STUDIA SOCJOLOGICZNE 2018, 4 (231)
ISSN 0039−3371, e-ISSN 2545–2770
DOI: 10.24425/122482

Anna Turner
Instytut Filozofii i Socjologii Polskiej Akademii Nauk
Marcin W. Zieliński
Instytut Filozofii i Socjologii Polskiej Akademii Nauk
Kazimierz M. Słomczyński
Ohio State University i Instytut Filozofii i Socjologii Polskiej Akademii Nauk

GOOGLE BIG DATA: CHARAKTERYSTYKA I ZASTOSOWANIE W NAUKACH SPOŁECZNYCH

W obliczu rewolucji technologii informatycznych badacze nauk społecznych mają przed sobą nie lada wyzwanie. Oto bowiem wraz ze zwiększającą się popularnością Internetu pojawiły się ogromne ilości danych zawierających opinie, poglądy i zainteresowania jego użytkowników. Chociaż analiza tych danych stawia przed badaczami poważne problemy metodologiczne, za ich użyciem przemawia fascynujący materiał powstający bez ingerencji badaczy. Dużą część tego materiału stanowią dane z najpopularniejszej na świecie wyszukiwarki Google. Co minutę jej użytkownicy ze wszystkich miejsc na świecie zadają ponad 3 miliony zapytań, które są następnie klasyfikowane i udostępniane za pomocą aktualizowanych na bieżąco narzędzi. W artykule tym omówione są próby adaptacji tych danych do potrzeb nauk społecznych, a także dotychczasowe badania na ten temat. Omówione są także praktyczne aspekty pracy z narzędziami Google'a: Google Trends oraz Google Keyword Planner. Artykuł jest przeznaczony przede wszystkim dla badaczy nauk społecznych zainteresowanych internetowymi źródłami Big Data oraz wykorzystaniem tych danych w pracy naukowej.

Słowa kluczowe: Google Trends, Google Keyword Planner, Big Data, dane internetowe, internetowe narzędzia analityczne

Google Big Data: Characteristics and Use in Social Sciences

Abstract

The IT revolution created serious challenges to researchers in the social sciences. The spreading popularity of the Internet resulted in a large quantity of data on opinions, beliefs, and interests of its users. Although researchers need to solve methodological problems

Anna Turner, IFiS PAN, e-mail: aturner@ifispan.waw.pl; Marcin W. Zieliński, IFiS PAN, e-mail: mzielinski@ifispan.waw.pl; Kazimierz M. Słomczyński, Ohio State University i IFiS PAN, e-mail: slomczynski.1@osu.edu

* Artykuł ten powstał w ramach grantu Narodowego Centrum Nauki (Preludium, 2017/25/N/HS6/01169, grant Anny Turner), przy poparciu programu Cross-national Studies: Interdisciplinary Research and Training Program (CONSIRT, www.consirt.osu.edu), którego autorzy są członkami. CONSIRT jest programem Polskiej Akademii Nauk (PAN) i Ohio State University (OSU), z siedzibą w Instytucie Filozofii i Socjologii PAN i Departamencie Socjologii OSU.

in order to analyze Internet data, these data constitute highly valuable material generated without researchers' involvement. A large part of this IT material is created by the Google search engine. Every minute the world-wide users of Google make over three million queries that are subsequently classified and made available through Google tools. In this article we describe attempts to adapt these tools to the needs of the social sciences and review recent research in this domain. We focus on practical issues of using two specific tools: the Google Trends and the Google Keyword Planner. This article is primarily addressed to researches in the social sciences who are interested in the IT sources of Big Data and intend to use this kind of data in their scientific endeavors.

Keywords: Google Trends, Google Keyword Planner, Big Data, IT data, IT analytical tools

Wstęp

W niedawnej edycji „Studiów Socjologicznych" ukazały się dwa artykuły poświęcone danym internetowym: „Potencjał Big Data w badaniach społecznych" (Żulicki 2017) i „Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: Perspektywa konstruktywistyczna" (Rodak 2017). Kontynuując dyskusję nad rolą Big Data w naukach społecznych, chcielibyśmy przybliżyć dane z wyszukiwarki Google oraz omówić możliwości ich analizy za pomocą wyspecjalizowanych i dostępnych narzędzi.

W literaturze zagranicznej możemy znaleźć co najmniej kilkaset artykułów poświęconych badaniom z wykorzystaniem danych z wyszukiwarki Google, opublikowanych w renomowanych czasopismach – takich jak „Nature", „Science", „PLOS One" czy „Public Opinion Quarterly" i takich, które tę renomę szybko zdobywają – „Big Data and Society". W tych czasopismach substantywne artykuły dotyczą analizy częstości poszukiwań konkretnych informacji oraz ustalania przyczyn, korelatów i konsekwencji takich zachowań dużych populacji odnoszących się do całych krajów lub ich części. Z kolei w wielu artykułach metodologicznych rozważane jest znaczenie danych z wyszukiwarki Google'a, ich opracowywanie i statystyczna obróbka.

Zainteresowanie danymi Google'a ma swe uzasadnienie w popularności tej wyszukiwarki. Co sekundę użytkownicy Google'a zadają około 55 tysięcy zapytań, co w skali dnia wzrasta do 4,7 miliarda. Google jest więc nie tylko ogromną bazą danych ale także aktualizowanym na bieżąco źródłem wiedzy na temat tego, co w danym okresie czasu jest dla użytkowników najbardziej interesujące i jakich informacji potrzebują. Istnieją narzędzia, które umożliwiają sprawdzenie *gdzie, czego i jak często szukano na całym świecie, co jest niewątpliwie ogromnym potencjałem do wykorzystania przez badaczy*[1].

[1] Narzędzia Google'a, które omawiamy w tym artykule mają przewagę nad narzędziami o podobnym działaniu, a oferowanymi przez inne firmy. Chociaż nie jest celem porównanie narzędzi oferowanych przez inne firmy, w zakończeniu tego artykułu odwołujemy się do analiz,

Mimo szybkiego przyrostu liczby artykułów o danych Google'a w czasopiśmiennictwie światowym, nie znaleźliśmy jak dotąd analogicznych artykułów – czy to substantywnych, czy metodologicznych – w czasopismach wydawanych w Polsce. Brak zobiektywizowanego zainteresowania takimi danymi wśród polskich naukowców jest przypuszczalnie spowodowany złożonym splotem różnych czynników. Z naszych rozmów z przedstawicielami nauk społecznych z różnych ośrodków naukowych w Polsce wynika, iż dwa spośród tych czynników są szczególnie istotne: niedostateczna wiedza, co dane Google'a w istocie oferują, oraz nieznajomość narzędzi do analizy tych danych[2]. Artykuł nasz jest odpowiedzią na te niedostatki dostarczając informacji o tym, czego można się dowiedzieć o wyszukiwaniach w Google w Polsce i na świecie, oraz z jakich narzędzi analitycznych Google'a można korzystać.

Wyszukiwania stron internetowych jako dane

Od dwóch dziesięcioleci Google sukcesywnie wzmacnia swoją pozycję jako firma, która współtworzy Internet, a poprzez kreowanie i dostarczanie nowych rozwiązań technologicznych ma realny wpływ na to, jak się poruszamy w wirtualnej rzeczywistości. Misja firmy jest niezmienna od momentu jej powstania. To uporządkowanie informacji z całego świata oraz udostępnianie ich do ogólnego użytku, a celem utrzymanie bezpłatnej i otwartej sieci (Google 2018a). John Batelle (2006) w swoim bestsellerze o roli Google'a w transformacji kulturowej wykazał, iż instytucja ta dostarcza bazy danych naszych intencji i stanowi repozytorium ludzkiej ciekawości, zainteresowań i pragnień. Jednak Google to także firma kontrowersyjna, wielokrotnie zmagająca się z zarzutami o nieetyczne zachowania. Fala krytyki spadła na Google po tym, jak w 2013 roku Edward Snowden ujawnił skalę nadużyć w sektorze informacyjnym. Największe firmy internetowe świata takie jak Google, Facebook, Microsoft i wiele innych przekazywały wszystkie dane swoich użytkowników (bez ich wiedzy i zgody) Amerykańskiej Agencji Bezpieczeństwa Narodowego (NSA). W roku 2017 zapadł niekorzystny dla Google wyrok, Komisja Europejska nałożyła na

które przeprowadziliśmy w tym zakresie. Od razu jednak trzeba podkreślić, iż czasowy i terytorialny zasięg narzędzi Google'a jest znacznie większy niż narzędzi oferowanych przez inne firmy.

[2] Rozmawialiśmy na ten temat bardziej lub mniej formalnie z kilkunastoma socjologami i politologami zorientowanymi na badania ilościowe w Warszawie, Poznaniu, Krakowie i Zielonej Górze. Rozmowy te nie pretendują do roli oficjalnych wywiadów, a nasi interlokutorzy nie stanowią reprezentacji swoich dyscyplin czy swoich ośrodków. Powołujemy się jednak na te rozmowy, gdyż wskazywane czynniki braku zainteresowania danymi z wyszukiwarki Google powtarzały się w prawie wszystkich rozmowach. Nie sądzimy, aby był to przypadek.

spółkę karę w wysokości 2,42 mld euro za złamanie przepisów o ochronie konkurencji. Poszło o rynek porównywarek cenowych, zarzucono i udowodniono Google stosowanie zakazanych praktyk polegających na promowaniu w wynikach wyszukiwania porównywarki należącej do Google przy jednoczesnym obniżaniu rankingów produktów należących do konkurencji (Komisja Europejska). W obliczu takich sytuacji wydaje się, że misja Google to jedynie slogan, który nie ma wiele wspólnego z praktykami firmy.

Nowe technologie na dobre zadomowiły się w życiu wielu ludzi wpływając na szereg aspektów codziennego życia. Już od lat więcej osób szuka informacji korzystając z Internetu niż z radia, telewizji i gazet (Lovink 2009; Gunter, Rowlands i Nicholas 2009; Vaidhyanathan 2011). Wpierw rozpatrzmy poszukiwanie informacji w wyszukiwarkach od strony użytkownika. Użytkownik dowolnej wyszukiwarki musi wiedzieć – przynajmniej ogólnie – jaka informacja go interesuje i w związku z tym formułuje pytanie w formie hasła, które może składać się z jednego lub wielu wyrazów. Po wprowadzeniu hasła do wyszukiwarki otrzymuje w jakiś sposób uporządkowany zbiór plików, zawierających hasło, które wybrał. Tak działają wszystkie wyszukiwarki, chociaż różnią się przeszukiwanymi zasobami.

Wyszukiwarka Google przetwarza – według różnych szacunków – od 70% do ponad 90% wszystkich wyszukiwań na całym świecie. Wyszukiwania mogą odbywać się – praktycznie – w dowolnym języku. Jedynie w trzech krajach bardziej popularne są inne wyszukiwarki, tylko częściowo ze względu na specyfikę językową: Naver w Korei Południowej, Yandex w Rosji i Baidu w Chinach. Poza tymi wyjątkami – związanymi nie tylko z językiem, ale i możliwościami kontroli użytkowników – zasięg Google jest zdecydowanie największy, nawet w krajach, w których powstają i konkurują inne wyszukiwarki uniwersalne[3]. Poza tym istnieją wyszukiwarki wyspecjalizowane w swych funkcjach[4].

W Google'u każde wyszukiwanie jest odnotowywane i może podlegać agregacji na poziomie lokalnym (np. miasta), regionalnym (np. województwa), krajowym (np. Polska), a także ponadpaństwowym (np. Europa). Zagregowane wyszukiwania należą do Big Data, ale są kategorią wyraźnie oddzielną od innych. Ze względu na mnogość definicji terminu Big Data, uważamy za istotne wyjaśnienie jak rozumiemy to pojęcie i dlaczego klasyfikujemy dane Google

[3] Według widzialni.pl, w Polsce w kwietniu 2018 roku procentowe frekwencje użycia różnych wyszukiwarek były następujące: Google 97,69, bing 1,54, Yahoo! 0,51, DuckDuckG0 0,13, YANDEX RU 0,03, Interia 0,03. (Widzialni.pl 2018). W tym czasie Google posiadał 91,0% udziału w rynku światowym. Metodologia obliczeń podana jest na stronie http://gs.statcounter.com/faq#global-stats-accurate. Dostęp 01.03.2018.

[4] Do tych wyszukiwarek należą, między innymi, A9 (odnajdowanie produktów w sklepach internetowych), Technokrati (indeksowanie blogów), BlogScope (wynajdywanie treści blogów), Picsearch (znajdowanie obrazów i zdjęć), oraz 123people (poszukiwanie ludzi).

jako Big Data. Ogólnie, pojęcie Big Data odnosi się do dużych, różnorodnych i zmiennych zbiorów danych, których przetwarzanie wymaga specjalnych narzędzi (Schönberger i Cukier 2013; Lazer, Kennedy, King i Vespignani 2014a; Nagler i Tucker 2015; Jenkins, Słomczyński i Dubrow 2016). W przypadku danych z wyszukiwarki Google'a wielkość zbiorów danych zwykle określa się jako iloczyn liczby użytkowników, haseł, które podlegają sprawdzeniu ilości czasu, dla którego dokonuje się ustaleń. Dla populacji 10 milionów wyszukiwanie 100 haseł przez 10 dni daje już 10^{10}. Różnorodność i zmienność danych wyraża się w wymiarze powiązań między hasłami i w wymiarze czasu. Obróbka danych wymaga wyspecjalizowanych narzędzi, które omówimy w następnej części tego artykułu.

Wyszukiwania Google to unikatowy rodzaj danych: są darmowe, łatwo dostępne, i nie byłyby możliwe do zebrania za pomocą tradycyjnych metod (Trevisan 2013, 2014). Bodaj największą ich zaletą jest ich niezakłócony charakter (Mellon 2013a; Zhu, Wang, Qin i Wu 2012; Ragas i Tran 2013; Kaisheng, Xin, Hao i Rongjun 2017). To użytkownik Internetu decyduje, czym jest zainteresowany i spontanicznie wyszukuje informacje na wybrane przez siebie tematy. Unika się tu sytuacji badawczej, gdy – jak pisał Stefan Nowak (2006: 86) – „w pewnych warunkach ujawnianie pewnych poglądów uchodzi za niewłaściwe lub nieprzyzwoite". Anonimowość zapewnia użytkownikowi poczucie prywatności i komfortu, a to czego szuka, nie jest obciążone ryzykiem oceny[5]. Dzięki temu uzyskujemy „obiektywny" – niezależny od badacza – materiał (Krzysztofek 2011), szczególnie istotny w obszarze tematów wrażliwych (Kaisheng, Xin, Hao i Rongjun 2017) – takich jak np. rasizm (Stephen-Davidowitz 2014), przemoc domowa, aborcja czy myśli samobójcze (Page, Chang i Gunnell 2011; Lester i Gunn III 2013).

W przypadku danych z internetowych wyszukiwarek mamy do czynienia z ich niewątpliwą zaletą w porównaniu z danymi zbieranymi podczas badań ankietowych (gdzie występuje efekt samego narzędzia badawczego i efekt ankieterski) lub z danymi z mediów społecznościowych (gdzie występuje efekt kreowania własnego wizerunku przed obserwatorami). Wszakże należy także pamiętać, że korzystając z danych z internetowych wyszukiwarek nie wiemy niczego istotnego o szukającym i nie jesteśmy w stanie w żaden sposób określić jego motywacji. Tak, na przykład, jeśli ktoś szuka informacji na temat przemocy domowej, przyczyn może być co najmniej kilka: faktyczne doświadczenie przemocy domowej i próba znalezienia pomocy, zainteresowanie tematem na skutek doniesień medialnych lub z powodów zawodowych, lub chęć pomocy osobie, o której wiemy, że takiej przemocy doświadcza. Są to jednak tylko nasze przypuszczenia.

[5] Niektóre tematy – jak pornografia dziecięca – są w wielu krajach zabronione przez prawo.

Dane z wyszukiwarki Google są wyjątkowe ze względu na powtarzające się w czasie i przestrzeni obserwacje na milionach użytkowników w 179 krajach. Umożliwia to badanie bardzo dużych grup zdefiniowanych przez położenie geograficzne, a także przeprowadzenie analiz porównawczych zarówno w różnych okresach czasu, jak i na poziomie wewnątrzkrajowym, krajowym i międzynarodowym. Biorąc pod uwagę to bogactwo materiału, dane Google'a były wielokrotnie wykorzystywane jako zmienne zależne i niezależne w różnych projektach badawczych z zakresu psychologii, ekonomii, kryminologii, zdrowia publicznego czy meteorologii (Askitas i Zimerman 2009; Carneiro i Mylonakis 2009; Lui, Metaxas i Mustafaraj 2011; Ripberger 2011; Sherman-Morris, Senkbeil i Carver 2011; Preis, Moat, Stanley i Bishop 2012; Nuti, Wayda, Ranasinghe, Wang, Dreyer, Chen i Murugiah 2014, Heiberger 2015; Gamma, Schleifer, Weinmann, Buadze i Liebrenz 2016, Wang, Zhang, Lu, Zhou, Chen i Niu 2018), a także z zakresu socjologii i nauk politycznych. W projektach tych dane z wyszukiwarki były wykorzystane między innymi do konstrukcji:

– wskaźnika ważności tematu debaty publicznej (*public agenda setting*) (Granka 2010; Scharkow i Vogelgesang 2011; Ragas, Tran i Martin 2013; Maurer i Holbach 2015; Hoskins, Trevisan, Oates i Mahlouty 2018);

– barometru istotności bieżących spraw (*issue salience barometer*) (Mellon 2013b; Ragas i Tran 2013; Maurer i Holbach 2015);

– wskaźnika nastawień opinii publicznej (*search queries as indicators of public opinion*) (Zhu, Wang, Qin i Wu 2012; Santos 2016).

W konstrukcji tych i innych wskaźników opartych na danych internetowych przyjmuje się, że im więcej zapytań jest wstawianych do wyszukiwarki, tym istotniejszy jest dany problem w badanej populacji. Wzrost istotności danego problemu może mieć różne źródła (Broder 2002), albo rodzaju „wydarzenie zewnętrzne, na które reagują członkowie populacji" albo „wydarzenie dotykające członków populacji, na które oni reagują". Kiedy ludzie nagle poszukują informacji na temat korupcji, to jest wysoce prawdopodobne, iż jakieś wydarzenie spowodowało, iż ten temat dla ludzi stał się ważny. Nie wyklucza to jednak sytuacji, w której wyszukiwania tego samego tematu są spowodowane przez osobiste doświadczenia użytkowników. To odpowiada drugiemu rodzajowi reakcji. Oczywiście poszukiwania w Internecie mogą też wynikać ze stałych zainteresowań członków populacji jakimiś konkretnymi sprawami. Chociaż to, co podlega wnioskowaniu z danych internetowych zależy od motywacji autorów zapytań, możemy tylko przyjmować specyficzne założenia dotyczące tych motywacji – niestety nie możemy ich weryfikować wykorzystując ten sam materiał. Wszakże weryfikacja tych założeń może pochodzić z innych badań niż wykorzystujących

dane z wyszukiwarki (Teo, Lim i Lai 1999). Należy podkreślić, że za wyjątkiem głośnych afer medialnych czy to na szczeblu lokalnym, krajowym czy międzynarodowym, które mogły spowodować zainteresowanie danym tematem, jest niesłychanie trudno przypisać dane wyszukiwanie do któregoś z wymienionych źródeł.

Google zbiera dane na temat swoich użytkowników, jednakże dane, jakie udostępnia w narzędziach są mocno okrojone. Używając opisanych w tym artykule narzędzi możemy się dowiedzieć, jak często i gdzie, ale nie wiemy, kto i z jakich powodów szukał czegoś w Google'u. Nawet ustalenie podstawowych cech demograficznych autora zapytań nie jest możliwe. Dane, które uzyskuje się poprzez narzędzia analityczne Google'a, są zmiennymi, które powstają w wyniku agregacji i tylko w ten sposób mogą być interpretowane (Lazarsfeld i Menzel 1961; Nowak 2006).

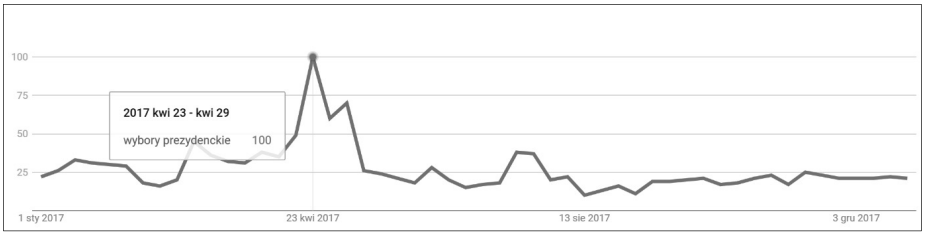
Narzędzia

Istnieją dwa narzędzia Google'a, przy użyciu których można sprawdzić ilość zapytań kierowanych do wyszukiwarki[6]: Google Trends (dostępne od 2006 roku, stosowane we wszystkich badaniach opisanych w literaturze nauk społecznych) oraz Google Planer Słów Kluczowych (dostępne od 2000 roku, jak dotąd szeroko stosowane jedynie w marketingu i badaniach marketingowych). Obydwa narzędzia łączy jedno – przedstawiają częstości tego, czego szukają użytkownicy wyszukiwarki. Jednakże między tymi narzędziami występują zasadnicze różnice, które tu streszczamy:

(1) *Sposób pomiaru* – w Google Trends liczba zapytań danego słowa nie jest liczbą całkowitą a znormalizowanym przez Google indeksem: „liczby przedstawiają, jak często hasło było wyszukiwane w odniesieniu do najwyższego punktu wykresu w danym czasie i regionie. Wartość 100 oznacza najwyższą popularność hasła. Wartość 50 oznacza, że popularność hasła była dwukrotnie mniejsza. Natomiast 0 świadczy o tym, że popularność hasła wynosiła mniej niż 1% najwyższej wartości" (Google 2018b). Na wykresie 1 podano częstości występowania hasła „wybory prezydenckie" w okresie od 1 stycznia do 3 grudnia 2017 roku.

[6] Istnieje także narzędzie Google Correlate, bardzo proste w użyciu, wpisujemy interesujące nas słowo kluczowe, określamy czas i przestrzeń i na podstawie tych danych mechanizm zwraca listę zapytań najlepiej skorelowanych z naszym. Wyniki są przedstawione za pomocą współczynnika korelacji r-Pearsona. Niestety od marca 2013 roku narzędzie nie jest już uaktualniane przez Google.

Wykres 1. Google Trends – panel wyników



W Planerze Słów Kluczowych natomiast mamy podaną średnią liczba wyszukiwań miesięcznych, a więc dane o wiele bardziej precyzyjne. Nasze analizy wykazały, że Google przyporządkowuje każde słowo do jednego z 82 segmentów wyszukiwań[7]. Dla przykładu, jeżeli dane słowo ma 100 wyszukiwań, może trafić albo do segmentu 90 albo 110, ale jeżeli ma 10 000 wyszukiwań, wtedy może trafić albo do segmentu niższego 9 900 albo do wyższego 12 100. Im wyższa liczba wyszukiwań, tym większa różnica między wartościami skrajnymi segmentu.

Wykres 2. Planer Słów Kluczowych – panel wyników

Search terms	Avg. monthly searches [?]
wybory prezydenckie	2,900

[7] Po przeanalizowaniu liczby wyszukiwań okazało się, że Google posiada 82 „koszyki liczby zapytań", nazwaliśmy je segmentami, które są proporcjonalnie zlogarytmizowane: 10, 20, 30, 40, 50, 70, 90, 110, 140, 170, 210, 260, 320, 390, 480, 590, 720, 880, 1000, 1300, 1600, 1900, 2400, 2 900, 3600, 4400, 6600, 8100, 9900, 12 100, 14 800, 18 100, 22 200, 27 100, 40 050, 49 500, 60 500, 74 000, 90 500, 110 000, 135 000, 165 000, 201 000, 246 000, 301 000, 368 000, 450 000, 550 000, 673 000, 823 000, 1 000 000, 1 220 000, 1 500 000, 1 830 000, 2 240 000, 2 740 000, 3 350 000, 4 090 000, 5 000 000, 6 120 000, 7 480 000, 9 140 000, 11 100 000, 13 600 000, 16 600 000, 20 400 000, 24 900 000, 30 400 000, 37 200 000, 45 500 000, 55 600 000, 68 000 000, 83 100 000, 124 000 000, 151 000 000, 185 000 000, 226 000 000, 414 000 000, 506 000 000, 923 000 000, 1 120 000 000, 3 760 000 000. Jeżeli zapytanie „Barack Obama" ma 74 000 miesięcznych wyszukiwań, to nie znaczy, że jest to dokładna liczba tych wyszukiwań. 74 000 jest pomiędzy segmentem 60 500 a 90 500 (co znaczy, że średnia miesięczna liczba wyszukiwań tego słowa jest bliska 74000, ale wciąż oscyluje pomiędzy 60 500 a 74 000 lub 74 000 a 90 500). Im większa liczba wyszukiwań, tym większe potencjalne różnice.

Podsumowując, żadne z narzędzi nie podaje dokładnej liczby wyszukiwań. Jednakże Planer Słów Kluczowych daje przybliżoną rzeczywistą wartość. Różnice pomiędzy narzędziami wynikają z ich przeznaczenia, Google Trends jest narzędziem bezpłatnym, publicznie dostępnym, a przez łatwość obsługi i atrakcyjność sposobu przedstawienia wyników bardzo często i szeroko stosowanym, także przez osoby, które nie mają doświadczenia w analizie danych. Nie jest zatem konieczne podawanie dokładnych wartości numerycznych, indeksy wyszukiwań w zupełności wystarczą, aby pokazać trend zainteresowania danym tematem. Planer Słów Kluczowych jest narzędziem o wiele bardziej zaawansowanym, dostępnym dla wąskiej grupy specjalistów, którzy zajmują się analizą danych, chociażby na potrzeby budowania kampanii reklamowych. W tym przypadku znormalizowany indeks nie byłby wystarczający, dlatego Google udostępnia uśrednione rzeczywiste wartości.

(2) *Dokładność* – Google Trends funkcjonuje najlepiej dla bardzo popularnych zapytań, jeśli chcemy sprawdzić słowo rzadziej wyszukiwane, szczególnie w języku innym niż angielski, możemy otrzymać informację o braku wystarczających danych do przedstawienia wyników. Planer Słów Kluczowych także i w tym przypadku jest bardziej precyzyjny. Otrzymujemy wyniki nawet dla słów, których miesięczna ilość wyszukiwań jest mniejsza niż 10.

(3) *Podobne zapytania* – obydwa narzędzia mają bardzo przydatną funkcjonalność do każdego zapytania, którego popularność chcemy sprawdzić. W efekcie otrzymujemy listę podobnych słów kluczowych, które także były wyszukiwane. Tak więc, np., gdy interesuje nas słowo „uchodźcy", Google Trends podpowiada maksymalnie 25 najpopularniejszych zapytań (w tym: uchodźcy w Polsce, imigranci, uchodźcy w Niemczech). Wszakże Planer Słów Kluczowych znów jest bardziej precyzyjny, oferuje prawie 800 dodatkowych zapytań do każdego słowa, które sprawdzamy.

(4) *Zakres dat* – w Google Trends prezentowane są dane od 2004 roku, w Planerze Słów Kluczowych z ostatnich 4 lat.

(5) *Pobieranie danych* – z obydwu narzędzi można pobrać dane ręcznie jako pliki w formacie CSV (Coma Separated Values) lub z wykorzystaniem API (Application Programming Interface), co znacznie przyspiesza prace kompletowania bazy słów kluczowych. Przeprowadzone przez nas testy wykazały, że z pomocą API możemy pobierać około 50 tysięcy słów kluczowych w ciągu dwóch dni.

(6) *Dostęp do narzędzi* – Google Trends jest narzędziem darmowym i publicznie dostępnym. Planer Słów Kluczowych jest narzędziem stworzonym z myślą o tworzeniu reklam w Google, co ma swoje konsekwencje. Po pierwsze, dostęp do platformy został jakiś czas temu przez Google mocno ograniczony i obecnie mogą z niej w pełni korzystać tylko użytkownicy, którzy posiadają

aktywne kampanie reklamowe Google AdWords[8]. Pozostali użytkownicy muszą przynajmniej posiadać konto Gmail, jednak wtedy zamiast ilości wyszukiwań, Google udostępnia zakresy liczbowe. Porównaliśmy, jak wyglądają wyniki dla zapytania „wybory prezydenckie" z aktywnym kontem AdWords i bez: w pierwszym przypadku to konkretna liczba 2,900; w drugim przedział: 1K-10K. Różnica jest ogromna, dlatego warto uzyskać dostęp do pełnej wersji narzędzia.

Trafność danych z wyszukiwarek

Trafność pomiaru (*validity*) oznacza zakres, w jakim miernik empiryczny właściwie odzwierciedla rzeczywiste znaczenie danego zjawiska, czyli jak dobrze mierzy pojęcie, które ma mierzyć. Trafność danych z wyszukiwarki Google była dotychczas szacowana w taki sam sposób, jak trafność innych wskaźników (Mellon 2013a,b). Warto więc rozważyć cztery zasadnicze typy trafności w zastosowaniu do tego rodzaju danych[9]:

(a) Trafność fasadowa *(face validity)* – to ustalenie listy słów kluczowych, których internauci mogli użyć, aby szukać informacji na dany temat. Na przykład, jeśli celem badania jest ustalenie stopnia zainteresowania tematyką wyborów prezydenckich w Polsce, to wyrażenia takie, jak „wybory prezydenckie", „wybory na prezydenta", „kandydaci na prezydenta", a także imiona i nazwiska wszystkich kandydatów spełniają kryterium *face validity*. Kompletując listę słów kierujemy się przede wszystkim zdrowym rozsądkiem, dobieramy słowa popularne, często używane w języku potocznym. Pomocna może być analiza słownictwa publikacji medialnych na interesujący nas temat, gdyż media z reguły inspirują użytkowników do wyszukiwania informacji. Dodatkowo możemy także sprawdzić, jakich określeń używa się badając to samo zagadnienie w badaniach sondażowych lub użyć materiału z wywiadów zogniskowanych (*focus groups*).

(b) Trafność treściowa *(content validity)* zostanie spełniona, kiedy wykluczone zostaną słowa, które są co prawda wyszukiwane przez użytkowników, ale nie odpowiadają wybranemu problemowi badawczemu. Jeszcze na etapie kompletowania listy należy się upewnić, czy wybrane przez nas słowa nie mają wielu znaczeń. Kontynuując przykład wyborów prezydenckich, gdyby oprócz

[8] Google przyczynił się także do powstania nowej dziedziny: „marketingu w wyszukiwarkach" (*search engine marketing*). Odpowiadając na potrzeby rynku, Szkoła Główna Handlowa w Warszawie od kilku już lat kształci przyszłych menedżerów na kierunku e-biznes o specjalności „Marketing Internetowy". Jak czytamy na stronie uczelni: „Praktyczny wymiar studiowania znajduje odzwierciedlenie m.in. w realizacji kampanii promocyjnej w systemie Google AdWords...".

[9] Kryteria te mogą mieć zastosowanie także do danych z innych wyszukiwarek, nie tylko z Google'a.

kandydata na prezydenta Andrzeja Dudy, istniał inny znany Andrzej Duda, to wyrażenie nie mogłoby zostać użyte w badaniu, ponieważ nie bylibyśmy w stanie odróżnić, kogo ma na myśli szukający. W takiej sytuacji należałoby użyć bardziej precyzyjnych określeń, np. „Andrzej Duda kandydat na prezydenta" czy „Andrzej Duda PiS". Podsumowując, wieloznaczność słowa, którego chcemy użyć wyklucza je z badania, ponieważ nie spełnia ono kryterium trafności treściowej (*content validity*).

Należy także pamiętać o dokładnym sprawdzeniu, z jakimi wyrażeniami wprowadzone słowo będzie w przeglądarce Google łączone. Na przykład, dla terminu „praca" Google automatycznie „doda" do naszej listy wszystkie wyrażenia, których także z „pracą" wyszukiwano, a zatem znajdą się „oferty pracy", „praca za granicą", „praca w Polsce", „praca chałupnicza", „praca dodatkowa", „praca od zaraz", „praca dla kobiet", „praca socjalna", „praca po godzinach", „praca domowa", „praca w ochronie", „praca w internecie", „praca – student", a nawet „składanie długopisów" i wiele innych. Wszystkie te słowa są związane z pracą, ale mierzą różne jej aspekty i mogą nie spełniać kryterium trafności treściowej (*content validity*). Wniosek jest taki, że należy unikać stosowania ogólnych definicji, które z założenia będą zawierały wiele terminów niepoprawnie adresujących pojęcie, które chcemy mierzyć.

(c) Trafność kryterialna (*criterion validity)* opisuje, jak dobrze nasza miara (dane Google'a) koresponduje z istniejącą zewnętrzną miarą tego konceptu. W literaturze badano zgodność danych Google'a z wynikami telefonicznych badań sondażowych stosując analizy korelacji (Scharkow i Vogelgesang 2011; Zhu, Wang, Quin i Wu 2012; Mellon 2013b). Mellon (2013b) opisał warunki konieczne, aby zapytania z wyszukiwarek internetowych mogły być wykorzystywane jako miary istotności problemu dla opinii publicznej. W analizie empirycznej posłużył się dwoma rodzajami danych. Tematy, które przez użytkowników badania sondażowego Gallup zostały wymienione jako najważniejsze zostały porównane z częstością wyszukiwań tych samych tematów w Google. Badania objęły dane za okres 6 lat w Stanach Zjednoczonych. Przeprowadzona analiza regresji wykazała silne zależności pomiędzy zmiennymi, a otrzymany model regresji tłumaczył 66%–76% zmienności zmiennej wyjaśnianej (częstości wyszukiwania danego hasła). W Polsce „ważność" wielu spraw jest systematycznie badana przez ośrodki zajmujące się sondażami opinii publicznej i dane te mogą posłużyć jako zmienne kryterialne wobec danych z wyszukiwarki Google.

(d) Trafność konstruktu *(construct validity)* – postulujemy, że dane Google spełniają kryterium trafności teoretycznej, jeśli rezultaty analiz potwierdzają teorię problemu badawczego, nad którym pracujemy. Zwykle trafność teoretyczna ustalana jest poprzez analizę czynnikową. Stosując ten rodzaj analizy stawiamy hipotezę, iż częstości konkretnych wyszukiwań wskazują na złożoną

zmienną, która jest właśnie mierzona za pomocą wskaźników. Na przykład, na podstawie danych Google'a zebranych przez Annę Turner (2017) poddaliśmy weryfikacji hipotezę, iż ustalone dla 57 krajów częstości wyszukiwań takich haseł, jak Snowden, Assange, WikiLeaks, NSA wyrażają zainteresowanie wyciekiem danych (*data leaks*). Istotnie, częstości te okazały się ze sobą względnie silnie skorelowane i analiza czynnikowa potwierdziła zasadność hipotezy, iż konstrukt „wyciek danych" jest trafny w sensie *construct validity*[10].

Dwa przykłady

Problem, czy warto używać danych Google'a w analizach naukowych, jest w zasadzie problemem empirycznym – warto, pod warunkiem, iż uzyskujemy znaczące i wzbogacające naszą wiedzę rezultaty. Jeżeli wyszukiwanie informacji w wyszukiwarce Google mierzy stopień zainteresowania danej populacji i (i = 1, 2, 3, ... , N) zjawiskiem Y, to należy oczekiwać, że znormalizowane częstości wyszukiwań $p(Y)_i$ korelują się z jakimiś istotnymi charakterystykami populacji X_i. Podamy dwa przykłady analiz, które spełniają ten warunek, a odnoszą się do innych populacji – wewnątrz jednego kraju (województwa w Polsce) oraz różnych krajów (próba 57 krajów).

Przykład 1. Dla badacza *przedsiębiorczości* w Polsce istotne jest to, ile wniosków o fundusze europejskie złożono w poszczególnych województwach. Dzięki narzędziom Google badacz ma możliwość sprawdzenia, czy mieszkańcy tych województw korzystali z wyszukiwarki jako źródła informacji na ten temat. Tabela 1 pokazuje, jak często szukano hasła *fundusze europejskie* w okresie 2014–2016 w poszczególnych województwach; wynik przedstawiono jako średnią miesięcznych wyszukiwań i średnią tę zważono przez współczynnik populacji internautów[11]. Ostatnia kolumna zawiera liczbę wniosków o fundusze europejskie złożonych w okresie 2014–2016.

[10] Dane do tej analizy obejmują okres od kwietnia 2013 do marca 2015 roku i dotyczą 57 krajów, co przy 24-miesięcznych pomiarach daje 1368 jednostek obserwacji. Wyszukiwania haseł były dokonywane w językach lokalnych. Korelacje między wyszukiwaniem poszczególnych haseł są wysokie: $0{,}435 \leq r \leq 0{,}862$. Analiza wykazała, iż składowe (komponenty czynnikowe) są również wysokie: Snowden 0,829, Assange 0,891, WikiLeaks 0,874 i NSA 0,720. Całkowita wyjaśniona wariancja: 69,1%.

[11] Średnia miesięczna liczba wyszukiwań hasła *fundusze europejskie* została zatem zważona przez liczbę użytkowników Internetu biorąc pod uwagę ich zróżnicowanie liczebnościowe w poszczególnych województwach, co można wyrazić wzorem $W_i = \frac{frP_i}{frS_i}$ gdzie W_i oznacza wagę dla i-tego województwa, frP_i częstość cząstkową korzystających z Internetu w województwie *i,* a frS_i częstość cząstkową liczby wyszukiwań hasła *fundusze europejskie* w tym województwie, gdzie $\sum_{i=1}^{16} frP_i = \sum_{i=1}^{16} frS_i = 1$.

Tabela 1. Liczba wyszukiwań dotyczących funduszy europejskich w Google a liczba złożonych wniosków o fundusze europejskie

Województwo	Średnia miesięczna liczba wyszukiwań zapytania: *fundusze europejskie* w Google w latach 2014–2016		Liczba wniosków o fundusze europejskie złożonych w latach 2014–2016***
	Dane źródłowe*	Dane ważone**	
Dolnośląskie	1300	1421	847
Kujawsko-pomorskie	720	937	84
Lubelskie	880	981	569
Lubuskie	210	500	54
Łódzkie	1000	1130	716
Małopolskie	1900	1556	776
Mazowieckie	5400	2686	1530
Opolskie	210	480	289
Podkarpackie	720	989	1114
Podlaskie	390	588	934
Pomorskie	1000	1160	1022
Śląskie	1600	2185	1886
Świętokrzyskie	390	562	1426
Warmińsko-mazurskie	390	671	679
Wielkopolskie	1600	1619	2017
Zachodniopomorskie	590	834	152

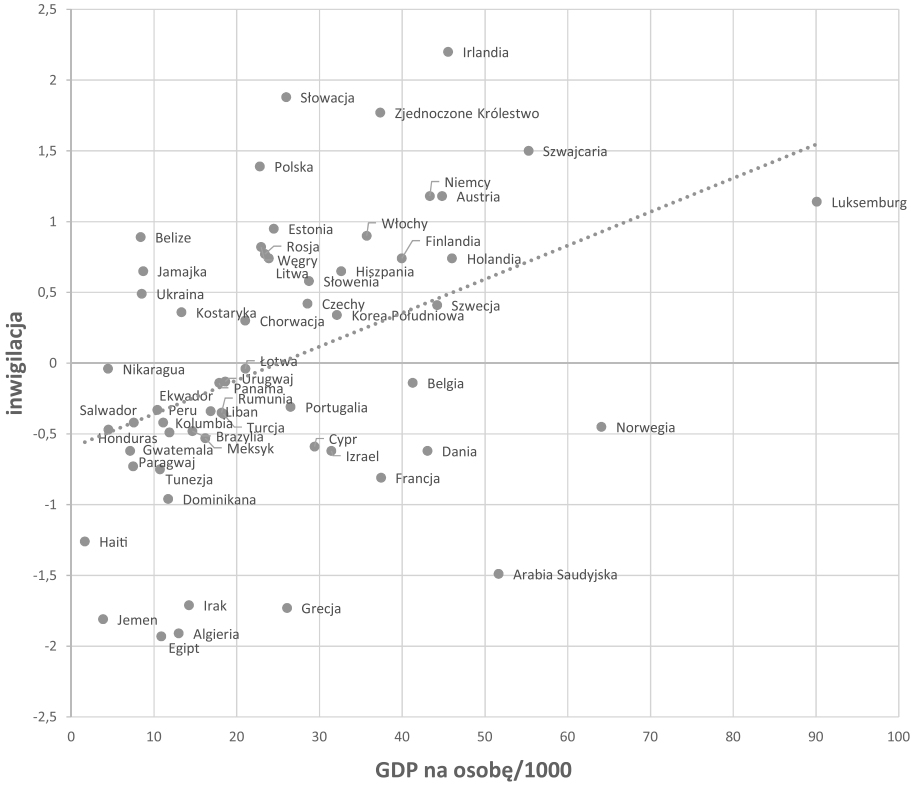
* Liczba wyszukiwań w Google Planer Słów Kluczowych.
** Liczba wyszukiwań w Google Planer Słów Kluczowych ważona przez liczbę internautów zamieszkujących dane województwo.
*** Liczba wniosków o fundusze europejskie, dane opublikowane przez Ministerstwo Rozwoju (2014).

Posiadając dane z dwóch niezależnych źródeł, można przeprowadzić analizę statystyczną regresji pomiędzy częstością zapytań o fundusze europejskie $p(Y)_i$ a liczbą złożonych wniosków o te fundusze X_i, gdzie i odnosi się województw. W omawianym przykładzie współczynnik korelacji Pearsona między tymi zmiennymi wynosi 0,64 i jest istotny statystycznie ($p < 0,01$). Niezależnie od tego, czy uznamy tę zależność za trywialną, czy nie – wskazuje ona na to, iż wyszukiwania w Google jako miara zainteresowania danej populacji specyficzną sprawą jest związana z zachowaniami tej populacji. Podkreślamy przy tym, że chodzi tu o związek na poziomie populacji, nie zaś jednostek.

Przykład 2. Turner (2017) zebrała dane na temat zainteresowania ludzi inwigilacją w różnych krajach. Używając tych danych, na podstawie wyszukiwań

takich haseł, jak „data security", „Internet surveillance" oraz „data protection law", stworzyliśmy indeks zainteresowania inwigilacją p(Y)i oraz sprawdziliśmy, czy ta zmienna zależy od zamożności kraju, mierzonego narodowym dochodem per capita Xi, gdzie indeks i odnosi się do krajów[12]. Wykres 3 obrazuje tę zależność.

Wykres 3. Zależność między dochodem narodowym na osobę a stopniem zainteresowania inwigilacją w różnych krajach



[12] W przypadku haseł „data security", „Internet surveillance" oraz „data protection law" dane, których użyliśmy (Turner 2017), były zbierane w 57 krajach w językach lokalnych. Częstości wyszukiwania tych haseł zostały poddane analizie głównych składowych. Dla „data security" i „data protection law" składowe (komponenty czynnikowe) są wysokie, $s \geq 0{,}906$. Mimo iż składowa dla „Internet surveillance" jest niska $s = 0{,}211$, całkowita wyjaśniona wariancja wynosi 57,6%. Tak więc przyjęliśmy, że częstości wyszukiwania omawianych haseł odzwierciedlają stopień zainteresowania inwigilacją i stworzyliśmy indeks ze średnią 0 i odchyleniem standardowym 1.

Istotnie, zgodnie z oczekiwaniami, opartymi na teoretycznych argumentach Ronalda Ingleharta (1977) większe zainteresowanie inwigilacją występuje w krajach o większym dochodzie na osobę. Ogólnie korelacja jest pozytywna i istotna statystycznie. Kraje o dochodzie na osobę powyżej 35 tysięcy dolarów na ogół odznaczają się większym zainteresowaniem inwigilacją. W tej grupie najbogatszych krajów powyżej jednego odchylenia standardowego na skali inwigilacji mają: Luksemburg, Szwajcaria, Austria, Niemcy, Irlandia i Wielka Brytania. Ale w tym samym przedziale skali inwigilacji znajdują się też kraje biedniejsze: Polska i Słowacja. Z kolei w grupie krajów najbogatszych można odnotować kilka przypadków niskiego zainteresowania inwigilacją, czego najbardziej wymownym przykładem jest Arabia Saudyjska i Norwegia. Tak więc, aby wyjaśnić skonstruowaną zmienną na podstawie wyszukiwarki Google należy wprowadzić inne zmienne.

Podane przykłady nie mają na celu analizy merytorycznej. Są one tylko ilustracją, w jakim kierunku mogą pójść przyszłe badania. Ogólnie rzecz biorąc, analiza konstruowanych zmiennych internetowych określonych na poziomie części kraju (np. województwa), czy na poziomie wyższego rzędu (np. kraju) wymaga posługiwania się dodatkowymi zmiennymi, pochodzącymi z innych źródeł, ale zdefiniowanymi na tym samym poziomie. Dla Europy statystyki regionalne są rozbudowane i mogą być z powodzeniem używane wraz z danymi internetowymi; źródłem może być baza danych Eurostatu (Eurostat 2018) W analizach krajowych można wykorzystywać różnorakie bazy danych, od tych, które produkowane są przez międzynarodowe agencje – typu UNESCO czy OECD – po takie, które pochodzą od organizacji akademickich – jak *Handbook of Social and Political Indicators*. Zwracamy też uwagę na szersze możliwości, mianowicie łączenia danych internetowych na poziomie krajowym z danymi surveyowymi. W projekcie *Survey Data Recycling* (Tomescu-Dubrow i Słomczyński 2017; Słomczyński i Tomescu-Dubrow 2018) poddano częściowej harmonizacji 1721 surveyów z 142 krajów (dane te są dostępne, Harvard Dataverse 2017).

Ograniczenia danych Google'a i możliwe błędy analiz

Potencjał badawczy danych internetowych jest ogromny, niemniej jak już pisaliśmy, głównym wyzwaniem jest to, że nie powstały one w wyniku działania narzędzi zaprojektowanych w celu tworzenia danych, które można wykorzystać do analizy naukowej (Lazer, Kennedy, King i Vespignani 2014a). Dyskusja dotyczy kilku problemów, między innymi nadal podnoszona jest kwestia reprezentatywności (Nagler i Tucker 2015), co wynika z prostego faktu, że nie wszyscy korzystają z Internetu, np. w Polsce 80% populacji ma dostęp do Internetu

i 90% używa wyszukiwarki jako źródła informacji przynajmniej raz w tygodniu, w Unii Europejskiej ta średnia wynosi odpowiednio 85% i 88% (Eurobarometer 2016). A zatem jeśli badamy próbę stworzoną z danych Google'a, zakładamy, że tylko zainteresowanie użytkowników Google jest mierzone i automatycznie tych, którzy Google nie używają, nasze badanie nie uwzględnia (Nagler i Tucker 2015), w efekcie nie wiemy, czy wyniki mogą być reprezentatywne dla całej populacji.

Zostały podjęte próby sprawdzenia tej kwestii, poprzez korelację wyników z telefonicznych badań sondażowych dotyczących ważnych problemów z ilością wyszukiwań tych samych tematów w Google w tym samym czasie (Scharkow i Vogelgesang 2011; Zhu, Wang, Quin i Wu 2012; Mellon 2013b). Wyniki potwierdziły istnienie korelacji jedynie pomiędzy niektórymi tematami, jednak – jak podkreślają autorzy – taki rezultat nie musi oznaczać braku reprezentatywności danych z wyszukiwarek. Wskazuje on tylko, że nawet określenie problemu jako ważnego nie zawsze przekłada się na szukanie informacji na ten temat w Google.

Należy także nadmienić, że pojawiają się głosy podważające zasadność wymogu reprezentatywności względem badanych populacji. Renomowani autorzy, Viktor Mayer-Schönberger i Kenneth Cukier, w książce *Big Data – rewolucja, która zmieni nasze myślenie, pracę i życie* (2013) twierdzą, że przywykliśmy myśleć o próbie reprezentatywnej jak o czymś uniwersalnym, gdy tymczasem koncepcja ta ma zastosowanie do rozwiązania konkretnego problemu, w konkretnym czasie i w konkretnych warunkach. Zamiast prób reprezentatywnych mamy całe zbiorowości. I tak w analizach danych pochodzących z wyszukiwarki Google, dotychczasowe założenie N = „próba reprezentatywna" zostaje zastąpione przez N = „wszyscy, którzy w danej sprawie chcą się czegoś dowiedzieć". To „wszyscy" należy rozumieć jako wszyscy z danej populacji: społeczności lokalnej, regionu kraju, całego kraju czy nawet regionu świata.

Bodaj najsłynniejszą porażką z użyciem danych z wyszukiwarek jest program Flu Trends, flagowy projekt Google'a (nie jedyną, zobacz także: Tran, Andel, Niederkrotenthaler, Till, Ajdacic-Gross i Voracek 2017). W 2008 roku zespół Google'a opublikował na łamach czasopisma „Nature" wyniki analiz przeprowadzonych w Stanach Zjednoczonych, które opierały się na prostym założeniu, że zanim chorujący odwiedzi lekarza (a jego wizyta zostanie zarejestrowana w systemie), będzie szukał informacji w Google, wpisując objawy grypy (Ginsberg, Mohebbi, Patel, Brammer, Smolinski i Brilliant 2009). Porównano więc ilość zapytań z ilością wizyt i wyniki prac okazały się absolutnie przełomowe: z dwutygodniowym wyprzedzeniem, z dużą dokładnością przewidywano wzrost zachorowań. Wynik ten był niesłychanie istotny nie tylko dla epidemiologów, ale także dla wszystkich badaczy zainteresowanych Big Data. Projekt był szeroko komentowany w mediach na całym świecie – okazało się

bowiem, że zgromadzone terabajty informacji internetowych mogą być użyteczne. I tak było aż do roku 2014, kiedy na łamach „Science" opublikowano artykuł jednoznacznie dokumentujący nie tylko mocno przeszacowaną w Flu Trends liczbę przypadków grypy w latach 2012–2013 (o prawie 140%), ale również w latach poprzednich (Lazer, Kennedy, King i Vespignani 2014a). Autorzy zwracali uwagę na niedokładności w kompletowaniu bazy słów kluczowych i w samym algorytmie analizującym dane. Google zareagował na krytykę poprawiając działanie narzędzia, a Ci sami autorzy przeprowadzili kolejne analizy uzupełniające, które co prawda potwierdziły bardziej dokładne wyniki, jednak szacunki wciąż były zawyżone o prawie 30% (Lazer, Kennedy, King i Vespignani 2014b). Naukowcy postulowali, aby zespół Google'a zaangażował do prac metodologów badań ilościowych, którzy dzięki swojej wiedzy i doświadczeniu mogliby pomóc w udoskonaleniu projektu. Niestety w roku 2014 program został przerwany[13]. Google Flu to niesłychanie ciekawy projekt z ogromnym potencjałem, ale bardzo niefortunnym zakończeniem. Niepowodzenie nie wynikało z tego, że udowodniono ponad wszelką wątpliwość, jakoby na podstawie danych Google'a nie dało się przewidywać nadchodzącej epidemii grypy, zwrócono natomiast uwagę, że przyczyną rozbieżności pomiędzy liczbą wyszukiwań a faktyczną liczbą zarejestrowanych zachorowań na grypę mogą być niewłaściwe metody badawcze, w tym mechanizmy zbierania danych. Słabością Google w tym projekcie była niechęć do dzielenia się szczegółami nawet z światowej sławy badaczami ilościowymi co, jak wiemy, przyniosło bardzo niepomyślny skutek.

Modelowym przykładem takiej współpracy jest projekt zrealizowany przez naukowców z Uniwersytetu Stanforda, Uniwersytetu Columbia i Microsoftu (White, Tatonetti, Shah, Nigam, Russ i Horvitz 2013). Biorąc pod uwagę fakt, że Internet jest miejscem, gdzie często wyszukuje się informacje na temat stanu zdrowia, postawiono następującą hipotezę, użytkownicy Internetu mogą na wczesnym etapie stosowania leków dostarczyć wskazówek co do niekorzystnego ich działania, poprzez wyszukiwanie informacji na ten temat. Zbieranie takich informacji metodami tradycyjnymi (na podstawie raportów od aptekarzy, lekarzy, firm farmaceutycznych) jest długotrwałe i nie zawsze precyzyjne. Postanowiono więc skorzystać z danych internetowych. Na potrzeby testów wybrano dwa leki: antydepresant (paroksetyna) i lek obniżający poziom cholesterolu we krwi (prawastatyna). Wybór ten nie był przypadkowy, powołano się na wstępne wyniki najnowszych ówcześnie raportów, według których jednoczesne przyjmowanie tych dwóch substancji mogło powodować hiperglikemię. Chciano sprawdzić, czy pacjenci przyjmujący obydwa lekarstwa będą doświadczać symptomów hiperglikemii i wyszukiwać informacji na ten temat w Internecie,

[13] Dane są nadal dostępne na stronie internetowej Google (Google 2018c).

jeszcze zanim występowanie tego skutku ubocznego zostało oficjalnie ogłoszone w roku 2011.

Badania przeprowadzono w 2010 roku w Stanach Zjednoczonych na grupie 6 milionów internautów, którzy zgodzili się wziąć w nich udział. W przeglądarkach badanych osób zainstalowano specjalną „wtyczkę", dzięki której przez okres 12 miesięcy była zbierana historia wyszukiwań w Google, Bing i Yahoo. Zebrano 82 miliony wyszukiwań, które następnie zanonimizowano i poddano analizom za pomocą zautomatyzowanych narzędzi zbudowanych przez zespół Microsoftu. W pierwszym etapie zidentyfikowano użytkowników, którzy szukali słów związanych z symptomami hiperglikemii (nazwy symptomów zdefiniowano na podstawie literatury medycznej). W drugim etapie podzielono użytkowników na trzy grupy: (1) tych, którzy szukali nazw obydwu leków, (2) tych, którzy szukali prawastatyny, oraz (3) tych, którzy szukali paroksetyny. Następnie policzono, ilu użytkowników z każdej grupy wyszukiwało dodatkowo symptomów hiperglikemii. Analiza wyników wyszukiwania wykazała, że osoby z grupy pierwszej, które szukały paroksetyny i prawastatyny, dwukrotnie częściej dokonywały wyszukiwań związanych z hiperglikemią niż osoby z grupy drugiej bądź trzeciej, które szukały hiperglikemii i tylko jednego z leków. Wykazano także, że różnice pomiędzy grupami utrzymują się na takim samym poziomie przez okres badanych 12 miesięcy.

Uwagi końcowe

Powstanie Google zrewolucjonizowało sposób, w jaki korzystamy z sieci, a wyszukiwarka stała się w wielu kwestiach jeśli nie podstawowym, to często jednym z pierwszych źródeł informacji. Niemal dziewięciu na dziesięciu respondentów będących użytkownikami Internetu twierdzi, że przynajmniej raz w tygodniu korzysta z wyszukiwarek internetowych, aby znaleźć informacje, 57% robi to codziennie lub prawie codziennie (Eurobarometer 2016). Powodów tej popularności jest co najmniej kilka: (a) dostępność – wyszukiwarka Google jest darmowa i banalnie prosta w obsłudze; (b) trafność – zaawansowane algorytmy potrafią przeszukiwać ogromne bazy danych tak by w ułamku sekundy zwrócić jak najlepiej dopasowane rezultaty; (c) bezpieczeństwo – Google ma różne zabezpieczenia antywirusowe; (d) udogodnienia techniczne – takie, jak wyszukiwanie głosowe.

Na potrzeby artykułu sprawdziliśmy kilka narzędzi o podobnym działaniu oferowanych zarówno przez Google (Trends, Keyword Planner), jak i przez inne firmy (w szczególności Bing Keywords Research, Yahoo Gemini Keywords Planner). Przeprowadzone analizy wykazały, że Bing i Yahoo podają tylko krótki okres historii wyszukiwań (6 miesięcy w Bing i 12 miesięcy w Yahoo).

W obu przypadkach brak jest dobrze działającego API pozwalającego na automatyczne pobieranie danych. Dokumentacja wyjaśniająca działanie narzędzi jest uboga. Poza tym tylko narzędzia Google'a oferują rozbicie geograficzne na regiony, województwa czy miasta. Z tego względu wybraliśmy narzędzia Google'a, które według naszych analiz oferują najwięcej możliwości badawczych.

Czego szukamy? Tego, co nas inspiruje, bawi, ciekawi i czego potrzebujemy – rozrywki, porad, informacji. Każde poszukiwanie pozostawia ślad, który narzędzia Google'a są w stanie przetworzyć w postaci danych. Prace nad adaptacją danych Google'a na potrzeby metodologii nauk społecznych trwają i jest to proces, podczas którego metodą prób i błędów uczymy się, jak zbierać dane, jak je interpretować, jakie metody analizy stosować, jakie problemy badawcze formułować. Wiedza doświadczonych uczonych jest tu równie niezbędna, jak wkład młodych badaczy, którzy nie znają świata bez Internetu, a internetowej rzeczywistości bez Google'a.

Wyszukiwanie informacji w Google jest procesem, który może być analizowany ze względu na swe matematyczne własności, odkrywane analogicznie do tych, jakie rządzą popularnością stron internetowych. Zgodnie z zasadami rządzącymi ruchem internetowym opisanym przez fizyka Alberta-Laszlo Barabasiego (2002) większość wyszukiwań w Internecie oparta jest na małej liczbie słów kluczowych. Wszakże, do tej pory, empiryczna wiedza o tych słowach kluczowych jest nikła. Nauki społeczne mogą przyczynić się nie tylko do inwentaryzacji popularnych słów kluczowych, ale i do wyjaśnienia, dlaczego one pojawiają się w Internecie, oraz jakie mają skutki dla zachowań ludzi. Barabasi to także twórca nowej nauki sieci (New Science of Network), który „na podstawie obiektywnych danych ma nadzieję odkryć ścisłe, matematyczne prawa opisujące ludzkie zachowania, które można użyć do prognozowania ludzkiego behawioru" (Krzysztofek 2011: 126). Od instrumentalnego spoglądania na relacje międzyludzkie dystansują się socjologowie. Anthony Giddens powołuje się na warunek uniwersalności, który występuje w naukach przyrodniczych, ale nie występuje w naukach społecznych: „nie ma żadnego twierdzenia dotyczącego zachowań ludzkich, które spełniałoby ten warunek" (za Krzysztofek 2011: 127). Zygmunt Bauman natomiast powiada: „tym, co odróżnia prawdy nauk przyrodniczych od prawd nauk o społeczeństwie, jest właśnie ten niesforny, nieujarzmiony i nieusuwalny element – ludzka subiektywność. Powoduje ona, że między badaczami a tymi, których badają, istnieje *tożsamość*, nie zaś ontologiczna i epistemologiczna *opozycja*" (2014: 39). Zdaniem Kazimierza Krzysztofka „rzeczywistość społeczna nie jest i nigdy nie będzie całkowicie dwuwartościowa. Istnieje bowiem jeszcze coś takiego jak logika rozmyta (*fuzzy logic*), którą się na co dzień posługujemy: nie tylko „tak-nie", ale także „trochę tak", „nieco tak", „zapewne tak" itp. Nie zdajemy sobie sprawy, jak wiele naszej wiedzy zdobywamy za pomocą takiej logiki" (2011: 135).

Tak samo, jak uczymy się i popełniamy błędy z Small Data, popełniamy je również z Big Data (piszą o tym m.in. Nagler i Tucker 2015; Jenkins, Słomczyński i Dubrow 2016). Aby minimalizować błędy, konieczna jest interdyscyplinarność obejmująca przedstawicieli nauk społecznych, analityków danych i specjalistów w zakresie software'u, a także współpraca z firmami internetowymi, które są w posiadaniu danych o ogromnej dla nauki wartości, ale udostępniają tylko część swojego warsztatu (piszą o tym m.in. Krzysztofek 2011; Mayer-Schönberger i Cukier 2013). Potrzebna jest także dyskusja nad stworzeniem aktów prawnych, aby umożliwić jednostkom naukowym dostęp do większej ilości danych niż zwykłemu użytkownikowi Internetu, zachowując warunki prywatności i anonimowości użytkowników. Apelujemy, aby taką dyskusję podjąć w kraju i na szerszym forum międzynarodowym. Niezależnie od tego zgłaszamy pilną potrzebę stworzenia instytucjonalnych warunków do intensyfikacji prac badawczych z użyciem danych internetowych w Polsce.

Bibliografia

Askitas, Nikos i Klaus F. Zimmermann. 2009. *Google Econometrics and Unemployment Forecasting.* „Applied Economics Quarterly" 55: 107–120.

Barabasi, Albert-Laszlo. 2002. *Linked: The New Science of Networks*. New York: Perseus.

Barabasi, Albert Laszlo. 2002. *Network Science*. Cambridge: Cambridge University Press.

Batell, John. 2006. *The Search: How Google and Its Rivals Re-wrote the Rules of Business and Transformed Our Culture.* London: Nicholas Brealey Publishing.

Bauman, Zygmunt. 2014. *Rozmowy o socjologii*. Warszawa: WN PWN.

Broder, Andrei. 2002. *A Taxonomy of Web Search*. „SIGIR Forum" 36 (2): 3–10.

Carneiro, Herman Anthony i Eleftherios Mylonakis. 2009. *Google Trends: A Web-based Tool for Real-time Surveillance of Disease Outbreaks.* „Clinical Infectious Diseases" 49: 1557–1564.

Eurobarometr. 2016. *Special Eurobarometr 447: Report on Online Platforms*. http://ec.europa.eu/information_society/newsroom/image/document/2016-24/ebs_447_en_16136.pdf. Dostęp 01.03.2018.

Eurostat. 2018. *Data base*. https://ec.europa.eu/eurostat/data/database. Dostęp 01.03.2018.

Gamma, Alex, Roman Schleifer, Wolfgang Weinmann, Anna Buadze i Michael Liebrenz. 2016. *Could Google Trends Be Used to Predict Methamphetamine-Related Crime? An Analysis of Search Volume Data in Switzerland, Germany, and Austria*. „PLOS ONE". DOI: 10.1371/journal.pone.0166566.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski i Larry Brilliant. 2009. *Detecting Influenza Epidemics Using Search Engine Query Data.* „Nature" 457 (7232): 1012–1014.

Google. 2018a. *Misja firmy*. https://www.google.com/intl/pl/search/howsearchworks/mission/. Dostęp 01.03.2018.

Google. 2018b. *Google Trends*. https://trends.google.pl/trends/explore. Dostęp 01.03.2018

Google. 2018c. *Google Flu Trends Data*. https://www.google.org/flutrends/about/. Dostęp 01.03.2018.

Granka, Laura. 2010. *Measuring agenda setting with online search traffic: Influences of online and traditional media*. Paper presented at the annual meeting of the „American Political Science Association". Washington, DC, 2–5 september 2010.

Gunter, Bertie, Ian Rowlands i David Nicholas. 2009. *The Google Generation: Are ICT Innovations Changing Information-seeking Behaviour?* Oxford: Chandos Publishing.

Harvard Dataverse. 2017. *SDR documentation*. https://dataverse.harvard.edu/dataverse/harvard?q=sdr. Dostęp 01.03.2018

Heiberger, H. Raphael. 2015. *Collective Attention and Stock Prices: Evidence from Google Trends Data on Standard and Poor's 100*. „PLOS ONE". DOI: 10.1371/journal.pone.0135311.

Inglehart, Ronald. 1977. *The Silent Revolution Changing Values and Political Styles among Western Publics.* Princeton: Princeton University Press.

Jenkins, J. Craig, Kazimierz M. Słomczyński i Joshua Kjerulf Dubrow. 2016. *Political Behavior and Big Data.* „International Journal of Sociology" 46(1): 1–7.

Kaisheng, Lai, Lee Yan Xin, Chen Hao i Yu Rongjun. 2017. *Research on Web Search Behavior: How Online Query Data Inform Social Psychology.* „Cyberpsychology, Behavior, and Social Networking" 20(10): 596–602.

Komisja Europejska. 2017. *Google ukarany*. https://ec.europa.eu/poland/news/170627_google_pl. Dostęp 01.03.2018.

Krzysztofek, Kazimierz. 2011. *W stronę maszyn społecznych. Jaka będzie socjologia, której nie znamy?* „Studia Socjologiczne" 2(201): 267–283.

Lazarsfeld, F. Paul i Herbert Menzel. 1961. *On the Relation between Individual and Collective Properties*. W: A. Etzioni (red.) *Complex Organizations. A Sociological Reader.* New York.

Lazer, David, Ryan Kennedy, Gary King i Alessandro Vespignani. 2014a. *The Parable of Google Flu: Traps in Big Data Analysis.* „Science" 343(6176): 1203–1205.

Lazer, David, Ryan Kennedy, Gary King i Alessandro Vespignani. 2014b. *Google Flu Trends Still Appears Sick: An Evaluation of the 2013–2014 Flu Season*. http://doi.org/10.2139/ssrn.2408560. Dostęp 01.03.2018.

Lester, David i John Gunn III. 2013. *Using Google Searches on the Internet to Monitor Suicidal Behavior*. „Journal of Affective Disorders" 148(2–3): 411– 412.

Lovink, Geert. 2009. *Society of the query: The Googlization of our lives*. W: K. Becker i F. Stalder (red.). *Deep Search: The Politics of Search Beyond Google*. Innsbrück: Studien Verlag, s. 45–53.

Lui, Catherine, Panagiotis Metaxas i Eni Mustafaraj. 2011. *On the predictability of the US elections through search volume activity*. https://pdfs.semanticscholar.org/. Dostęp 01.03.2018.

Mayer-Schönberger, Viktor i Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think.* London: John Murray.

Maurer, Marcus i Thomas Holbach. 2015. *Taking Online Search Queries as an Indicator of the Public Agenda The Role of Public Uncertainty*. „Journalism and Mass Communication Quarterly" 93(3): 572–586.

Mellon, Jonathan. 2013a. *Where and When Can We Use Google Trends to Measure Issue Salience?* „Political Science and Politics" 46(2): 280–290.

Mellon, Jonathan. 2013b. *Internet Search Data and Issue Salience: The Properties of Google Trends as a Measure of Issue Salience.* „Journal of Elections, Public Opinion and Parties" 24: 45–72.

Ministerstwo Rozwoju. 2014. *Fundusze europejskie 2014-2020 – Polska na tle innych krajów*. http://www.miir.gov.pl/media/23805/FE2014-20_Polska_na_tle_innych_krajow.pdf. Dostęp 7.11.2018.

Nuti, Sudhakar V., Brian Wayda, Isuru Ranasinghe, Sisi Wang, Rachel P. Dreyer, Serene I. Chen i Karthik Murugiah. 2014. *The Use of Google Trends in Health Care Research: A Systematic Review*. „PLOS ONE". DOI: 10.1371/journal.pone.0109583.

Nagler, Jonathan i Joshua A. Tucker. 2015. *Drawing Inferences and Testing Theories with Big Data*. „PS Political Science and Politics" 48(1): 84–88.

Nowak, Stefan. 2006. *Metodologia badań społecznych*. Warszawa: WN PWN.

Page, Andrew, Shu-Sen Chang i David Gunnell. 2011. *Surveillance of Australian Suicidal Behavior Using the Internet?* „Australian and New Zealand Journal of Psychiatry" 45(12): 1020–1022.

Preis, Tobias, Helen Susannah Moat, H. Eugene Stanley i Steven R. Bishop. 2012. *Quantifying the Advantage of Looking Forward*. „Scientific Reports", DOI: 10.1038/srep00350.

Ragas, Matthew W. i Hai L. Tran. 2013 *Beyond Cognitions: A Longitudinal Study of Online Search Salience and Media Coverage of the President.* „Journalism i Mass Communication Quarterly" 90: 478–499.

Ragas, Matthew W., Hai L. Tran i Jason A. Martin. 2013. *Media-induced or Search-driven? A Study of Online Agenda-setting Effects During the BP Oil Disaster.* „Journalism Studies" 15(1): 48–63.

Ripberger, Joseph T. 2011. *Capturing Curiosity: Using Internet Search Trends to Measure Public Attentiveness.* „Policy Studies Journal" 39(2): 239–259.

Rodak, Olga. 2017. *Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: Perspektywa konstruktywistyczna*. „Studia Socjologiczne" 3 (226): 209–236.

Scharkow, Michael i Jens Vogelgesang. 2011. *Measuring the Public Agenda Using Search Engine Queries*. „International Journal of Public Opinion Research" 23: 104–113.

Sherman-Morris, Kathleen, Jason Senkbeil i Robert Carver. 2011. *Who's Googling What? What Internet Searches Reveal about Hurricane Information Seeking*. „Bulletin of the American Meteorological Society" 92(8): 975–985.

Santos, Renato. 2016. *Are Our Students Really Interested in Science? Or Does Google Trends Show a Social Desirability Bias in Brazilian Public Opinion Surveys?* „Acta Scientiae" 18 (2).

Stephen-Davidowitz, Seth. 2014. *The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data.* „Journal of Public Economics" 118(1): 26–40.

Słomczyński, Kazimierz M. i Irina Tomescu-Dubrow. 2018. *Basic Principles of Survey Data Recycling.* Chapter 43. W: T.P. Johnson, B.-E. Pennell, I. Stoop i B. Dorer (red.). *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts (3MC)*. New York: Wiley.

Thompson, Teo, S. H., Vivien K. G. Lim i Raye Y. C. Lai. 1999. *Intrinsic and Extrinsic Motivation in Internet Usage.* „OMEGA International Journal of Management Science" 27(1): 25–37.

Tomescu-Dubrow, Irina i Kazimierz M. Słomczyński. 2016. *Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling*. „International Journal of Sociology" 46(1): 58–72.

Tran S. Ulrich, Rita Andel, Thomas Niederkrotenthaler, Benedikt Till, Vladeta Ajdacic-Gross i Martin Voracek. 2017. *Low Validity of Google Trends for Behavioral Forecasting of National Suicide Rates.* „PLOS ONE". DOI: 10.1371/journal.pone.0183149.

Trevisan, Filippo. 2013. *Search Engines and Social Science: A Revolution in the Making*. *Google Forum UK*. Swindon, UK: Economic and Social Research Council.

Trevisan, Filippo. 2014. *Search Engines: From Social Science Objects to Academic Inquiry Tools*. „First Monday" 19(11) http://firstmonday.org/ojs/index.php/fm/article/view/5237. Dostęp 01.03.2018.

Trevisan, Filippo, Andrew Hoskins, Sarah Oates i Dounia Mahlouly. 2018. *The Google voter: search engines and elections in the new media ecology*. „Information, Communication i Society" 21(1): 111–128.

Turner, Anna. 2017. *Public Interest in Data Leaks and Data Surveillance Before and After Snowden. Google Big Data in Cross-National Perspective.* Warszawa: Pracownia Wydawnicza Andrzej Zabrowarny.

Vaidhyanathan, Siva. 2011. *The Googlization of Everything (And Why We Should Worry).* Berkeley: University of California Press.

Wang J, Zhang T, Lu Y, Zhou G, Chen Q, Niu B. 2018. *Vesicular stomatitis forecasting based on Google Trends.* „PLOS ONE". DOI: 10.1371/journal.pone.0192141.

White, Ryan W., Tatonetti, P. Nicholas, Shah H. Nigam, Altman B. Russ i Eric Horvitz. 2013. *Web-scale pharmacovigilance: listening to signals from the crowd.* „Journal of the American Medical Informatics Association" 20: 404–408.

Widzialni.pl. 2018. *Najpopularniejsze wyszukiwarki internetowe w Polsce i na świecie*. https://www.widzialni.pl/blog/najpopularniejsze-wyszukiwarki-internetowe-w-polsce-i-na-swiecie/. Dostęp 01.03.2018.

Zhu, Jonathan J. H., Xiaohua Wang, Jie Qin i Lingfei Wu. 2012. *Assessing Public Opinion Trends based on User Search Queries: Validity, Reliability, and Practicality.* The annual conference of the „World Association for Public Opinion Research", Hong Kong, 14–16 June 2012.

Żulicki, Remigiusz. 2017. *Potencjał Big Data w badaniach społecznych.* „Studia Socjologiczne" 3 (226): 175–207.