Kamil Wais

Synergia badań społecznych, analiz statystycznych i nowych technologii – nowe możliwości i zastosowania

Abstrakt:

Nowe technologie w konsekwentny i nieunikniony sposób zmieniają otaczającą nas rzeczywistość. Następuje to również dzięki coraz większym ilościom generowanych danych i ich nowym rodzajom. Duża część z tych danych jest wartościowym produktem, nadającym się do wykorzystania w projektach analitycznych i badawczych. Ogromnym potencjałem sprzyjającym rozwojowi badań społecznych dysponują zwłaszcza technologie internetowe. Rozwijają się również szeroko dostępne narzędzia analityczno-statystyczne. Problemem pozostaje jednak brak wysoko wykwalifikowanych, interdyscyplinarnie wykształconych analityków i badaczy. Dotyczy to szczególnie takich, którzy swobodnie łączą kompetencje informatyczno-programistyczne z umiejętnościami analityczno-statystycznymi i głębokim, humanistycznym rozumieniem problemów społecznych. Można jednak wskazać już przykłady udanego połączenia nowych technologii i metod statystycznych w służbie badań społecznych, które niosą ze sobą ogromny potencjał w dostarczeniu danych kluczowych z punktu widzenia potrzeb prowadzenia innowacyjnych badań naukowych i tworzenia polityk publicznych opartych na danych.

Abstract:

New technologies consistently and inevitably changing our reality. It also occurs due to increasing quantities of generated data and its new types. A large part of that data is a valuable product, suitable for use in analytical and research projects. Especially Internet technologies have great potential conducive to the development of social research. There are also developing widely available statistical and analytical tools. The problem that remains is the lack of highly qualified and interdisciplinary educated analysts and researchers. This applies particularly to those who freely combine IT and programming competences with analytical and statistical skills and with a deep, humanistic understanding of social problems. However, there are already examples of successful combination of new technology and statistical methods in the service of social studies, that have a huge potential in providing key data for the purpose of conducting innovative research and development of evidence-based policies.

Słowa klucze: *big data*, R, *data science*, internetowy panel badawczy, polityka oparta na danych
Key words: data, R, data science, access panel, evidence-based policy

Nowe technologie, analizy statyczne, badania społeczne mogą i często mają duży wpływ na jakość naszego życia. Nie zawsze jednak rosnący potencjał pozytywnego wpływu tych elementów jest w pełni wykorzystywany, choć rozwój współpracy w ramach dyscyplin naukowych z nimi związanych mógłby go łatwo uwolnić. Pytanie, które warto zadać, brzmi: jaki potencjał niosą te elementy i czy ich interdyscyplinarne połączenie pozwoli osiągnąć efekt synergii, którego skutki będą większe, niż prosta suma poszczególnych elementów triady. Żeby odpowiedzieć na to pytanie, warto przeanalizować poszczególne składowe pod kątem tego, co mogą one wnieść do tego synergicznego układu.

Nowe technologie jako źródła nowych danych

Nowe technologie, zwłaszcza technologie internetowe i mobilne, skutecznie przekształcają rzeczywistość, choć dyskusyjne bywa, czy jest to zmiana na lepsze. Zmiany technologiczne są jednak nieuniknione i pozostaje tylko próbować je okiełznać i wykorzystać. Trzeba jednak w tych zmianach dostrzec również rozwój źródeł i coraz szerszych strumieni generowanych danych. Danych cyfrowych, a więc stosunkowo łatwych w obróbce, z których można wydobyć nieraz niezwykle istotną i potrzebną wiedzę. Istotna jest skala tego zjawiska, populacja świata przekroczyła pod koniec 2011 roku 7 miliardów ludzi, a liczba użytkowników Internetu wynosi obecnie prawie 2,5 miliarda, a więcok. 35% ludzkości. Rośnie więc szybko i konsekwentnie liczba jednostek, które biorą udział w procesie produkcji nowych danych (Worldometers 2012; ITU 2012: 6).

Zwiększa się ilość danych generowanych nie tylko przez komputery stacjonarne, ale i coraz częściej przez urządzenia mobilne, w tym smartfony. Liczba tych urządzeń na świecie, będących w użytkowaniu w trzecim kwartale 2012 roku po raz pierwszy przekroczyła 1 miliard. Potrzeba było na to 16 lat, ale przewidywania mówią, że kolejny miliard wejdzie do użytku w przeciągu najbliższych trzech lat – do 2015 roku (Bicheno 2012). Zwiększenie się liczby użytkowników smartfonów oznacza, że więcej osób będzie korzystać z różnorodnego oprogramowania. Przewidywania mówią, że liczba pobieranych rocznie aplikacji mobilnych ma zwiększyć się z ponad 30 miliardów w 2011 roku do 200 miliardów w 2016 roku (IDC 2012). Warto pamiętać, że oznacza to, iż znacząco zwiększa się liczba osób, które poprzez codzienne użytkowanie oprogramowania desktopowego i mobilnego generują ogromne ilości danych. Część tych danych może być pełnowartościowa z punktu widzenia badacza i analityka, i co ważne są one generowane w sposób zupełnie naturalny i niewymuszony z punktu widzenia samego użytkownika. Przykładem mogą tu być serwisy WWW oraz aplikacje desktopowe i mobilne o charakterze dzienniczków umożliwiających monitorowanie i agregację danych dotyczących: kaloryczności spożywanych posiłków; spalonych kalorii; zmian w wadze ciała;

uprawianej aktywności sportowej; cyklów aktywności, odpoczynku, snu; czasu spędzonego w pracy, przed komputerem, w Internecie; gospodarowania budżetem domowym; poruszania się w przestrzeni (dane geolokalizacyjne); użytkowania samochodów; czytelnictwa książek; oglądalności filmów itp. Przykłady te można by mnożyć, rodzaj gromadzonych danych jest zależny tylko od wyobraźni twórców takich aplikacji. Można się spodziewać, że jeśli gromadzenie i przetworzenie danego typu danych będzie stanowić jakąś wartość dodaną dla użytkownika, to aplikacja do tego służąca z czasem powstanie i będzie użyteczna. W konsekwencji mamy tu do czynienia ze zjawiskiem *crowd-sourcingowego* (społecznościowego) generowania danych, których badacz nie musi wywoływać – one już istnieją. Pozostają jeszcze kwestie umiejętnego doboru próby, umowy partnerskiej z takim nowym „pre-respondentem", która pozwoli uzyskać dostęp do danych w celach badawczych na ściśle określonych zasadach oraz technicznego pozyskania i obróbki takich surowych danych.

Aplikacje mobilne są więc ogromnym źródłem różnorodnych danych, ale to nie one są określane mianem *disruptive technology*, czyli technologii, która ma potencjał wywołać istotną, zauważalną dla wszystkich zmianę rzeczywistości (NIC 2008: i). SRI Consulting Business Intelligence (SRIC-BI) zidentyfikowało 102 takie technologie i wybrało 6 z nich, które do 2025 roku mają największą szansę znacząco wpłynąć na pozycję międzynarodową i sytuację wewnętrzną USA. Wśród zwycięskiej szóstki znalazły się technologie związane z procesami starzenia się społeczeństw, magazynowaniem energii, biopaliwami, czystymi technologiami węglowymi, usługami opartymi na robotyce oraz tzw. „Internet Rzeczy" (*Internet of Things – IoT*), definiowany w raporcie jako technologia informacyjna nakierowana na zwiększenie połączeń pomiędzy ludźmi i rzeczami (NIC 2008: i-ii). Chodzi tu o zastąpienie podstawowego paradygmatu technologii informacyjno-komunikacyjnych (ICT) „kiedykolwiek, gdziekolwiek, z kimkolwiek" (*anytime, any place, anyone*), paradygmatem „kiedykolwiek, gdziekolwiek, z czymkolwiek" *(anytime, any place, anthing*). W praktyce oznacza to pojawienie się trzeciego wymiaru komunikacji: między osobami a rzeczami oraz między samymi rzeczami (ITU 2005: 2). Staje się to możliwe dzięki włączeniu w sieć Internetową dowolnych urządzeń z własnym adresem IP. Mowa więc tu nie tylko o komputerach, smartfonach czy tabletach, ale także o urządzeniach codziennego użytku: pralkach, lodówkach, systemach sterowania domem (oświetleniem, klimatyzacją, zużyciem wody), telewizorach, sprzęcie medycznym, biurowym, samochodach itp. Zgodnie z prognozami w 2020 r., jak podaje E. Bendyk, spodziewać się można już 50-100 mld urządzeń podłączonych do Internetu. Prawo Boba Metcalfe'a mówiące, że wartość sieci wzrasta proporcjonalnie do kwadratu liczby jej węzłów, wydaje się mieć zastosowanie także do jej potencjału generowania danych – „kolejne węzły, kolejni użytkownicy zwiększają potencjał całości w sposób nieliniowy" (Bendyk 2012: 21). Pomijając inne implikacje można stwierdzić, że Internet Rzeczy jest kolejnym, ogromnym źródłem danych, generowanych w sposób ciągły. Dodatkowo, będą one miały charakter obiektywny, „twardy", w odróżnieniu od danych deklaratywnych podlegającym różnym błędom subiektywnej oceny.

Rozwój źródeł danych ma, z punktu widzenia badawczo-analitycznego, dwojakie konsekwencje we wzroście wolumenu danych cyfrowych i zwiększającym się potencjale badań opartych o dane jednoźródłowe (*single-source data*). Idea danych jednoźródłowych zakłada wykorzystanie w procesie badawczym danych z różnych źródeł i różnych typów, ale generowanych przez tę samą jednostkę / użytkownika / respondenta. I tak staje się możliwe przeprowadzenie projektów badawczych opartych jednocześnie na różnych rodzajach danych np. deklaratywnych zebranych w trakcie wywiadów CAWI (*Computer Aided Web Interviewing*), zastanych pochodzących z otwartych danych publicznych oraz z pomiarów pasywnych realizowanych przez odpowiednie oprogramowanie na komputerach

stacjonarnych lub urządzeniach mobilnych. Wszystkie te dane dotyczą wówczas jednego respondenta, co pozwala między innymi zweryfikować jego deklaracje, sposób postrzegania rzeczywistości, opinię o swoich zachowaniach z rzeczywistymi zachowaniami i zdarzeniami z przeszłości. To z kolei pozwala na konstrukcję dokładniejszych wskaźników i budowę lepszych modeli predykcyjnych. Przykładowo, pytanie ankietowe o ilość czasu spędzonego średnio tygodniowo przed komputerem, będzie zawsze obarczone zrozumiałą niedokładnością i zapewne niedoszacowaniem ze strony respondenta, ale połączenie tych deklaracji z dokładnym pomiarem pasywnym z aplikacji monitorującej aktywność użytkownika komputera, da już dobry obraz faktycznych zachowań respondenta, w zestawieniu z tym, jak są one postrzegane przez niego samego. Głównym wyzwaniem w przypadku danych jednoźródłowych jest zapewnienie możliwości przypisania ich do konkretnego respondenta oraz połączenia poziomego zebranych danych, a więc konieczność istnienia wspólnego identyfikatora jednostki badania we wszystkich wykorzystywanych źródłach danych. Wrażliwą kwestią jest też sprawa ochrony prywatności osób, których one dotyczą, co w przypadku publicznie dostępnych danych zastanych, pozyskanych z Internetu, bywa nieraz kwestią dyskusyjną, która winna być wnikliwie rozpatrywana w przypadku każdego projektu badawczego (Markham, Buchanan 2012).

Wspomniane wyżej pojęcie „otwarte dane publiczne" *(open data*), jak precyzuje Nadolny, odnosi się „nie tyle do konkretnego typu danych, ile do sposobu ich funkcjonowania w obiegu publicznym, a więc dostępności, czytelności i możliwości użytkowania do własnych celów" (2012: 37). Światowym liderem otwartych danych publicznych jest amerykańska platforma rządowa data.gov, która do końca 2012 roku udostępniła już prawie 380 tys. zbiorów danych surowych i geoprzestrzennych, ponad 1,2 tys. aplikacji rządkowych, ponad 200 aplikacji obywatelskich oraz ponad 100 aplikacji mobilnych, bazujących na otwartych danych publicznych (DATA.GOV 2012). Wsparciem rozwoju inicjatyw związanych z *open-data* zajmuje się, m.in. londyński *Open Data Institute* (ODI) założony przez Timothy Bernersa-Lee (twórca *WorldWideWeb*) i Nigela Shadbolta (eksperta w zakresie sztucznej inteligencji). Misją ODI jest „pobudzanie ewolucji kultury otwartych danych w celu tworzenia ekonomicznych, środowiskowych i społecznych wartości, co z kolei będzie prowadzić do odblokowania podaży, generowania popytu, tworzenia i rozpowszechniania wiedzy służącej rozwiązaniu lokalnych i globalnych problemów". Finansowe fundamenty funkcjonowania ODI zabezpieczone są przez brytyjski rząd na okres 5 lat kwotą 10 milionów funtów (ODI 2012).

Spowodowany rozwojem nowych technologii wzrost wolumenu danych powoduje powstanie zbiorów określanych jako *big data*. Pojęcie to jest definiowane jako zbiór danych o rozmiarze tak dużym, że wykracza poza możliwości typowego oprogramowania bazodanowego w zakresie jego wczytywania, przechowywania, przetwarzania i analizowania. (Manyika i in. 2011: 1). Ta miękka definicja sformułowana przez Instytut McKinsey'a sugeruje, że mimo zwiększania się możliwości popularnych pakietów statystycznych, zawsze będzie jakaś umowna granica, poza którą wolumen danych przeznaczonych do analizy będzie tak duży, że będzie wymagał specjalnych rozwiązań statystyczno-informatycznych i specjalistycznych kompetencji analitycznych. Zwłaszcza, że przewidywany wzrost danych globalnych szacowany jest na 40% rocznie (Manyika i in. 2011: vi, 17). Biorąc to pod uwagę, nie dziwi fakt, że w 2012 roku przedsiębiorstwa zainwestowały około 4,3 miliarda dolarów na technologie związane z *big data*. Przewiduje się, że wydatki te uruchomią efekt domina w postaci nowych inicjatyw i udoskonaleń, który to efekt spowoduje kolejne inwestycje w tym obszarze. Inwestycje te osiągną w 2013 roku już 34 miliardy dolarów, by w ciągu 5 lat dojść do poziomu 232 miliardów dolarów w całości wydatkowanych na technologie związane z *big data* (Gasper 2012).

Narzędzia i kompetencje analityczno-statystyczne

Problemem staje się więc coraz częściej nie brak danych, ale brak narzędzi i kompetencji pozwalających te dane przetwarzać i analizować. Wspomniany już raport Instytutu McKinsey'a stwierdza, że w samych tylko Stanach Zjednoczonych prognozowane zapotrzebowanie w 2018 roku na specjalistów z pogłębionymi kompetencjami analitycznymi może być 50-60% większe niż będzie w stanie dostarczyć system edukacji. W liczbach bezwzględnych oznacza to brak od 140 do 190 tys. specjalistów, których kształcenie jest trudne i długotrwałe. Dodatkowo ma zabraknąć również 1,5 miliona menedżerów i analityków, którzy „będą w stanie zadawać właściwe pytania i efektywnie wykorzystywać rezultaty analiz *big data*". Autorzy raportu są przy tym przekonani, że z brakiem takich specjalistów będzie zmagać się cały świat, nie tylko USA (Manyika i in. 2011: 10-11).

Naprzeciwko tym problemom wychodzą trzy nowe trendy: rozwój i upowszechnienie języka programowania statystycznego „R", rozwój dyscyplin naukowych, takich jak *data science* oraz coraz popularniejsze masowe kształcenie i samokształcenie kompetencji analityczno-statystycznych.

Prace nad nowym językiem programowania statystycznego zostały rozpoczęte w 1993 roku na Uniwersytecie Auckland w Nowej Zelandii. Jego twórcy – Ross Ihaka i Robert Gentleman – wykorzystali doświadczenia z językiem „S" powstałym w laboratoriach Bella. W 1995 roku R został upowszechniony na wolnej licencji *open-source* (*GNU / General Public License*). Decyzja o udostępnieniu społeczności użytkowników za darmo kodu źródłowego, jak i samego oprogramowania była krytycznym momentem w rozwoju języka (Smith 2010: 2). Potencjał R leży przede wszystkim, w jego szerokich możliwościach oraz zaangażowanej społeczności użytkowników i osób pracujących nad rozwojem języka, którzy do rdzenia programu dopisują wciąż nowe pakiety funkcji przeznaczone do specjalistycznych zastosowań. R okazuje się być jednym z najbardziej zwięzłych języków programowania, pozwalającym na szybką implementację algorytmów (McLoone 2012), co pozwala na sprawne wdrażanie nowych funkcji i zastosowań. Liczba pakietów R w połowie 2001 roku w głównym repozytorium (CRAN) wynosiła około stu. Obecnie przekroczyła ona 4100, a dodatkowo około 2000 pakietów zlokalizowanych jest w innych repozytoriach. Średnio dziennie przybywają więc około 2 pakiety rozwiązujące jakieś problemy analityczne (obliczenia własne; Muenchen 2012). Jest to przyrost tym bardziej imponujący, jeśli uwzględni się fakt, że implementacja nowej metody statystycznej do komercyjnych pakietów statystycznych jak SAS i SPSS zajmuje średnio 5 lat (Muenchen 2009: 2).

Jedynym, jak się wydaje, istotnym ograniczeniem R'a jest brak bezpośredniego przystosowania do pracy z *big data*, ze względu na to, że dane, na których wykonuje obliczenia, przechowywane są w całości w pamięci operacyjnej komputera. Jest to jednak ograniczenie pozorne, które można rozwiązać stosując biblioteki funkcji dedykowane do pracy ze zbiorami typu *big data*, czy wspomnianą już wcześniej komercyjną wersję R'a, która jest w stanie na przykład prowadzić obliczenia modelu regresyjnego na bazie danych liczącej 123 miliony wierszy (Rickert 2011: 12). Niezależnie w najnowszej, podstawowej wersji R'a (2.15.2) rozszerzono ilość obsługiwanej pamięci na systemach 64-bitowych z 16 GB do 32 GB, co pozwala na pracę już z bardzo dużymi zbiorami danych i jak można przypuszczać, jest wystarczające do większości zastosowań akademickich (R-Project 2012).

Obecnie język R jest używany przez ponad 2 miliony analityków, zarówno w środowisku akademickim, jak i wiodących firmach sektora prywatnego (Google, Facebook, LinkedIn) (Rickert 2011: 16). Analiza liczby wzmianek o R, pojawiających się w publikacjach naukowych w temacie lub metodzie analizy na podstawie danych z serwisu Google Scholar, zwiększa się skokowo od około 2003 roku, przy jednoczesnym spadku popularności komercyjnych pakietów SPSS i SAS (Muenchen 2012). Wyniki badania ankietowego, przeprowadzonego wśród ponad 1,3 tysiąca data-minerów z ponad 60 krajów, wskazują, że wykorzystanie R'a rośnie i w 2011 roku był już wykorzystywany przez 47% wszystkich respondentów (RexerAnalytics 2011). Jak twierdzi Norman Nie – dawniej współzałożyciel SPSS, a dziś prezes firmy Revolution Analytics z siedzibą w Palo Alto, która specjalizuje się w dostarczaniu komercyjnych wersji R wspierających pracę z *big data* - współcześnie „R jest najbardziej potężnym i elastycznym językiem programowania statystycznego na świecie" (Smith 2010: 1).

Mamy więc do czynienia z rozwojem narzędzi przy jednoczesnym wzroście zapotrzebowania na kompetencje analityczno-statystyczne. Zapotrzebowanie to spotyka się z odpowiedzią, przejawiającą się zwiększeniem zainteresowania kursami programistyczno-statystycznymi. Widoczne jest to szczególnie w kontekście rozwoju masowej, otwartej edukacji internetowej MOOC (*Massive Open Online Course*). Największe i najbardziej zaawansowane platformy MOOC w swojej ofercie dydaktycznej mają kursy poświęcone metodom statystycznym, programowaniu w języku R oraz innym tematom związanym z analizą danych. Platforma coursera.org zrzeszająca obecnie ponad 2,1 miliona kursantów oferuje m.in. kursy: *Computing for Data Analysis, Computational Methods for Data Analysis, Scientific Computing, Data Analysis, Passion Driven Statistics, Data Management for Clinical Research, Introduction to Data Science, Statistics One, Machine Learning, Web Intelligence and Big Data*. Wszystkie oferowane kursy trwają kilka tygodni, oparte są o starannie opracowaną metodykę nauczania i prowadzone są przez uznanych wykładowców z wiodących uczelni wyższych z całego świata. W pierwszej edycji kursu *Statistics One*, wprowadzającego do statystyki i analiz statystycznych w R, a prowadzonego przez psychologa – Andrew Conway'a - profesora Uniwersytetu Princeton, wzięło udział 70 tys. kursantów. Natomiast kurs programowania statystycznego w języku R, prowadzony przez Rogera D. Penga – profesora biostatystyki w Johns Hopkins Bloomberg School of Public Health – przyciągnął do pierwszej edycji 40 tys. kursantów (COURSERA 2012a).

Oprócz platformy Coursera, kursy wprowadzające do statystyki oferuje m.in. platforma Udacity, autorstwa Sebastiana Thun'a – profesora Uniwersytetu Stanforda, współtwórcy znanego projektu Google aut jeżdżących bez kierowcy (UDACITY 2012a). Sebastian Thun wraz z Peterem Nrovigiem – dyrektorem badań Google – są twórcami kilkutygodniowego kursu „Wprowadzenie do sztucznej inteligencji", w którego pierwszej edycji, opisywanej m.in. przez New York Times, wzięło udział 160 tys. studentów z ponad 190 krajów (UDACITY 2012b; Lewin 2012). Kursy zarówno na platformie Coursera, jak i na platformie Udacity, oferowane są za darmo. Kursy statystyczne o bardziej formalnym, matematycznym charakterze znajdują się w ofercie komercyjnej platformy Aleks: *Introduction to Statistics, Business Statistics, Statistics for the Behavioral Sciences*. Ukończenie tych kursów pozwala otrzymać akredytację Amerykańskiej Rady Edukacji (*American Council on Education* – ACE), której rekomendacje uznawane są w ponad 2000 amerykańskich uczelniach (ALEKS 2012). ACE prowadzi również prace ewaluacyjne, które pozwolą przyznawać rekomendacje tej instytucji wybranym kursom ukończonym na platformie Coursera (COURSERA 2012b).

Warto tutaj wspomnieć, że te największe platformy MOOC, powstawały często jako spin-offy (spółki powstałe z inicjatywy uczelni i blisko z nimi współpracujące w celu transferu wiedzy ze świata

akademickiego) najbardziej znanych uniwersytetów i uczelni technicznych; np. edX to efekt połączonych wysiłków Uniwersytetu Harvarda i MIT (*Massachusetts Institute of Technology*), natomiast Coursera, to spin-off Uniwersytetu Stanforda (Lwin 2012). Niezależnie jednak od inicjatora danej platformy MOOC, ich potencjał jest rozpoznawany także w sektorze prywatnym. Fundacja Billa i Melindy Gatsów przyznała w listopadzie 2012 r. 12 grantów wspierających rozwój MOOC na łączną kwotę 3 milionów dolarów. Prawie 1/3 tej sumy otrzymała Amerykańska Rada Edukacji (ACE) na przetestowanie funkcjonowania akredytacji kursów MOOC i zbadania nowych modeli biznesowych w dziedzinie szkolnictwa wyższego. Pół miliona dolarów zostało przekazanych 9 uczelniom na opracowanie „wprowadzających" kursów MOOC dla przyszłych studentów (GatesFundation 2012). Jeśli rozwój MOOC spełni pokładane w nim nadzieje, to przełoży on się na wzrost wiedzy i umiejętności analityczno-statystyczno-programistycznych u szerokiego grona odbiorców tych usług edukacyjnych.

Niezależnie od potrzeby masowego kształcenia szeroko rozumianych kompetencji analityczno -statystycznych, dużą rolę w najbliższej przyszłości odegrają zapewne osoby wyspecjalizowane w nowej dyscyplinie jaką jest „nauka o danych" (*Data Science*). Według Davida Smitha – osoby uznanej przez Forbes za jedną z 10 najbardziej wpływowych w temacie *Big Data - Data Science* to połączenie umiejętności informatyczno-programistycznych z analizą statystyczną i głębokim zrozumieniem zarówno samych danych, jak i badanego problemu (*RevolutionAnalytics* 2012). Jakkolwiek pierwsze dwie składowe wymagają kompetencji o charakterze ścisłym, to ta trzecia pozostawia miejsce dla kompetencji o bardziej humanistycznym charakterze, pozwalających wniknąć w istotę problemu społecznego, którego rozwiązanie, być może, ukryte jest za właściwie dobranymi zestawami danych.

Efekt synergii w służbie badań społecznych

Powiązania pomiędzy informatyką i statystyką są naturalne i bardzo silne. Znacznie dalej od tych dyscyplin wydają się być nauki społeczne. Nawet socjologowie zajmujący się badaniami ilościowymi zdają się bardziej cenić ugruntowaną metodologię tradycyjnych technik badawczych, niż zgłębiać potencjał badań społecznych opartych na nowych technikach i nowych technologiach. A przecież, jak to dobrze ujął P. Kaczmarek-Kurczak, „świat potrzebuje ludzi, którzy potrafią znajdować właściwy kurs pomiędzy skałami na coraz bardziej burzliwym oceanie informacji. Potrzebuje nawigatorów".Tacy nawigatorzy umiejętnie poruszający się w rosnących strumieniach danych, rozumiejącynową rzeczywistość (także tę technologiczno-internetową) i potrafiący ją przystępnie objaśniać, mogliby się z powodzeniem rekrutować również (choć nie tylko) z grona badaczy społecznych, jako osób, którym nieobca jest „analiza informacji, wyjaśnianie zjawisk, tworzenie cząstkowych rozwiązań problemów, diagnoza i cierpliwe eksperymenty". Zapotrzebowanie na tego typu działalność będzie się zapewne zwiększać, gdyż „[...] umiejętność selekcji informacji i budowania z nich pewnych teorii, dotyczących sposobu funkcjonowania świata, były do niedawna cechą specyficzną naukowców, jednak dziś rośnie znaczenie tego rodzaju kompetencji w bardzo szerokim spektrum zawodów i obszarów gospodarki [...] również na najwyższych szczeblach zarządzania coraz silniej pojawia się presja na rozwój umiejętności teoretycznych i analitycznych" (Kaczmarek-Kurczak 2012: 39).

Można się spodziewać, że z czasem takie kompetencje zostaną szerzej docenione także przy tworzeniu polityk publicznych i doprowadzą do realizowania postulowanego uprawiania polityki opartej na faktach (*evidence-based policy*). Polityka ta została zdefiniowana jako taka, która „[...] faktycznie rozwiązuje problemy; jest przyszłościowa i oparta bardziej na dowodach niż na potrzebie reakcji na krótkoterminowe naciski; która zwalcza przyczyny, a nie symptomy; której miernikiem są rezultaty, a nie sama aktywność; która jest bardziej elastyczna i innowacyjna niż zamknięta i biurokratyczna [...]" (tłum. własne: UK Cabinet Office 1999). Synergiczne połączenie badań społecznych, nowych technologii i metod statystycznych może dać efekt w postaci wartościowych danych, które dadzą obiektywne fundamenty do tworzenia takich polityk i podejmowania decyzji o nie opartych (*data-driven approach*).

Interesującym przykładem projektu, który ma trudny do przecenienia potencjał w dostarczaniu wartościowej wiedzy (także dla tworzenia polityk publicznych), a jednocześnie integrującym wcześniej opisane elementy w celu realizacji badań społecznych, jest holenderski MESS Project. Jego pełna nazwa to: „An Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences". MESS to projekt początkowo finansowany ze środków krajowych, w ramach grantu Holenderskiej Organizacji Badań Naukowych (NWO), a później objęty stałym dofinansowaniem ze środków rządowych (Das 2012: 8; LISS 2012a). Głównym celem projektu MESS jest „budowa infrastruktury gromadzącej dane, która zintensyfikuje badania w naukach społecznych, oparte na nowych technologiach i badaniach ankietowych" (Das 2012: 10).

Rdzeniem projektu jest panel o nazwie LISS (Longitudinal Internet Studies for the Social Sciences), składający się z 5 tys. gospodarstw domowych (obejmujących 8 tys. osób), w których choć jedna osoba mówi po holendersku. Pozostałe gospodarstwa obejmuje inny panel poświęcony specjalnie imigrantom i ich problemom, liczący prawie 2 tys. gospodarstw (Das 2012: 17; LISS 2012c). W panelu LISS próba ponad 10 tys. adresów, na której prowadzono rekrutację probabilistyczną, została oparta o losowanie proste z operatu dostarczonego przez holenderski odpowiednik polskiego GUS-u (Statistics Netherlands). Z każdym z wylosowanych gospodarstw została podjęta próba kontaktu za pomocą wywiadu telefonicznego lub ankieterskiego, jeśli baza nie zawierała znanego numeru telefonu stacjonarnego. Jeśli kontakt telefoniczny nie został nawiązany, po 15 próbach następowała seria do 8 prób kontaktu ze strony ankietera. Duża liczba powrotów ankieterskich pozwoliła zapewnić wskaźnik realizowalności próby na poziomie 48% (Das 2012: 13-14).

W czasie pierwszej rekrutacji do panelu, przeprowadzonej w 2007 roku, około 15% gospodarstw domowych w Holandii nie posiadało dostępu do Internetu, a część nie miała dostępu szerokopasmowego. Dostęp szerokopasmowy został zapewniony we wszystkich gospodarstwach, które go nie posiadały, a tym, które nie miały również sprzętu komputerowego, dostarczono tzw. *simPC* – małe i proste urządzenia pozwalające na wypełnienie ankiet on-line i nawigowanie po nich za pomocą dużych przycisków przystosowanych do wygody osób starszych (Das 2012: 12).

Wykorzystanie panelu oparte jest na zasadzie otwartego dostępu dla świata akademickiego. Badacze, zarówno krajowi jak i zagraniczni, mogą zgłaszać swoje propozycje projektów badawczych. Projekty te, jeśli przejdą pozytywną ocenę rady naukowej, zostają zrealizowane bezpłatnie. Do czerwca 2012 r. na 144 zgłoszone propozycje, zaakceptowano już 99, a 24 jest w trakcie recenzji (LISS 2012d). Liczba zgłaszanych projektów wzrasta i pochodzi z różnych dyscyplin i różnych środowisk akademickich np. z Uniwersytetów Harvarda, Stanforda i Michigan. Wszystkie pozyskane z badań dane są rozpowszechnianie w środowisku akademickim drogą internetową (Das 2012: 14; LISS 2012b).

Połowa czasu przeznaczonego na wywiady zarezerwowana jest na główne badanie w ramach projektu, powtarzane corocznie. Jest ono w wielu miejscach kompatybilne z różnymi ogólnokrajowymi i międzynarodowymi badaniami społeczno-ekonomicznymi, co pozwala porównywać wyniki badań z różnych technik badawczych i analizować relacje pomiędzy poruszaną problematyką, po jednoźródłowym przypisaniu danych z różnych badań do pojedynczych gospodarstw domowych. Główne badanie panelowe umożliwia śledzenie zmiany i ciągły monitoring warunków życia członków panelu i sytuacji gospodarstwa domowego (Das 2012: 15). Możliwe jest również jednoźródłowe połączenie danych panelowych z dodatkowymi danymi uzyskanymi z rejestrów krajowych Statistics Netherlands, w tym tych dotyczących dochodów, wynagrodzenia, ubezpieczenia społecznego itp. Takie źródło daje możliwość reprezentatywnych studiów na danych pozbawionych braków odpowiedzi, obciążenia techniki czy nieprecyzyjności danych deklaratywnych; jednocześnie redukuje obciążenie respondenta, gdyż nie ma potrzeby gromadzenia danych, które już raz zostały zgromadzone przez administrację publiczną i mogą być na odpowiednich warunkach i za zgodą respondenta przetworzone do celów badawczych (Das 2012: 17).

Szczególnie interesujące są innowacyjne próby połączenia wywiadów CAWI z danymi uzyskanymi za pomocą samoobsługowych urządzeń dokonujących pomiarów biomarkerów np. cholesterolu lub wagi ciała i poziomu tłuszczu w wadze ciała. Pomiary te wykonywane są przy pomocy czułych urządzeń i przesyłane do bazy panelu przez Internet bezpośrednio z urządzenia, bez pośrednictwa respondenta. W ramach badania pilotażowego weryfikowano w ten sposób różnice z danymi deklaratywnymi oraz wpływ raportowania wyników częstych pomiarów respondentowi na jego zdrowie i zachowania prozdrowotne. Inną innowacją jest zastosowanie techniki badań TUR (*Time Use Research*), która pozwala na pozyskanie danych z codziennie wypełnianych kwestionariuszy i dzienniczków dziennej aktywności. Wykorzystanie smartfonów i aplikacji mobilnych pozwala na bardziej precyzyjne dane oraz na wykorzystanie danych z pomiarów pasywnych (jak dane geolokalizacyjne) lub materiały audiowizualne (Scherpenzeel, Sonck, Fernee 2012). Dalsze plany obejmują eksperymenty z wykorzystaniem urządzeń mierzących ciśnienie krwi i tzw. akcelerometrów – małych, nieinwazyjnych urządzeń noszonych przez respondenta, mierzących precyzyjnie zużytkowaną energię na różne aktywności i dostarczających danych o wzorcach aktywności fizycznej – dziennych i tygodniowych. Ma to zastosowanie np. w badaniach ograniczeń w aktywności osób starszych i jej skutkach (Das 2012: 16-17).

Koszt wdrożenia takiego projektu, szczególnie w początkowej fazie, wymaga znaczących środków finansowych pokrywających koszty rekrutacji, wynagrodzeń panelistów, zaplecza informatycznego i badań pilotażowych. Jednakże, jak twierdzą jego autorzy, „kiedy bierze się pod uwagę ogromną ilość interdyscyplinarnych danych, które są efektywnie pozyskiwane z dużej próby w bardzo skuteczny sposób, to jest to nadal tańsze, niż gromadzenie danych przy pomocy tradycyjnych metod. Centralne koszty eksploatacji mogłyby być znacznie mniejsze, gdyby prosić badaczy o zapewnienie własnego budżetu na ich badania. Jednakże otwarty charakter dostępu do tej infrastruktury jest jej kluczowym elementem. Innowacyjne pomysły nie mogą być hamowane przez brak infrastruktury lub budżetu". Dlatego też, zarówno powstanie projektu, jak i koszty jego funkcjonowania pokrywane są ze środków budżetu centralnego (Das 2012: 19).

Potencjał projektu MESS został doceniony w innych krajach europejskich. Inicjatywy wzorowane na holenderskim panelu pojawiły się między innymi we Francji (*Étude Longitudinal par Internet Pour les Sciences Sociales*) i w Niemczech (*German Internet Panel*) i są aktywnie wspierane przez twórców

holenderskiego LISS (Das 2012: 20). Pozostaje pytanie, czy taki projekt ma szanse bytu w polskich warunkach. Kluczową sprawą jest tu kwestia penetracji Internetu w danej populacji, ponieważ im niższa, tym większe ryzyko skrzywienia wyników próby i/lub większe koszty rekrutacji probabilistycznej. Aby przeprowadzić porównanie penetracji Internetu, można skorzystać w tym celu ze wskaźnika odsetka gospodarstw domowych wyposażonych w łącze szerokopasmowe, który według danych Eurostatu dla Holandii wyniósł w 2012 roku 83%, a więc zauważalnie więcej niż w Polsce (67%). Przy czym trzeba pamiętać, że wdrażanie projektu MESS rozpoczęło się w 2007 roku, a jego finansowanie było zapewnione już w 2006 r. Oznacza to, że w chwili opracowywania założeń projektu łącza szerokopasmowe posiadało w Holandii 66% gospodarstw domowych, a więc odsetek porównywalny z obecną sytuacją w Polsce (EUROSTAT 2012). Analogiczny projekt mógłby więc być już w Polsce przygotowywany, zwłaszcza że projekt MESS bazował na wcześniejszych doświadczeniach z realizacją badań internetowych na panelu, realizowanych już od 2000 roku na panelu CentERpanel (Das 2012: 10).

Niestety w Polsce nie ma jeszcze ugruntowanej wiedzy i doświadczeń dotyczących tego typu projektów. Dlatego też, widoczna jest w kraju potrzeba kumulowania szerokiej wiedzy i doświadczeń, łączących interdyscyplinarny dorobek z pogranicza nowych technologii, metod statystycznych i badań społecznych. Taka kumulacja, umożliwi z czasem realizację w kraju nowatorskich projektów badawczych, które w pełni będą wykorzystywać istniejący potencjał technologiczny i warsztatowy a jednocześnie dostarczą wiedzy kluczowej w procesach decyzyjnych i w dalszym rozwoju nauki. Nie będzie to jednak możliwe bez kształcenia w tym kierunku pokolenia młodych badaczy, którzy będą otwarci na wyzwania narzucane przez taką interdyscyplinarność, a zarazem będą swobodniej obracać się w dynamicznym świecie nowych technologii.

Bibliografia:

ALEKS 2012, *Aleks. Course Products* [on-line], Dostępne w Internecie: http://www.aleks.com/about_aleks/course_products

Bendyk E. 2012, *Punkt przełomu. Trendy rozwojowe o zasięgu globalnym i regionalnym*, Warszawa: MGG Conferences

Bicheno S. 2012, *Global Smartphone Installed Base Forecast by Operating System for 88 Countries 2007 to 2017*, Strategy Analytics.

COURSERA 2012a, *Coursera* [on-line]. Dostępne w Internecie: https://www.coursera.org/about/pedagogy

COURSERA 2012b, *American Council on Education to Evaluate Credit Equivalency for Coursera's Online Courses* [on-line]. Dostępne w Internecie: http://blog.coursera.org/post/35647313909/american-council-on-education-to-evaluate-credit

Das M. 2012, *Innovation in online data collection for scientific research: the Dutch MESS project*, "Methodological Innovations Online" 2012, 7(1)

DATA.GOV 2012, *DATA.GOV. An Official Web Site of the United States Government* [on-line]. Dostępne w Internecie: http://www.data.gov

EUROSTAT 2012, *Households having access to the Internet, by type of connection*, European Commision. Dostępne w Internecie: http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/data/main_tables

Gasper T. 2012, *Big Data Right Now: Five Trendy Open Source Technologies* [on-line], TechCrunch. Dostępne w Internecie: http://techcrunch.com/2012/10/27/big-data-right-now-five-trendy-open-source-technologies/

GatesFundation 2012, *Massive Open Online Courses (MOOCs)* [on-line], Bill&Melinda Gates Fundation. Dostępne w Internecie:

http://www.gatesfoundation.org/postsecondaryeducation/Pages/massive-open-online-courses.aspx

IDC 2012, *Worldwide and U.S. Mobile Applications Download and Revenue 2012–2016 Forecast: The Appification of Everything Goes Global, International Data Corporation.*

ITU 2005, *The Internet of Things. ITU Internet Reports 2005. Executive Summary*, Geneva: International Telecommunication Union

ITU 2012, *Measuring the Information Society*, Geneva: International Telecommunication Union

Kaczmarek-Kurczak P. 2012, *Mistrzowie nawigacji*, „Polityka", nr 2(2).
Poradnik psychologiczny Polityki. Tom 9

Lewin T. 2012, *Collece of Future Could be Come one, Come All* [on-line], "The New York Times". Dostępne w Internecie: http://www.nytimes.com/2012/11/20/education/colleges-turn-to-crowd-sourcing-courses.html?pagewanted=all&_r=0

LISS 2012a, *CentERdata. Institute for data collection and research. LISS Panel* [on-line]. Dostępne w Internecie: http://www.lissdata.nl/lissdata/

LISS 2012b, *LISS Panel. Data Archive* [on-line]. Dostępne w Internecie: http://www.lissdata.nl/dataarchive/

LISS 2012c, *LISS Panel. About the Panel* [on-line]. Dostępne w Internecie: http://www.lissdata.nl/lissdata/About_the_Panel

LISS 2012d, *LISS Panel. Proposals. Approval rate* [on-line]. Dostępne w Internecie:

http://www.lissdata.nl/lissdata/Proposals/Approval_Rate

Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh Ch., Hung Byers A. 2011, *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute.

Markham A., Buchanan E. 2012, Ethical *Decision-Making and Internet Research: Version 2.0. Recommendations from the AoIR Ethics Working Committee. Final Draft*, Association of Internet Researchers.

McLoone J. 2012, *Code Length Measured in 14 Languages* [on-line], Wolfram Blog. Dostępne w Internecie: http://blog.wolfram.com/2012/11/14/code-length-measured-in-14-languages

Muenchen R. A. 2009, *R for SAS and SPSS Users*, New York: Springer

Muenchen R. A. 2012, *The Popularity of Data Analysis Software* [on-line]. Dostępne w Internecie: http://r4stats.com/articles/popularity

Nadolny M. 2012, *Data Economy. Gospodarka oparta na danych, Trendy rozwojowe i zmiany gospodarcze w regionie*, Warszawa: MGG Conferences

NIC 2008, *Disruptive Civil Technologies. Six Technologies With Potential Impacts on US Interests Out to 2025, Prepared by SRI Consulting Business Intelligence under the auspices of the National Intelligence Council*, Conference report, CR 2008-07.

ODI 2012, *Open Data Institute. Knowledge for everyone* [on-line]. Dostępne w Internecie: http://www.theodi.org/

Revolution Analytics 2012, *The Rise of Data Science in the Age of Big Data Analytics: Why Data Distillation and Machine Learning Aren't Enough* [on-line], Revolution Analytics. Dostępne w Internecie: http://www.revolutionanalytics.com/news-events/free-webinars/2012/rise-of-data-science

Rexer Analytics 2011, *5th Annual Rexer Analytics Data Miner Survey* [on-line]. Dostępne w Internecie: http://rexeranalytics.com/Data-Miner-Survey-Results-2011.html

Rickert J. 2011, *Big Data Analysis with Revolution R Enterprise*, Revolution Analytics.

R-Project 2012, *R News. Changes in R Version 2.15.2* [on-line]. Dostępne w Internecie: http://cran.r-project.org/src/base/NEWS.html

Scherpenzeel A., Sonck N., Fernee H. 2012, *Time use data collection using Smartphones: Results of a pilot study among experienced and inexperienced users, 6th Internet SurveyMethodology Workshop*, Ljubljana.

Smith D. 2010, *R is Hot. How Did a Statistical Programming Language Invented in New Zealand Become a Global Sensation?, Executive White Paper*, Revolution Analytics.

UDACITY 2012a, *Udacity. Introduction to Statistics (ST101)* [on-line]. Dostępne w Internecie: http://www.udacity.com/overview/Course/st101/CourseRev/1

UDACITY 2012b, *Udacity Blog. Why a Functional Verification Course?* [on-line]. Dostępne w Internecie: http://blog.udacity.com/2012/11/why-functional-verification-course.html

UK Cabinet Office 1999, *Modernising Government. Presented to Parliament by the Prime Minister and the Minister for the Cabinet Office by Command of Her Majesty* [on-line]. Dostępne w Internecie: http://www.archive.official-documents.co.uk/document/cm43/4310/4310.htm

Worldometers 2012, *Worldometers. Real Time World Statistics* [on-line]. Dostępne w Internecie: http://www.worldometers.info

Kamil Wais – doktor nauk humanistycznych, adiunkt w Katedrze Metod Ilościowych w Ekonomii Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie, specjalizujący się w badaniach społecznych wspieranych technologiami internetowymi (*access panels*, CAWI, *single-source data*, pomiary pasywne).