem próby. Ze względów technicznych, rekrutacja do badania on-line z założenia wykorzystuje tylko jedną warstwę Internetu – najczęściej www. Takie badanie pomija więc od razu pewną grupę użytkowników, którzy korzystają z innych niż www narzędzi i aplikacji internetowych, np. odbierających tylko pocztę za pomocą odpowiedniego oprogramowania albo porozumiewających się komunikatorami internetowymi.

Odrębną kwestią jest zapewnienie on-line reprezentatywności próby. Choć badania on-line realizowane na dużą skalę dają wyniki bardzo zbliżone do uzyskiwanych na grupie reprezentatywnej, próby nieograniczone (tworzone na zasadzie samodoboru respondentów) nie mogą być uznane za reprezentatywne dla całej populacji, podobnie jak próby celowe. Konieczne jest zatem tworzenie próby metodą losowania systematycznego opartego na identyfikatorach *cookie* systemu trackingowego (gdzie jedynym wymogiem jest odpowiednio duży zasięg witryn objętych tym samym systemem trackingowym tak, by w okresie badania witryny, na których jest ono prowadzone były odwiedzane przez wystarczająco dużą grupę użytkowników).

Niedoskonałość badań kwestionariuszowych prowadzonych on-line ujawnia się także w tym, że umożliwiają one określenie tylko odsetka populacji użytkowników Internetu odwiedzających daną witrynę, nie zapewniając jednak informacji o liczbie internautów ogółem, nie dają tym samym podstawy do określenia liczby użytkowników serwisu www.

Dzięki kwestionariuszowym badaniom użytkowników Internetu można pozyskać ogólną wiedzę o korzystaniu z sieci, głównych odwiedzanych witrynach, częstotliwości korzystania, czasie przeznaczanym na przeglądanie stron

Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.

internetowych itp. Niedostępne są jednak nawet podstawowe wskaźniki obrazujące popularność poszczególnych witryn, jak np. liczba obejrzanych stron, częstotliwość powracania na daną witrynę (liczba sesji/wizyt) czy czas spędzony na przeglądaniu zasobów serwisu.

Analiza logów witryny www

Metoda analizy plików rejestru dostępu (tzw. logów) bazuje na wbudowanej w serwery www możliwości zapisywania wszystkich kierowanych do nich żądań transmisji zlokalizowanych na nich zasobów. Zaletą analizy logów jest brak uciążliwości tej metody pomiaru dla użytkowników serwisu i dla samej witryny. Niewątpliwą zaletą jest także uwzględnianie wszystkich żądań transmisji, niezależnie od tego, od użytkownika o jakich charakterystykach (np. w jakim wieku) i pracującego na jakim systemie operacyjnym (i z jaką przeglądarką stron www) one pochodziły.

Pomiar oglądalności witryn internetowych „na poziomie technicznym" w pierwszym rzędzie napotyka jednak właśnie na problemy definicyjne. Z punktu widzenia przeciętnego użytkownika Internetu, tzw. odsłona (*page-view*), czyli jedna z głównych jednostek pomiarowych ruchu na witrynie www, dokonuje się wówczas, gdy w przeglądarce stron internetowych na komputerze użytkownika pojawia się określona treść. Z punktu widzenia użytkownika nie ma przy tym znaczenia, skąd owa treść została zaczerpnięta i w jaki sposób (za pomocą jakiego protokołu internetowego) była dostarczona do przeglądarki www. Tymczasem w zdecydowanej większości, technicznych ze względu na swoją genezę, definicji strona określana jest jako „plik zapisany za pomocą Hypertext Markup Language (HTML)"[5]. Liczba stron pobranych (*number of pages*) wyliczona na podstawie statystyk serwera teoretycznie powinna być zatem równa liczbie stron poprawnie załadowanych do przeglądarek wszystkich użytkowników, a zatem im wyświetlonych. Jednak rozwój funkcjonalności programów służących użytkownikom do oglądania witryn internetowych zaowocował tym, że obsługują one poprawnie nie tylko protokół HTTP, służący transmisji plików HTML i obiektów z nimi powiązanych, ale również umożliwiają wyświetlenie w sposób analogiczny zawartości katalogu ogólnodostępnego serwera FTP czy treści do-

[5] Whatis.com Encyclopedia of Technology Terms, http://searchwebservices.techtarget.com/sDefinition/0,,sid26_gci212738,00.html, czas dostępu: 2005-04-24, hasło: page (tłumaczenie własne).

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

stępnych na – stosunkowo rzadko już występujących – serwerach Gopher. Jednocześnie od 1991 roku, czyli od momentu kiedy CERN upublicznił stworzony przez Tima Berners-Lee standard World Wide Web, składnia języka HTML została mocno rozszerzona, umożliwiając między innymi tworzenie stron www z wykorzystaniem technologii ramek (*frames*), w które zaczytywane są dodatkowe dokumenty HTML. Powoduje to, że liczba żądań transmisji dokumentów typu HTML odnotowanych w pliku logu nie jest równa liczbie stron www wyświetlonych w przeglądarce użytkownika, gdyż dokument „technicznie" składający się nawet z kilkunastu plików HTML internauta widzi jako jedną stronę. Dodatkowe zniekształcenia wyników pomiaru pojawiają się jako efekt generowania stron HTML już po stronie przeglądarki www w komputerze użytkownika, np. w wyniku działania skryptów, a nie transmisji dokumentu HTML, a także wskutek posługiwania się „pustymi" z punktu widzenia użytkownika (bo niewidzialnymi dla niego) przeładowaniami dokumentów HTML koniecznymi do poprawnego funkcjonowania witryny www.

Pomiar oglądalności witryny dokonany metodą analizy logów pozwala na określenie, które części serwisu cieszą się największym, a które najmniejszym zainteresowaniem. Możliwe jest generowanie szeregów czasowych i przedstawianie danych w przekrojach kolejnych dni miesiąca, tygodnia czy godzin doby. Pomiar oglądalności tą metodą jest w stanie dostarczać także danych na temat tego, z jakich systemów operacyjnych i jakich przeglądarek stron www korzystają osoby odwiedzające witrynę, z jakich krajów pochodziły żądania transmisji i w wyniku jakich zapytań do wyszukiwarek stron internetowych użytkownicy przenosili się na strony serwisu.

Popularne na rynku oprogramowanie do analizy zapisów plików rejestrów dostępu jest w stanie wskazywać strony, przez które użytkownicy najczęściej wchodzili na witrynę i na których najczęściej swoją wizytę kończyli. Przy nieco bardziej zaawansowanej analizie możliwe jest również wyznaczanie także pełnych „ścieżek" przemieszczania się użytkowników po witrynie.

W przeciwieństwie do badań kwestionariuszowych, metoda analizy logów nie jest w stanie dostarczać żadnych informacji na temat charakterystyk użytkowników witryny. Jeżeli jednak witryna jest w stanie zapisywać żądania transmisji w jednym z dwóch powszechnie akceptowalnych formatów (*Common Logfile Format, Combined Logfile Format*), możliwe jest uzyskanie dzięki niej informacji na temat ruchu na witrynie, niezależnie od jej wielkości i popularności. Statystyki oparte na analizie logów dotyczą jednak głównie aspektów technicznych funkcjonowania serwera, a zakres dostępnych zmiennych jest relatywnie ograniczony. Stosunkowo niewielką użyteczność marketingową mają dane

Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.

na temat liczby żądań transmisji (*hits*), wielkości transferu wyrażonej w KB czy liczby plików ogółem dostarczonych w wyniku żądań transmisji do przeglądarek użytkowników. Poprawnie skonfigurowany (dla danej witryny) program analizujący logi jest jednak w stanie dostarczyć precyzyjnej informacji na temat liczby odsłon stron serwisu (*number of pages*). Nieco mniej dokładna jest natomiast liczba wizyt (*visits*), gdyż wyliczana jest ona na podstawie numerów IP użytkowników, podczas gdy – ze względu na powszechną praktykę dynamicznego przyznawania numeru IP użytkownikom w sieciach lokalnych, maskowanie IP wewnętrznych i istnienie systemów typu proxy/w3cache – jednym numerem IP identyfikowanych jest często wielu użytkowników. Tymczasem to właśnie liczba unikalnych komputerów (*sites*) jest w analizie logów jedynym przybliżeniem liczby użytkowników odwiedzających określoną witrynę.

Odrębnym problemem w wykorzystaniu analizy logów jest porównywalność wyników uzyskanych przez różne witryny. W zależności bowiem od konfiguracji systemu, w plikach rejestrów dostępu zapisywane będą, lub nie, określone informacje, które następnie będą lub nie będą odzwierciedlone w statystykach zbiorczych. Sprawia to, że tak naprawdę w pełni porównywalne są jedynie statystyki z analogicznie skonfigurowanych serwerów www. Co więcej, wyniki analizy logów, ze względu na fakt, iż bazują one w całości na danych wewnętrznych, mogą być z powodzeniem i w stosunkowo prosty sposób fałszowane, co *de facto* uniemożliwia ich porównywanie na rynku.

Pomiar oglądalności systemem trackingowym

Metodą podobną do analizy logów jest pomiar oglądalności witryn systemem trackingowym, gdyż dane również i w tym przypadku są wynikiem analizy pewnego rodzaju logów lub zapisów w bazie danych. Zupełnie różna jest natomiast metoda gromadzenia danych źródłowych do owej analizy. Log w systemie trackingowym jest bowiem najczęściej efektem odwołania skryptu osadzonego na stronie www serwisu. Takie rozwiązanie powoduje jednak zarówno zwiększone obciążenie serwera (między innymi dodatkowym żądaniem transmisji, które musi być przez serwer obsłużone i zwiększeniem ilości danych transmitowanych przez serwer), jak i pewne zwiększenie obciążenia terminala użytkownika. Wymaga również dodatkowego wysiłku po stronie webmastera i administratora serwisu www, którzy muszą czuwać, by odpowiednie skrypty zamieszczone były w kodzie HTML wszystkich jego stron.

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

Potencjalnym zagrożeniem i źródłem błędów w pomiarach trackingowych jest pominięcie pewnych stron www i niezamieszczenie na nich skryptów, czego skutkiem jest wykluczenie tych stron witryny z pomiaru trackingowego. Analogiczny efekt wystąpić może w przypadku niepoprawnego wklejenia skryptu do źródła HTML strony. Z kolei – jeżeli system trackingowy bazuje na identyfikatorach przyznanych dla określonej strony lub grupy stron – umieszczenie na stronach skryptów z błędnym identyfikatorem powoduje zawyżenie wskazań dla pewnej części witryny i zaniżenie (albo całkowite wyzerowanie) wyniku dla stron, które w rzeczywistości tym identyfikatorem powinny być mierzone. Zawyżać statystyki odwiedzalności witryny, szczególnie w zakresie liczby odsłon stron, może zwielokrotnienie skryptu pomiarowego na stronie. Z kolei wyniki strony zaniżać może blokowanie przez użytkowników wykonywania na stronach serwisu skryptów pomiarowych.

Pewnego rodzaju zniekształcenia wyników powoduje także sama konstrukcja pomiaru trackingowego oglądalności, który w swoim zamyśle powstał jako pomiar audytowy, zewnętrzny. Oznacza to, że dane pierwotne, będące wynikiem odwołań skryptów ze stron witryny, gromadzone są na systemie zewnętrznym w stosunku do objętej pomiarem witryny. Zatem wszelkie zakłócenia w funkcjonowaniu sieci prowadzić mogą do nieuwzględnienia części odwołań w zapisach logów systemu trackingowego, a tym samym skutkować „gubieniem" pewnej części ruchu faktycznie występującego na witrynie. Z kolei w nieuprawniony sposób zwiększać wyniki witryny mogą przypadki wykorzystania skryptów monitoringowych systemu trackingowego przez osoby trzecie, np. w wyniku zaczerpnięcia kodu HTML wraz z osadzonym skryptem i posłużenia się nim jako wzorcem do stworzenia kolejnej witryny, zupełnie niezależnej od strony poddanej pomiarowi.

Czynnikiem mogącym wpływać na pomiar jest także sposób umieszczenia skryptu w kodzie HTML strony www. W przypadku, gdy skrypt wklejony jest na samym początku strony, zaraz za znacznikami <HTML> i <BODY>, zaczyna się on wykonywać od razu, jeszcze przed wyświetleniem właściwej zawartości na ekranie komputera użytkownika. Efektem tego, zwłaszcza w przypadku wolnego transferu, może być wpis w logu systemu trackingowego, oznaczający załadowanie (wyświetlenie) strony, podczas gdy zniecierpliwiony oczekiwaniem użytkownik przerwał transmisję i w rzeczywistości strony nie obejrzał. Z kolei umiejscowienie skryptu na samym końcu kodu HTML, tuż przed znacznikami </BODY> i </HTML>, skutkować może jego niewykonaniem, nawet jeżeli strona została w całości wyświetlona na komputerze użytkownika, lecz zdążył on przed uruchomieniem skryptu kliknąć w jeden z linków lub wpisać w pole adresu przeglądarki inny adres i rozpocząć przechodzenie na stronę kolejną.

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

Zdolność systemów trackingowych do dostarczania dużo większej ilości danych jest między innymi wynikiem wykorzystania w nich tzw. *cookies*, czyli „[...] niewielkich informacji tekstowych, wysyłanych przez serwer www i zapisywanych na twardym dysku użytkownika"[6]. Technologia *cookies*, zwanych także znacznikami kontekstu klienta, „[...] pozwala zapamiętywać informacje o użytkownikach korzystających z usług internetowych"[7], tym samym pozwalając na przybliżenie liczby użytkowników korzystających z danej witryny.

Liczba niepowtarzalnych *cookie*, odnotowanych w systemie trackingowym na danej witrynie, określana jest mianem liczby unikalnych użytkowników. *De facto* więc pomiar popularności serwisu www systemem trackingowym mierzy nie liczbę rzeczywistych osób, ale liczbę niepowtarzalnych *cookie* nadanych przez system (jest to jednak miara jednakowa dla wszystkich serwisów mierzonych systemami trackingowymi). Gdyby nie występowało zjawisko kasowania *cookie* i nadawania użytkownikom kolejnych, liczba niepowtarzalnych *cookie* byłaby wiernym odzwierciedleniem liczby niepowtarzalnych przeglądarek stron internetowych na komputerach użytkowników. Pewna część użytkowników kasuje jednak pliki *cookie* w wyniku reinstalacji systemu operacyjnego, zmiany komputera czy ze względu na ustawienia systemu itp., co prowadzi do zawyżenia liczby *cookie* ponad liczbę fizycznych osób odwiedzających witrynę. Jednak z jednego komputera może korzystać więcej niż jedna osoba, a wszyscy użytkownicy posługujący się tą samą przeglądarką będą identyfikowani tym samym *cookie*. Zaniżać liczbę *cookie* odnotowywanych przez system trackingowy będzie również blokowanie *cookie* przez przeglądarki użytkowników.

Wykorzystanie przez systemy trackingowe technologii *cookie* pozwala, pomimo niedoskonałości tej metody, na zmierzenie podstawowych czterech parametrów charakteryzujących oglądalność serwisu www: liczby unikalnych użytkowników, liczby odsłon stron (rozumianych jako wyświetlenie strony www użytkownikowi, a w praktyce będących wykonaniem skryptu pomiarowego), liczby wizyt (będących ciągiem odsłon stron wykonanych przez tego samego użytkownika, pomiędzy którymi odstęp nie jest większy niż ustalona jednostka czasu) oraz czasu spędzanego na witrynie, będącego sumą różnic pomiędzy czasem ostatniej i pierwszej odsłony wizyt wszystkich użytkowników. Pozwala również na obliczenie na ich podstawie wielu wskaźników pochodnych, takich jak np. liczba wizyt na użytkownika, liczba odsłon stron na użytkownika, liczba odsłon na wizytę, czas na jedną stronę itp.

---

[6] Wikipedia. http://pl.wikipedia.org/wiki/Cookies, czas dostępu: 2005-04-29, hasło: ciasteczka.
[7] Digipedia. http://definicje.digipedia.pl/def/314499713.html, czas dostępu: 2005-04-29, hasło: cookie.

Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.

Krzysztof Kapera, Mariusz Kuziak

Pomiar oglądalności witryn panelem użytkowników Internetu

W przypadku panelu użytkowników Internetu, pomiar oglądalności witryn internetowych realizowany jest dzięki monitoringowi zachowań użytkownika poprzez specjalny program zainstalowany na komputerze użytkownika. Pomiar ten może być wykonywany na dwóch poziomach:

- na poziomie protokołu internetowego, gdzie zbierane są wszystkie dane o ruchu wchodzącym do komputera użytkownika, niezależnie od tego, czy pochodzą one z przeglądarki, komunikatora internetowego, programu pocztowego, klienta FTP itp.,
- na poziomie przeglądarki stron www, gdzie zbierane są dane o stanie przeglądarki, np. o zakończeniu z sukcesem transmisji strony www.

W naturalny sposób obie te metody pomiaru dostarczać będą różniących się wyników.

Istotnym ograniczeniem zbierania danych o ruchu na witrynach internetowych za pomocą oprogramowania monitorującego jest rodzaj wykorzystywanego przez użytkownika oprogramowania, zarówno systemu operacyjnego, jak i – w przypadku programu monitorującego aktywność przeglądarki – wykorzystywanej przeglądarki stron internetowych. Opracowanie jednej tylko wersji oprogramowania monitorującego, np. dla systemów operacyjnych z rodziny Windows, eliminować będzie z założenia użytkowników korzystających z innych systemów, jak np. Linux czy MacOS. Z kolei wersje oprogramowania działające na różnych systemach operacyjnych, ze względu na różnice pomiędzy nimi, skutkować mogą różnymi odczytami nawet przy jednakowych zachowaniach użytkowników.

Potencjalne błędy ujawniają się również ze względu na chęć internauty do zainstalowania na swoim komputerze nieznanego mu oprogramowania. Obawy użytkownika mogą dotyczyć charakteru danych zbieranych przez to oprogramowanie (np. kwestii, czy nie zbiera ono także informacji o nielegalnym oprogramowaniu zainstalowanym na komputerze) albo tego, czy jego zainstalowanie nie spowoduje spowolnienia działania komputera i połączenia internetowego. Źródłem błędów może być też możliwość instalacji oprogramowania przez użytkownika. Osoba korzystająca z Internetu w pracy może nie mieć możliwości pobierania plików wykonywalnych z Internetu albo nie posiadać zgody przełożonych lub służb informatycznych przedsiębiorstwa na instalację oprogramowania na firmowym komputerze, a osoba korzystająca z Internetu grzecznościowo

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

– nie otrzymać zgody właściciela komputera na taką instalację. Negatywnie na możliwość pozyskiwania danych wpływa również korzystanie przez sporą liczbę internautów z terminali publicznych, ogólnodostępnych, jak np. uczelniane ośrodki informatyczne, biblioteki czy kawiarenki internetowe. Ze względu na fakt, że dane charakteryzujące korzystanie z Internetu łączone są z profilami społeczno-demograficznymi użytkowników, instalowanie oprogramowania monitorującego w miejscach, gdzie z Internetu korzysta wiele osób mija się z celem, gdyż prowadzi do zaburzenia pomiaru (np. znacząco zwiększając wskaźniki aktywności internetowej odnotowywane na danym komputerze).

Zaawansowana technologicznie metoda badań, jaką jest zbieranie danych o zachowaniach użytkowników on-line poprzez działanie oprogramowania monitorującego, narażona jest jednak nie tylko na powyższe błędy techniczne. Dane zebrane przez to oprogramowanie muszą być bowiem przesyłane do komputera centralnego, co może powodować bezpowrotną utratę części z nich w przypadku zakłóceń funkcjonowania sieci albo też znaczne opóźnienie w dostarczeniu pewnej ich części do przetwarzania (skutkującej koniecznością ich pominięcia), w przypadku, gdy dane nie zostaną przesłane przed rozłączeniem się użytkownika z siecią, a do momentu następnego połączenia z Internetem upłynie zbyt wiele czasu.

Odrębnym problemem i źródłem błędów są kwestie związane z rekrutacją i utrzymaniem panelu użytkowników Internetu. Ze względu na to – co jest niewątpliwą zaletą tej metody – że zbierane dane dotyczą całej aktywności użytkownika i bazują na jego faktycznych zachowaniach, a nie pamięci, w statystykach odnotowywane są także adresy tych stron, których w przypadku badań kwestionariuszowych internauta z bardzo dużym prawdopodobieństwem, by nie pamiętał. Daje to z jednej strony możliwość pozyskiwania danych nawet dla małych witryn i dla całych agregatów witryn (np. wielu setek serwisów www zrzeszonych w jedną sieć reklamową), z drugiej sprawia, że – by wyniki były wystarczająco szczegółowe (o dużej rozdzielczości) dla tych witryn – konieczne jest stworzenie i utrzymanie panelu bardzo licznego. To naturalnie pociąga za sobą spore koszty, zwłaszcza, jeżeli rekrutacja do panelu miałaby odbywać się off-line, metodami tradycyjnymi.

Częściowym rozwiązaniem problemu kosztów rekrutacji jest prowadzenie rekrutacji on-line, konieczne jest jednak wówczas posłużenie się reprezentatywną metodą doboru próby, taką jak np. losowanie systematyczne użytkowników po *cookie*. Ze względu na fakt systematycznego kasowania *cookie* przez pewną cześć użytkowników, trzeba jednak rekrutować użytkowników, których *cookie* jest starsze niż ustalony, średni dla populacji czas ponawiania wizyt w Interne-

**Krzysztof Kapera, Mariusz Kuziak: Metody pomiarów i źródła błędów w badaniach oglądalności witryn internetowych,
[w:] Wykorzystanie Internetu w badaniach marketingowych, praca zbiorowa pod red. Ewy Zeman-Miszewskiej, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego w Katowicach, Katowice 2005, s. 105-116.**

cie. Takie podejście eliminuje jednak z próby wszystkie te osoby, których regularność kasowania *cookie* jest krótsza niż ten okres.

Z kolei na utrzymanie liczebności panelu na określonym poziomie wpływać mogą wszelkiego rodzaju operacje wykonywane przez użytkowników Internetu na ich komputerach, skutkujące uszkodzeniem programu monitorującego aktywność, wyłączeniem (czasowym lub całkowitym) jego działania lub odinstalowaniem. Powoduje to konieczność ciągłego dorekrutowywania do panelu kolejnych uczestników. Proces ten powinien być jednak prowadzony ze sporym wyprzedzeniem w stosunku do przewidywanego spadku, gdyż – ze względu na efekt nowości mogący skrzywiać zachowania – dane pozyskane od nowego panelisty w początkowym okresie muszą zostać wyeliminowane z analizy.