Maciej Brosz
Uniwersytet Gdański

Grzegorz Bryda
Uniwersytet Jagielloński

Piotr Siuda
Uniwersytet Kazimierza Wielkiego

https://doi.org/10.18778/1733-8069.13.2.01

Od redaktorów: *Big Data* i CAQDAS a procedury badawcze w polu socjologii jakościowej

Abstrakt Świat życia codziennego zalewany jest ogromną liczbą różnych skwantyfikowanych i zdigitalizowanych danych. Mogą one być przetwarzanie i traktowane jako materiał badawczy – również jakościowy. Zastosowanie w badaniach jakościowych wielkich zbiorów danych (*Big Data*) modyfikuje postępowanie na wszystkich etapach procesu badawczego: od projektowania badania aż po formułowanie wniosków końcowych. Czy osadzenie w badaniach jakościowych modelu *Big Data* prowadzi ku ateoretyczności badań? Jakie skutki niesie rezygnacja z próby losowej na rzecz kompletności danych? Celem artykułu jest zasygnalizowanie owych zmian i krótkie ich scharakteryzowanie z uwzględnieniem znaczącej roli różnego typu oprogramowania (zwłaszcza CAQDAS), a co za tym idzie – analiz, które można przeprowadzać.

Słowa kluczowe *Big Data*, CAQDAS, komputerowa analiza danych jakościowych, danetyzacja, przetwarzanie danych, badania jakościowe

Maciej Brosz, dr, socjolog, adiunkt w Zakładzie Socjologii Stosowanej w Instytucie Filozofii, Socjologii i Dziennikarstwa Uniwersytetu Gdańskiego, właściciel firmy Q&Q Zakład Realizacji Badań Społecznych. Zainteresowania naukowe: socjologia zamieszkiwania, jakość życia w środowisku miejskim, rewitalizacja zdegradowanych obszarów miejskich, komputerowe wspomaganie badań jakościowych, programowanie w języku R.

Adres kontaktowy:
Uniwersytet Gdański
Instytut Filozofii, Socjologii i Dziennikarstwa
ul. Jana Bażyńskiego 4, 80-309 Gdańsk
e-mail: maciej.brosz@ug.edu.pl

Grzegorz Bryda, dr, adiunkt w Instytucie Socjologii UJ, w Zakładzie Socjologii Komunikacji Społecznej, Kierownik Pracowni analiz i badań jakościowych CAQDAS TM Lab. Zainteresowania: teoria socjologiczna, kogniwistyka, *big data*, metodologia badań jakościowych, informatyka społeczna, CAQDAS, TextMining i NLP w analizie danych jakościowych, modelowanie procesów społecznych. Współpracuje z instytucjami publicznymi i prywatnymi w zakresie metodologii badań społecznych oraz wielowymiarowej analizy danych ilościowych i jakościowych.

Adres kontaktowy:
Instytut Socjologii, Uniwersytet Jagielloński
ul. Grodzka 52, 31-044 Kraków
e-mail: grzegorz.bryda@uj.edu.pl; pracownia.caqdastm@gmail.com

Piotr Siuda, doktor socjologii, prodziekan ds. nauki oraz adiunkt na Wydziale Administracji i Nauk Społecznych Uniwersytetu Kazimierza Wielkiego w Bydgoszczy. Interesuje się społecznymi aspektami Internetu oraz metodami badań socjologicznych. Autor kilku monografii naukowych; publikował w wielu periodykach naukowych, w tym w wysoko punktowanych pismach, posiadających *impact factor* („European Journal of Cultural Studies", „International Journal of Cultural Studies"). Koordynator szeregu projektów badawczych, w tym zadań: „Dzieci sieci", „Dzieci sieci 2.0", „Prosumpcjonizm pop-przemysłów". Członek The Association of Internet Researchers oraz Polskiego Towarzystwa Socjologicznego. Prowadzi sympozja i szkolenia z pisania artykułów naukowych. Strona domowa: http://www.piotrsiuda.com; blog: http://www.piotrsiuda.pl.

Adres kontaktowy:
Uniwersytet Kazimierza Wielkiego
Wydział Administracji i Nauk Społecznych
ul. Ogińskiego 16, 85-092 Bydgoszcz
e-mail: piotr.siuda@ukw.edu.pl

Współcześnie twierdzi się o nastaniu ery *Big Data* (por. Chen, Zhang 2014) – obserwujemy gwałtowny przyrost liczby informacji gromadzonych w formie elektronicznej, jak i rozwój technologii dotyczących pozyskiwania danych, ich zapisu oraz magazynowania w postaci repozytoriów, hurtowni, archiwów czy też różnorodnych korpusów danych tekstowych. Zmianie ulega także sposób wykorzystywania danych, tak ilościowych, jak i jakościowych. Współcześnie dane te w formie zdigitalizowanej można spotkać w każdym obszarze życia codziennego, począwszy od baz danych dotyczących transakcji bankowych, informacji z kas fiskalnych, rejestrów użycia kart kredytowych, zestawień rozmów telefonicznych, przez statystyki urzędowe, archiwa danych statystycznych, danych społecznych (sondażowych, jakościowych), aż po rejestry medyczne, biologiczne i tym podobne[1]. Warto podkreślić, iż źródłem *Big Data* jest nie tylko Internet, kojarzony zwykle z tej skali zasobami. Wielkie zbiory pozostają w domenach archiwów państwowych, instytucji samorządowych, korporacji, fabryk, przedsiębiorstw, szpitali, czyli wszędzie tam, gdzie rejestrowane są ślady ludzkiej działalności. Wraz z rozwojem *Big Data* doskonalone są narzędzia ich przetwarzania oraz analizowania (specjalne miejsce w tym artykule poświęcone zostanie narzędziom CAQDAS, czyli oprogramowaniu wspomagającemu proces analizy danych jakościowych).

Pojęcie *Big Data* ukuto w takich naukach jak astronomia czy genetyka, które jako pierwsze w XXI wieku zmierzyły się z olbrzymią ilością danych. Obecnie koncepcja *Big Data* przenosi się na wszelkie obszary działalności ludzkiej. Termin ten nie ma jednoznacznej definicji. Początkowo oznaczał taki wzrost liczby informacji, że pamięci komputerów służących do ich przetwarzania stały się niewystarczające, co zmusiło do zaprojektowania nowych narzędzi. Obecnie uważa się, że termin *Big Data* obejmuje wszystko to, co może być realizowane badawczo w dużej skali celem zyskania nowej wiedzy lub zawartości w znaczący sposób zmieniających nasze życie codzienne, choć nie ma ostatecznej zgody co

[1] W systemach *Big Data* pozyskujemy i analizujemy zarówno dane miękkie (opinie, sądy, komentarze), jak i dane twarde (zdarzenia, fakty, transakcje, zachowania).

do zasięgu owej skali. Na pewno natomiast *Big Data* stanowi wyzwanie dla naszego sposobu życia i interakcji ze światem.

Wielkie dane są także wyzwaniem dla badaczy jakościowych, gdyż ich zaistnieje modyfikuje postępowanie na wszystkich etapach procesu badawczego. Celem przedstawianego tutaj artykułu jest zasygnalizowanie owych zmian i krótkie ich scharakteryzowanie z uwzględnieniem znaczącej roli różnego typu oprogramowania (zwłaszcza typu CAQDAS). W kolejnych częściach tekstu zajmiemy się kolejnymi etapami badawczymi. Zaczniemy od projektowania badań, starając się odpowiedzieć na pytanie, czy *Big Data* i inne nowe technologie prowadzą do powstania nowego paradygmatu badawczego rzutującego na to, jak przystępuje się do badań w polu socjologii jakościowej. Następnie pokrótce omówimy narzędzia i techniki badań jakościowych opartych na *Big Data* i różnego typu oprogramowaniu, po czym przejdziemy do etapu analizy danych gromadzonych z ich wykorzystaniem.

Warto podkreślić, że przedstawiany artykuł należy traktować w kategoriach wstępu do całego prezentowanego numeru „Przeglądu Socjologii Jakościowej". Zawarte w tym numerze publikacje są pokłosiem zorganizowania jednej z grup tematycznych w trakcie XVI Ogólnopolskiego Zjazdu Socjologicznego PTS, „Solidarność w czasach nieufności", w Gdańsku we wrześniu 2016 roku (grupa „Big Data, CAQDAS i nowe technologie w polu socjologii jakościowej"). Właśnie dlatego artykuł zamyka krótkie streszczenie wszystkich tekstów z tego numeru.

Proces projektowania badań – kilka słów o nowym paradygmacie badawczym

Czy pojawienie się na badawczym horyzoncie socjologicznych dociekań jakościowych tak zwanego *Big Data* oznacza inne podejście do konstruowania badań, czy drastycznie zmienia się proces ich projektowania? Jakie istotne przekształcenia wiążą się z koniecznością wykorzystywania odpowiedniego oprogramowania (np. typu CAQDAS) umożliwiającego przetwarzanie ogromnych ilości danych? Czy użycie wspomnianych nowych technologii powoduje pojawienie się zupełnie nowego paradygmatu badawczego?

Odpowiedź jest pozytywna – *Big Data* i nowe technologie oznaczają nowe epistemologiczne spojrzenie na proces projektowania naukowych dociekań, co wynika przede wszystkim z faktu, że wiedza generowana w ich toku ma pochodzić nie z testowania różnych teorii przez odwoływanie się do odpowiednich danych empirycznych. Bezpośrednim źródłem poznania stają się same dane – to z nich wypływać ma wiedza o świecie społecznym. Zmianę tę doskonale ujął Rob Kitchin (2014), odnosząc się nie do samych badań jakościowych, ale do nauki jako takiej. Jego zdaniem mamy do czynienia z nowym paradygmatem o zakresie multidysycyplinarnym, przy czym paradygmat ujęty został tutaj nie w kategoriach Kuhnowskich, czyli jako powszechnie obowiązujący sposób postrzegania tego, w jaki sposób uprawiać naukę. Paradygmaty naukowe powstawać mają w wyniku zmiany formatów danych, z którymi badacze mają do czynienia. Rewolucje naukowe nie wynikają z wyczerpywania się starych paradygmatów, ich niemożności odpowiadania na kluczowe pytania nurtujące akademików – rewolucje są wynikiem powstawania nowych metod analitycznych (por. Hey, Tansley, Tolle 2009).

Ujmując sprawę w ten sposób, nie można zapominać, że samo pojęcie paradygmatu naukowego podlega sporej krytyce – zwłaszcza w naukach społecznych, a więc w polu mocno zróżnicowanym pod względem podejść badawczych (por. Masterman 1970). Patrzenie na owe nauki, w tym socjologię, jako determinowane przez jakiś jeden nadrzędny paradygmat zdaje się nieuprawnionym upraszczaniem bardzo zróżnicowanego obszaru, sztucznym nakładaniem ram niepasujących do wielości sposobów patrzenia na świat społeczny. Mamy zatem do czynienia z rozbieganiem się teoretycznego ujmowania nauki z tym, jak wyglądają badania naukowe w praktyce – w wypadku terminu paradygmatu nie uwzględnia się w należyty sposób procesów ciągłej ewolucji różnych nurtów akademickich.

Mimo omawianych kontrowersji, zaryzykować można stwierdzenie, że, jeśli chodzi o zaistnienie *Big Data* oraz wykorzystanie różnego typu oprogramowania, można mówić o tak znaczącej zmianie, że zastosowanie pojęcia paradygmatu jest na miejscu. Termin ten ma służyć podkreśleniu skali zachodzącej zmiany. Nowa rewolucja, którą obserwujemy, ma być rewolucją danych oraz metod analitycznych i skutkować znaczącą zmianą praktyk badawczych w ramach nauk społecznych, w tym zmianą sposobów projektowania badań.

„Śmierć teorii" czy „wzbogacenie" obecnych procedur

Bardziej radykalni zwolennicy wykorzystania nowych technologii ogłosili już nawet „śmierć teorii" (por. Anderson 2008; Pigliucci 2009; Kitchin 2014; Parks 2014). Ciągle wzrastająca liczba danych, a także idące za owym wzrostem doskonalenie technik ich analizowania (lepsze, wydajniejsze oprogramowanie) czynić mają teorię zbędną w odkrywaniu praw rządzących społeczeństwami ludzkimi. *Big Data* automatycznie produkują znaczącą wiedzę na temat skomplikowanych zjawisk społecznych – niepotrzebne jest stosowanie się do procedur procesu badawczego, łącznie z formułowaniem hipotez. Dane mówić mają same za siebie, wolne od teoretycznych ograniczeń, co powoduje, że badacze społeczni uwolnić się powinni od obsesji poszukiwania przyczynowości na rzecz poszukiwania korelacji – mniej istotne staje się odpowiadanie na pytanie „dlaczego?", na znaczeniu natomiast zyskuje szukanie odpowiedzi na pytanie „co?". Celem ma być nie tyle odkrycie przyczyn zjawisk i procesów, lecz powiązań, relacji między nimi. Zbędne jest konstruowanie teoretycznych modelów – nauki społeczne powinny poświęcić się zwiększaniu jakości zbieranych (wielkich) danych oraz doskonaleniu narzędzi ich analizowania. Warto podkreślić, że tego rodzaju podejście dominować zaczyna przede wszystkim w świecie biznesu i badaczy zachowań konsumenckich[2] – kiedy przeniesiemy

[2] Dzieje się tak z prostego powodu – patrzenie na dane jako „mówiące same za siebie" stanowi dla biznesu świetną podstawę sprzedaży swoich produktów. *Big Data* mają oferować uzyskanie dogłębnej, obiektywnej i przynoszącej zyski informacji bez zaangażowania nauki oraz naukowców. Dobrym przykładem są systemy polecania produktów klientom sklepów internetowych. Weźmy chociażby internetowe księgarnie, gdzie poszczególnemu użytkownikowi „podsuwa się" książki nie w oparciu o czynniki kulturowe czy konwenanse związane z czytaniem, ale w oparciu o wzorce zakupowe wszystkich konsumentów danej e-księgarni. To, czy danej osobie X spodoba się konkretna pozycja książkowa domniemywa się w oparciu o śledzenie zwyczajów zakupowych innych osób kupujących podobne (lub takie same) książki jak ta osoba. Stwierdzenie takich nabywczych uwarunkowań jest w tym wypadku celem – zupełnie zbędna jest wiedza na temat powodów występowania takich, a nie innych zależności.

je na pole rozważań akademickich, w tym na pole socjologii jakościowej, wydaje się ono wiązać z wieloma uproszczeniami.

Wielkie dane nie powstają przecież znikąd – są zawsze wynikiem działań ludzkich ukierunkowanych na zdobywanie konkretnych informacji, a używane sposoby analizowania, a także wykorzystywane algorytmy zależą od decyzji konkretnych badaczy. Wykrywanie zależności widocznych w wypadku konkretnych danych nie zachodzi zatem w próżni i zawsze jest wynikiem wcześniejszych odkryć, teorii, ale też doświadczeń czy umiejętności poszczególnych osób. Dane (nie tylko te wielkie) nigdy nie mówią same za siebie, nie mogą się „oswobodzić" od interpretacyjnych ram nakładanych na nie przez badaczy. Interpretacje zależą zaś od indywidualnych predyspozycji, przekonań czy postaw naukowców i nawet jeśli proces zbierania oraz analizowania danych jest mocno zautomatyzowany, jest on zawsze „osadzony" w konkretnych wartościach i kontekstualizowany w obrębie danego pola badawczego. Pozyskiwanie *Big Data* i wykorzystywanie różnych narzędzi służących ich „obróbce" nie jest procesem tak obiektywnym, jak moglibyśmy sądzić. Interpretacje są przecież także wynikiem decyzji dotyczących tego, jak owe dane zbierać i analizować, jakie nowe technologie wykorzystywać i tym podobne. Ze względów opisanych wyżej powinniśmy patrzeć na *Big Data* zupełnie inaczej niż proponują to zwolennicy tezy o „śmierci teorii". Jak przekonuje wspomniany wcześniej Kitchin (2014), zmiana dotycząca nowego paradygmatu rzeczywiście polega na uzyskiwaniu zrozumienia danego zjawiska w oparciu o dane, a nie w oparciu o teorię, ale nie traci się jej z oczu. Do projektowania badań podchodzi się w sposób indukcyjny (od szczegółu do ogółu), choć wyjaśnianie przez indukcję nie jest końcem procesu badawczego. Można powiedzieć, że wykorzystanie *Big Data* jest dopiero wstępem do formułowania hipotez i wdrażania metody dedukcyjnej (od ogółu do szczegółu)[3]. *Big Data* służą identyfikowaniu potencjalnych pytań badawczych, mających być potem weryfikowanymi w toku dalszych badań. Dużą rolę gra w tym wypadku sama teoria – to jak dane są generowane i jak zostaną użyte wynika z przyjęcia pewnych założeń podpartych wiedzą teoretyczną. To ona podpowiada, jak należy podjeść do konkretnych danych, aby uzyskać wartościowe informacje. Podsumowując, można powiedzieć, że mamy do czynienia ze zmodyfikowaniem tradycyjnego procesu badawczego w taki sposób, aby uwzględniał on nową drogę budowania teorii – drogę uwzględniającą wykorzystanie *Big Data*.

Przy okazji warto zaznaczyć, że badacz jakościowy podejmujący ową drogę i projektujący dane badanie musi zdawać sobie sprawę z konieczności „otworzenia się" na inne dyscypliny naukowe. Wydobycie wartościowych informacji z wielkich danych ze zrozumiałych względów wymaga interdyscyplinarnego podejścia w sferze metodologii. Socjolog pracujący z *Big Data* może nie być świadomy możliwości tkwiących w różnego typu oprogramowaniu – z pewnością przydatna jest w tym wypadku wiedza informatyczna czy statystyczna. Potrzebę interdyscyplinarności widać zresztą także na poziomie teoretycznym – z racji tego, że wielkie dane bardzo często są tak bogate w szczegóły, że ich analiza i interpretacja odwoływać się musi do teoretycznych doświadczeń wielu dyscyplin.

[3] Można zatem w tym względzie zauważyć pewne analogie do klasycznej teorii ugruntowanej.

„Jasne" strony nowego paradygmatu

Przy tym wszystkim przy projektowaniu badań wykorzystujących *Big Data* w polu socjologii jakościowej należy być świadomym, że dane te mogą mieć bardzo duże znaczenie dla rozwiązania powszechnie znanego dylematu metodologicznego. Chodzi o zapewnienie realizmu badawczego z jednej strony, a z drugiej strony o zachowanie kontroli nad warunkami przeprowadzania badania (por. Chang, Kauffman, Kwon 2013). Zwykle wybór konkretnej metody badawczej wiąże się z „opowiedzeniem się" albo za realizmem, albo za ściślejszą kontrolą. Na przykład obserwacja socjologiczna o charakterze uczestniczącym niejawnym (badacz staje się pełnoprawnym członkiem obserwowanej grupy, nie informując jej członków o tym, że są przedmiotem obserwacji) zapewnia wysoki realizm badania, natomiast zupełnie niemożliwe staje się kontrolowanie warunków, w jakich ono się odbywa. Rzeczy mają się odwrotnie, jeśli chodzi o, na przykład, tradycyjne eksperymentalne badania laboratoryjne, gdzie dąży się do wyeliminowania przypadkowości oraz zredukowania wpływu czynników zewnętrznych mogących zniekształcać otrzymane wyniki. Wykorzystanie *Big Data* oraz różnego oprogramowania wspierającego ich analizę umożliwia zażegnanie opisywanego konfliktu. Wielkie dane można zbierać w oparciu o wcześniej ustalone założenia teoretyczne, skupiać się na różnych ich aspektach, manipulować zmiennymi tak, aby uzyskać dane, na których akurat zależy badaczom. Można zatem tak zaprojektować badanie, aby znaleźć dane spełniające wcześniej sformułowane założenia eksperymentalne, a więc uzyskać sporą dozę kontroli. Jedocześnie w wypadku badań z wykorzystaniem *Big Data* nie ma potrzeby „kopiowania" rzeczywistego świata społecznego w sztucznym otoczeniu. Możemy „obserwować" zjawiska przebiegające w naturalny sposób, zbierać informacje reprezentujące ludzkie działania i interakcje, gromadzić cyfrowe ślady ludzkiej działalności – na przykład tweety, opinie internautów, kliknięcia na aukcjach sieciowych i tak dalej. Wykorzystanie *Big Data* pozwala uzyskać dogłębną wiedzę o jednostkach czy społecznościach; możliwe staje się także uchwycenie dynamiki wielu różnych zjawisk społecznych[4]. Patrząc ogólnie na jakościowe badania socjologiczne oparte na *Big Data*, warto zaznaczyć, że badacze uzyskują dostęp do danych zupełnie nowego rodzaju, a także korzystają z rozlicznych ułatwień w dostępie do danych do tej pory trudno osiągalnych. Ponadto można nimi łatwo zarządzać przy pomocy różnych narzędzi, na przykład oprogramowania typu CAQDAS. Niejednokrotnie wykorzystanie omawianych

[4] Warto podkreślić, że o przydatności *Big Data* można mówić na trzech poziomach socjologicznych dociekań – na poziomie makro, mezo oraz mikro, chociaż dla badacza jakościowego istotne są w tej mierze dwa ostatnie poziomy. Jeśli chodzi o pierwszy, najszerszy w swoim zasięgu, wielkie dane pomóc mogą chociażby w odkrywaniu wzorów międzynarodowego przepływu ludności (na przykład migrantów), ale też zależności między państwami, gospodarkami czy też różnymi sektorami przemysłu. Analizy na poziomie mezo wykorzystujące *Big Data* mogą skupiać się na śledzeniu zachowań poszczególnych jednostek używających urządzeń mobilnych na danym obszarze geograficznym; innym przykładem jest zbieranie danych na temat komunikacji i zachowań użytkowników poszczególnych portali typu *socialnetworking* (np. Facebook). Mikroanalizy natomiast obejmować mogą zagadnienia dotyczące aktywności sieciowej poszczególnych internautów (np. tekstualne wzorce wykrywane we wpisach na danym blogu).

nowych technologii w polu badań jakościowych uzupełnia i dopełnia bardziej tradycyjne podejścia badawcze, umożliwia konstruowanie badań o znacznie szerszym zasięgu oraz takich, które udzielają odpowiedzi na pytania, na które odpowiedzi nie udałoby się uzyskać bez pomocy owych technologii (por. Shah, Cappella, Neuman 2015: 9).

Warto zaznaczyć, że mimo zasygnalizowanych wyżej niewątpliwych zalet *Big Data* wielu badaczy ma negatywne nastawienie, jeśli chodzi o możliwość ich wykorzystania w badaniach jakościowych (por. Ramsay 2010). Trzeba bowiem pamiętać, że *Big Data* mogą mieć charakter redukcjonistyczny. Rezultatem ich wykorzystania mogą być analizy ignorujące szerszy kontekst społeczny, na przykład uwarunkowania kulturowe konkretnych zjawisk społecznych. Mamy zatem do czynienia ze „słabą", jedynie powierzchowną analizą, zamiast z wnikliwym poznaniem danego zagadnienia – głębia zrozumienia ma być zastąpiona skalą danych redukujących skomplikowane i wielowymiarowe struktury społeczne do liczb. Dobry przykład podał w swoim artykule wspomniany wcześniej Kitchin (2014), kiedy wspomniał o projekcie analizującym język używany przez użytkowników serwisu internetowego Twitter. Stworzona przez badaczy swoista językowa mapa Twittera pokazała wzorce geograficznej koncentracji różnych społeczności etnicznych w mieście Nowy Jork. Dociekania nie były jednak w stanie odkryć sposobów, w jaki powstają takie zgromadzenia, a także tego, jakie są konsekwencje ich istnienia. Rozstrzygnięcie tych kwestii wymaga przecież oparcia się na teorii oraz „głębokiej" wiedzy kontekstualnej. Podany przykład ma potwierdzać, że zjawiska społeczne są zbyt skomplikowane oraz przypadkowe, aby dało się je zredukować do praw i formuł. Ludzie często nie zachowują się w sposób racjonalny – ich życie pełne jest sprzeczności, paradoksów oraz nieprzewidzianych wydarzeń. Dodatkowo ogromne zróżnicowanie stylów życia czy kultur powoduje, że redukowanie zróżnicowanego świata społecznego do uniwersalnych modelów nie jest uprawnione.

Przyznając takiemu podejściu rację, należy od razu zaznaczyć, że nie można twierdzić, że analizy dokonywane w oparciu o *Big Data* są pozbawione wartości. Jest wręcz odwrotnie, choć rzeczywiście należy uznać, że wytworzone informacje są bardzo specyficzne. Potrzebują osadzenia w kontekście społecznym, a także tego, o czym mowa była wcześniej, czyli oparcia w teorii. Potrzebują być może także wsparcia się na informacjach wytworzonych w toku badań prowadzonych zgodnie ze starym paradygmatem, czyli tych bazujących na danych gromadzonych w znacznie mniejszej skali. Należy przecież pamiętać, że *Big Data*, CAQDAS oraz wszelkie nowe technologie nie stanowią o radykalnym zerwaniu z przeszłością badań jakościowych. Można tutaj raczej mówić o metodologicznym postępie, nad charakterem którego wciąż trzeba się zastanowić. Cały czas bowiem zachodzi potrzeba szerszej krytycznej refleksji nad epistemologicznymi konsekwencjami użycia *Big Data*, zwłaszcza w polu badań jakościowych. Poza wszelką wątpliwością jest tylko to, że omawiany nowy paradygmat opiera się na dostępności nowych narzędzi i rozwiązań technologicznych umożliwiających proces analizy danych – to właśnie tymi narzędziami i rozwiązaniami zajmiemy się bliżej w kolejnej części tekstu.

Metody i techniki prowadzenia badań z wykorzystaniem *Big Data* i innych nowych technologii

O metodach i technikach prowadzenia badań z wykorzystaniem *Big Data*, CAQDAS i innych nowych technologii z pewnością można orzec, iż jest to obszar zróżnicowania oraz przenikających się tradycji i nowych rozwiązań lub idei. Stąd też próba uchwycenia stanu obecnego wobec stanu poprzedniego obejmuje równolegle istniejące rozwiązania i praktyki użytkowe – te nowoczesne i złożone, jak i wykorzystywane od kilkunastu lat, prostsze i mniej skomplikowane. Obok możliwości stwarzanych przez współczesne narzędzia wspomagające proces analityczny należy też uwzględnić umiejętności samych użytkowników. Cóż z tego, że dysponujemy zaawansowanymi narzędziami analizy danych – programami, pakietami i algorytmami, skoro poziom kultury informatycznej socjologów jest dalece niewystarczający. Socjolodzy, jak i przedstawiciele innych nauk społecznych, sięgający po nowoczesne oprogramowanie są przygotowani do jego wykorzystywania w stopniu bardzo ograniczonym. Umiejętności pisania skryptów, prostych programów czy posługiwania się konsolą, wierszem poleceń nie są tymi, na które kładzie się nacisk w ramach studiów kierunkowych. Socjolog to wszak nie informatyk. Jednakże faktem jest, iż w orbicie zainteresowań socjologów znalazły się narzędzia wymagające takich właśnie kompetencji.

Analityczny przegląd metod i technik prowadzenia badań z wykorzystaniem omawianych nowych technologii napotka wiele trudności. Próby ich pokonania polegają na przyjmowaniu konwencji typologicznych. Techniki uchwycone w prezentowanym tu krótkim przeglądzie procedur jakościowej (pytanie o ich ewentualną niejakościowość celowo zostaje pominięte i ujęte w nawias) analizy danych wywodzą się z dwóch nurtów. Wskazanie pierwszego (**ujęcie przedmiotowe**) polega na zdekodowaniu akronimu CAQDAS, odnoszącego się do bogatej i rozwijającej się grupy programów użytkowych wykorzystywanych na wiele sposobów w analizach danych jakościowych. W tym miejscu wiele mówi sama nazwa: oprogramowanie wspomagające proces analizy danych jakościowych. Drugi nurt (**ujęcie funkcjonalne**) zdefiniować można poprzez zidentyfikowanie typów przeprowadzanych analiz lub ich celu z uwzględnieniem rozmiaru opracowywanego materiału. Przyjęta na potrzeby dokonywanego przeglądu stosowanych metod i technik prowadzenia badań logika prezentacji odpowiada procesowi poznawania narzędzi CAQDAS i w tym sensie nawiązuje do diachronicznej formuły opisu procesu stawania się użytkownikiem tego typu rozwiązań.

Przetwarzanie danych: podejście manualne

Wśród sposobów wykorzystywania oprogramowania w badaniach jakościowych poczesne miejsce zajmuje ten polegający na funkcjonalnym wyeliminowaniu tradycyjnych nośników informacji: notatnika, papieru, odręcznego kodowania z wykorzystaniem specjalnych arkuszy. Podstawową funkcjonalnością omawianego tu oprogramowania jest klasyfikowanie informacji i nadawanie im zgodnego z preferencjami użytkownika oznaczenia; funkcję tę określa się mianem *code and retrieve*. Zastosowanie oprogramowania CAQDAS pozwala

na wprowadzenie do procesu obróbki i przetwarzania danych charakterystycznego dla techniki komputerowej porządku. Przetwarzane dane porządkowane są w osobliwą budowlę przypominającą drzewo katalogów i podkatalogów, żywcem przypominającą strukturę UNIXopodobnego systemu operacyjnego. Ten typ operacji można wykonać w każdym z dostępnych programów – od najprostszych na przykład OpenCode, RQDA, QDA Miner lite, po te o znacznie większych możliwościach: NVivo, Atlas.ti, QDA Miner i inne. Wyróżnikiem najprostszego modelu używania oprogramowania wspomagającego proces przetwarzania i analizy danych jest manualny tryb kodowania, obejmujący lekturę analizowanego materiału oraz odręczne (choć zapośredniczone interfejsem programu) kodowanie. Ten typ pracy z danymi jest możliwy w sytuacji, gdy objętość zgromadzonego materiału nie przekreśla szans na ich uważne, kilkukrotne przeczytanie. Mowa zatem o nie więcej niż kilkuset stronach tekstu.

Celem procesu kodowania jest zidentyfikowanie kluczowych elementów treści oraz powiazań między nimi. Rozbudowane pole poszukiwań staje się doskonałym obszarem zastosowań dla podstawowych operacji, które człowiekowi zajęłyby godziny, dni i tygodnie, zaś komputerowi sekundy. Użytkownik oprogramowania szybko identyfikuje udogodnienia związane z wyszukiwaniem, porównywaniem, odpytywaniem zbudowanego zbioru danych. Opisywany moment przybliża kolejny model wykorzystania oprogramowania związany z wprowadzeniem wspomagania procesu kodowania i oparcia go o częściowo automatyczne procesy.

Przetwarzanie danych: semiautomatyka

Rozwój metod automatycznego lub półautomatycznego przetwarzania materiału znajduje swe przyczyny w zwiększającej się objętości materiału poddawanego analizie lub ograniczeniach czasowych, uniemożliwiających dokładną lekturę materiału. Jeden i drugi powód w równym stopniu przyczyniają się do powstawania udogodnień pozwalających stworzyć odpowiedni zbiór danych do planowanych analiz.

Proces obróbki danych wiąże się z wykorzystaniem narzędzi wyszukiwania, budowania zapytań, kwerend w obrębie bazy danych z uwzględnieniem szczególnych warunków definiowanych przez analityka. Podejście to pozwala bez konieczności całościowej lektury zidentyfikować fragmenty tekstu spełniające warunki zapytania. Podstawowa implementacja semiautomatycznego przetwarzania tekstu pozwala na częściową redukcję materiału, który należałoby opracować manualnie. Podejście to pozwala jednak na coś więcej niż tylko na wspomnianą redukcję. Wprawne posługiwanie się językiem bazodanowych zapytań oraz operatorów logicznych (znanych każdemu humaniście) w połączeniu z podstawowymi algorytmami maszynowego uczenia się (*machine learning*) umożliwia stworzenie na podstawie definiowanych przez analityka pomiarów wejściowych (np. ręczne zakodowanie kilku słów występujących w określonej relacji) reguł, za pomocą których program przeprowadzi dalszy proces obróbki danych. Programy, które są wyposażone w tak działające moduły, to między innymi QDA Miner z modułami WordStat, SimStat, w nieco mniejszym wymiarze Atlas.ti oraz NVivo. Wśród wymienianych narzędzi można także wymienić program (choć trafniejszym określeniem jest język programowania) R z zainstalowanym pakietem tm (*textmining*) lub ekwiwalentnym.

Przetwarzanie danych: podejście *Big Data*

Jak już wspomnieliśmy we wcześniejszej części artykułu, obserwowany w ostatnim dziesięcioleciu dynamiczny rozwój cyfrowych technik generowania, przetwarzania i gromadzenia informacji spowodował pojawienie się w rzeczywistości społecznej wielkich zbiorów danych. Przetwarzanie *Big Data* eliminuje, jak już wspomniano, możliwość tradycyjnego oznaczania, indeksowania czy kodowania. Sam proces wstępnego rozpoznania wzorca, pominąwszy czasochłonność, niósłby ze sobą względu na skalę zbioru danych zbyt duże ryzyko nieuprawnionego selektywnego odczytania, bowiem rekordy bazy danych można liczyć w dziesiątkach, setkach tysięcy czy wręcz w milionach. Identyfikacji wzorców dokonuje się przy pełnej automatyzacji w oparciu o generowane na potrzeby badań lub istniejące słowniki pozwalające rozpoznawać określone sekwencje danych. Równolegle do wspomnianego podejścia rozpoznawania wzorców wykorzystuje się też analizy o charakterze statystycznym. Uwzględnienie częstości występowania słów czy indeksów oraz prawdopodobieństwa wystąpień określonych struktur językowych pozwala na wysnuwanie wniosków nie tylko odnośnie profilu syntaktycznego tekstu, ale także semantycznego. Stąd też dla analiz określanych połączonym mianem CAQDAS i *Big Data* fundamentalne znaczenie ma wywodzący się z nauk informatycznych zespół praktyk programistycznych i analitycznych określany jako przetwarzanie języka naturalnego (NLP – *natural language processing*).

Wskazanie programów zagospodarowujących poszczególne techniki *Big Data* jest nie lada wyzwaniem. Wiąże się to z faktem, iż przedsięwzięcia badawcze tego rodzaju obejmują wiele procedur. Złożoność procesu wyjawi już samo wskazanie, iż chodzi o pozyskanie danych (np. *webscraping*, *webcrawling/onlinecrawling*), ich wstępne odczytanie i przygotowanie (dzielenie na kolumny, strony, dekodowanie znaków – analiza składniowa, tzw. *parsing*), identyfikację wzorców (model nadzorowany lub nienadzorowany, z wykorzystaniem algorytmów, np. LDA, STM, lub z wykorzystaniem algorytmów klastrowania), wizualizację zagregowanych danych w formie chmur tagów, dendrogramów czy z wykorzystaniem map różnego rodzaju. Środowiskiem programistycznym, do którego sięga wielu badaczy, jest to związane z językiem programowania R. Różnorodność pakietów uzupełniających funkcjonalność programu pozwala na zbudowanie narzędzia dostosowanego do potrzeb analityka.

Celem przyświecającym zaprezentowanemu tu krótkiemu przeglądowi metod i technik wykorzystywanych w badaniach jakościowych ze wspomaganiem komputerowym nie była kompleksowa systematyka istniejących rozwiązań. Przegląd ten spełnić miał zadanie wyznaczenia punktu lub punktów zaczepienia dla badacza poszukującego sposobu na pokonanie trudności przewidywanych w planowanych działaniach badawczych. Tego typu odbiorcy należy się też pewna podpowiedź. Otóż do tej pory nie stworzono takiego programu, który wszystko wykonałby za badacza. Droga, jaką trzeba pokonać

pomiędzy pytaniem, hipotezą, falsyfikacją lub ich odpowiednikami w różnych tradycjach metodologicznych, nadal pozostaje domeną socjologa-badacza. Elementem owej drogi jest oczywiście również analizowanie zebranych danych – zagadnieniu temu poświęcimy kolejną część artykułu.

Big Data – ich źródło oraz proces analizy

Jak zauważyliśmy już w prezentowanym artykule, pojawienie się *Big Data* oznacza zmianę metod pozyskiwania, gromadzenia, zapisywania i analizowania informacji, które wpływają na sposób rozumienia i organizacji społeczeństwa. Świat życia codziennego zalewany jest ogromną liczbą różnych skwantyfikowanych i zdigitalizowanych danych społecznych, ekonomicznych, kulturowych, geograficznych, technologicznych, których liczba rośnie codziennie lawinowo. Informacja staje się wartością wtedy, gdy możemy ją przekuć w formę danych, by potem poddać szczegółowej obróbce i analizie. Koncepcję tę określa się mianem danetyzacji, czyli takiego sposobu przetwarzania informacji, który umożliwia ich późniejsze wykorzystanie i dalszą analizę danych (Mayer-Schoberger, Cukier 2014: 103–132). Doskonałym przykładem danetyzacji są rozwijane w naukach humanistycznych i społecznych korpusy tekstowe wykorzystywane do analiz lingwistycznych, a także zbiory danych tekstowych pochodzące z mediów społecznościowych (Facebook, Twitter), umożliwiające na przykład analizę dyskursu czy analizę sentymentu. Przekształcanie danych tekstowych: książek, dokumentów, zdań czy słów w dane pozwala na rozwijanie różnorodnych sposobów ich użycia, analizowania oraz tworzenia modeli teoretycznych i predykcyjnych. Bez cyfryzacji i danetyzacji nie byłoby *Big Data*, a także analiz typu CAQDAS w badaniach jakościowych. Jednakże *Big Data* to coś więcej niż proces digitalizacji danych.

Big Data jest zaliczane do działu informatyki nazywanej sztuczną inteligencją, a dokładniej tak zwanych systemów uczących się, jednakże w praktyce podejście to nie polega na uczeniu komputerów tak, by myślały jak ludzie. Sednem *Big Data* jest zdolność do przewidywania, szacowania prawdopodobieństwa wystąpienia określonych zdarzeń, zachowań lub sytuacji dzięki zastosowaniu matematyki do przetwarzania ogromnej liczby danych ustrukturyzowanych i nieustrukturyzowanych. Jednocześnie systemy analityczne *Big Data* są tak zaprojektowane, by rozwijać własne modele predykcyjne przez monitorowanie zdarzeń i procesów dostarczających nowych danych z otoczenia. Przykładowo Amazon zarekomenduje idealną książkę, Google wyszuka odpowiednią stronę, Facebook wie, kogo/co lubimy, LinkedIn odgadnie, kogo znamy. *Big Data* oznacza zdolność do korzystania z informacji zawartej w danych w nowatorski sposób, który ułatwia lepsze zrozumienie rzeczywistości oraz tworzenie dóbr i usług o znacznej wartości, a także przewidywanie, na podstawie modeli analitycznych, ścieżek przebiegu procesów społecznych i biznesowych.

Opierając się na koncepcji *Big Data*, możemy przetwarzać wszystkie dane dotyczące badanego zjawiska/procesu, jakie tylko jesteśmy w stanie o tym zjawisku/procesie zebrać. Obecnie ograniczenia w gromadzeniu i analizie danych nie są już takim problemem. Analizie podlegają wszystkie z/gromadzone (zdigitalizowane) dane bez potrzeby dobierania próby losowej i szacowania błędu ekstrapolacji wyników. Użycie wszystkich danych pozwala dostrzec szczegóły, z których wcześniej nie zdawaliśmy sobie sprawy, ponieważ byliśmy przyzwyczajeni do redukowania liczby danych. Koncepcja *Big Data* nie wymaga schematu doboru próby, zarówno losowej, jak i nielosowej. Większe spektrum danych nie wymaga także zachowania dużej dokładności[5]. Gdy możliwości pomiaru są ograniczone, skupiamy się na tym, co istotne. Wzrost skali danetyzacji sprzyja niedokładności pomiaru, ale w konsekwencji zwiększa się możliwość wnioskowania. Dokładność, precyzja „pomiaru" wymaga dobrze przygotowanych danych, co sprawdza się raczej w przypadku małej liczby danych. W koncepcji *Big Data* rezygnujemy ze sztywnej precyzji na rzecz ogólnej tendencji, poznania kierunku rozwoju jakiegoś zjawiska czy procesu. Nie oznacza to jednak rezygnacji z precyzji, ale nie jest ona priorytetem. W koncepcji *Big Data* rezygnujemy z niedoskonałości próby losowej na rzecz kompletności danych. Tracimy na dokładności, ale zyskujemy lepsze zrozumienie, wgląd w określone zjawisko. W *Big Data* istotne jest bowiem poszukiwanie i rozumienie związków między danymi, których do tej pory nie byliśmy w stanie pojąć. Zmiana skali dostępności danych spowodowała zmianę ich statusu. Zmiana ilości doprowadziła do zmiany jakości, prób zrozumienia istoty badanego zjawiska czy procesu, struktury danych, wydobycia wiedzy zawartej w danych bez potrzeby sięgania w głąb, poznawania szczegółowych cech czy istoty jakiegoś zjawiska lub procesu (choć – zgodnie z założeniami nowego paradygmatu opisywanego na początku tekstu – z uwzględnieniem informacji kontekstowych oraz teorii). *Big Data* wymaga od badacza interakcji z danymi. Kiedy pozwalamy „przemówić danym", odkrywamy powiązania, których istnienia nie podejrzewaliśmy. W erze analogowej zbieranie i analiza danych pochłaniały zazwyczaj dużo czasu, nowe pytania badawcze wymagały konieczności ponownego odtworzenia procesu zbierania danych i analizy. W erze *Big Data* wraz z digitalizacją danych i możliwością przetwarzania analogowych informacji w sposób zrozumiały dla komputerów z jednej strony nastąpił postęp w dziedzinie zarządzania danymi, ich analizy i tworzenia modeli analitycznych, zaś z drugiej, jeśli chcemy analizować duże ilości danych, musimy się pogodzić z niedokładnością i niepewnością naszego wnioskowania.

[5] Dzięki mniejszej liczbie błędów wynikających z doboru próby losowej możemy zaakceptować większą liczbę błędów pomiaru, a tym samym mniejszą dokładność pomiaru.

CAQDAS, *Data*/*TextMining* a proces analizy danych jakościowych

Przywołanemu wcześniej zjawisku danetyzacji świata życia codziennego towarzyszy rozwój nowych algorytmów, technik analitycznych oraz technologii informatycznych w zakresie przetwarzania i analizy danych, w tym lingwistyki komputerowej i sztucznej inteligencji. Danetyzacja wymaga metodologii służących rozwijaniu modeli umożliwiających kompleksową analizę zjawisk lub procesów. Kluczowe znaczenie odgrywa w tym rozwoju eksploracja danych (ang. *Data Mining*), określana także jako drążenie danych, pozyskiwanie wiedzy, wydobywanie danych, ekstrakcja wiedzy zawartej w danych. *Data Mining* to podstawowy etap procesu

odkrywania wiedzy w bazach danych (ang. KDD, *Knowledge Discovery in Databases*). Logika KDD zawiera się w sekwencji następujących etapów: zrozumienia danych, wyboru danych do analizy, wstępnego przetworzenia danych, przekształcenia danych do analizy, przeprowadzenia eksploracji w celu odkrycia struktury wzorców i zależności, konstruowania modeli analitycznych, oceny stopnia dopasowania modeli do danych, a następnie oceny i interpretacji wyników pod kątem uzyskanej wiedzy. Nie ma jednoznacznej, ogólnie przyjętej definicji eksploracji danych. Większość definicji zwraca jednak uwagę na trzy rzeczy: analizę dużych zbiorów danych (*Big Data*), poszukiwanie struktury zależności między danymi i wizualizację jako formę reprezentacji wyników. Dane w koncepcji *Big Data* nie są traktowane jako coś statycznego, jako takie, których przydatność kończy się wraz z ich zgromadzeniem, lecz jako struktury dynamiczne, dlatego też mogą być wielokrotnie wykorzystywane w inteligentny sposób z użyciem zaawansowanych algorytmów i technik analitycznych odnoszących się właśnie do metod i technik eksploracji oraz klasyfikacji. Z racji tego, że analizie poddaje się wszystkie dane, to wszelkiego rodzaju odstępstwa od normy, przypadki nietypowe stają się ważnymi informacjami w zrozumieniu istniejących w zbiorze danych zależności.

W analizach typu *Big Data* kluczową rolę odgrywa wspomniany już w pierwszej części artykułu zwrot w kierunku poszukiwania korelacji, przy jednoczesnej rezygnacji z poszukiwania przyczynowości. Dotychczasowe analizy w badaniach jakościowych ograniczały się do sprawdzania niewielkiej liczby hipotez, które formułowane były zgodnie z logiką dedukcyjną, przed zebraniem danych[6] lub indukcyjną, rodziły się w trakcie procesu analizy danych (tak jak na przykład w metodologii teorii ugruntowanej). W analizach typu *Big Data* (także w badaniach jakościowych) nie musimy się skupiać na poszukiwaniu zależności przyczynowo-skutkowych, naszym celem jest odkrywanie relacji między zdarzeniami, faktami, sądami, zachowaniami i tym podobnymi, które umożliwiają poznanie określonego zjawiska lub procesu. Skupienie się na poszukiwaniu korelacji między danymi nie prowadzi do wyjaśnienia dlaczego coś się dzieje, ale pozwala stwierdzić, co się dzieje, z czym mamy do czynienia, jaka jest skala zjawiska. Punktem wyjścia w analizie jest przede wszystkim zrozumienie danych i relacji między nimi (odkrywanie struktury relacji w procesie analizy), a dopiero w konsekwencji poszukiwanie wyjaśnienia zależności między nimi. To ostatnie jest oczywiście ważne przy wspomnianym na początku artykułu odrzuceniu tezy o „śmierci teorii". Opisany proces zrozumienia danych powinien być następnie wsparty poszukiwaniem informacji „kontekstowych" lub/i znaleźć podbudowę w teorii.

W podejściu *Big Data* wykorzystuje się między innymi techniki statystyczne (statystyki opisowe, tabele kontyngencji, analizę czynnikową, dyskryminacyjną, hierarchiczną analizę skupień, regresję logistyczną itp.), techniki uczenia maszynowego, sieci neuronowe, algorytmy indukcyjne, genetyczne czy drzewa klasyfikacyjne w celu odkrywania wiedzy zawartej w danych i tworzenia wielowymiarowych modeli predykcyjnych. Współcześnie procesy eksploracji danych znajdują na przykład zastosowanie w analizie danych o ruchu internetowym (analiza logów), rozpoznawaniu sygnałów obrazu, mowy, pisma, sensu wyrazów i zdań, struktur chemicznych, stanu zdrowia człowieka, wspomaganiu diagnostyki medycznej, biologii i badaniach genetycznych, analizie operacji bankowych, prognozowaniu wskaźników ekonomicznych, pogody, plam na Słońcu, aż po zagadnienia z zakresu kognitywistyki, doświadczeń psychologicznych, analizy sposobu rozumowania i kategoryzacji, poruszania się i planowania i tym podobne.

Jak wcześniej wspominaliśmy, *Big Data* dotyczy trzech zmian w podejściu do analizy informacji, które uzupełniają się i wzmacniają wzajemnie: możliwości analizowania dużej liczby danych z określonej dziedziny, braku konieczności ograniczania się do mniejszych zbiorów (stosowania doboru próby), gotowości do zajmowania się nieuporządkowanymi danymi płynącymi z rzeczywistego świata i nieprzywiązywania zbyt dużej wagi do ich dokładności. Dlatego w analizach typu *Big Data* eksploracja danych poprzedza eksplanację, a zrozumienie tego, co tkwi w danych, potrzebę poszukiwania relacji przyczynowych. Logika tego podejścia wydaje w pełni odpowiadać myśleniu badaczy jakościowych, z racji tego, że w badaniach jakościowych mamy najczęściej do czynienia z dużą swobodą pozyskiwania danych, a same dane jakościowe są zwykle danymi nieustrukturyzowanymi. Niestety wciąż brakuje w środowisku analityków i badaczy jakościowych w Polsce pogłębionej refleksji nad analizami *Big Data*, a także możliwościami wykorzystywania metod i technik eksploracji danych jakościowych oraz odkrywania wiedzy w obszarze badań jakościowych.

Podobnie rzecz ma się z CAQDAS. W ciągu ostatnich dwóch dekadach, wraz z rozwojem technologii informatycznych, zwiększa się świadomość badaczy jakościowych dotycząca korzystania z oprogramowania CAQDAS, szczególnie w analizie wywiadów socjologicznych (Bryda 2014a). Pomimo że rdzeń współczesnej analizy danych jakościowych stanowią wciąż procedury teorii ugruntowanej, zaimplementowane w wielu programach CAQDAS[7], to dzięki procesowi digitalizacji danych i danetyzacji samych badań jakościowych, czego przykładem jest tworzenie korpusów dokumentów tekstowych czy archiwów danych jakościowych, większe znaczenie w procesie analiz jakościowych zaczęła odgrywać analiza treści (Berelson 1952; Holsti 1969; Brent 1984; Weber 1990; Krippendorf 2004), wzbogacona o najnowsze osiągnięcia w dziedzinie lingwistyki komputerowej. Jeśli prześledzimy pojawianie się nowych funkcjonalności w programach CAQDAS na przestrzeni ostatnich kilkudziesięciu lat, to zobaczymy, że rozwój wspomaganej komputerowo analizy danych jakościowych w kierunku *DataMining* czy *TextMining* (Wiedemann 2013; Bryda 2014b) nie byłby możliwy bez rozwoju technik ilościowej i jakościowej analizy treści, metod mieszanych (Tashakkori, Teddlie 2003), a także metodologii eksploracji danych tekstowych i odkrywania wiedzy (Hand, Mannila, Smyth 2005; Larose 2006; 2008). W badaniach jakościowych procesowi temu towarzyszy wyraźny zwrot metodologiczny w kierunku

[6] „Hipotezy" powstają w drodze eksploracji danych, jako efekty identyfikacji systematycznych relacji pomiędzy zmiennymi w sytuacji, gdy natura tych relacji nie jest z góry określona. Stąd drążenie danych utożsamia się zazwyczaj z podejściem indukcyjnym do odkrywania wiedzy.

[7] Teoria ugruntowana wytyczyła nie tylko wzorce i procedury przeprowadzania analiz jakościowych, ale jej założenia metodologiczne stały u podstaw rozwoju wielu obecnych funkcjonalności programów CAQDAS.

paradygmatu *mixed-methods*. Jego wyrazem jest przechodzenie od klasycznej analizy danych jakościowych (*Qualitative Analysis*), przez *Qualitative Content Analysis*, w kierunku pogłębionej eksploracji danych jakościowych (Bryda 2014b) i *TextMining* wykorzystującej techniki statystyczne i algorytmy z dziedziny inteligencji komputerowej czy przetwarzania języka naturalnego (Bryda, Tomanek 2014). *TextMining* ma korzenie w rozwijającej się od kilkunastu lat metodologii *Data Mining*, ale obecnie staje się podstawą wielu analiz jakościowych i rozwoju funkcjonalności we wspomaganej komputerowo analizie danych jakościowych (Ho Yu, Jannasch-Pennell, DiGangi 2011). Rozwój CAQDAS w kierunku wykorzystania zaawansowanych metod eksploracji i odkrywania wiedzy w danych (głównie tekstowych) jest możliwy nie tylko dzięki zastosowaniu nowych technologii informatycznych, ale przede wszystkim dzięki ewolucji świadomości analitycznej badaczy jakościowych i metodologii prowadzenia analizy danych jakościowych w kierunku *Big Data*, gdzie główną rolę odgrywa poszukiwanie korelacji i prawdopodobieństwo.

Big Data, CAQDAS w praktyce badawczej

Wpływ *Big Data*, CAQDAS i nowych technologii na proces badań jakościowych, sposób zbierania i analizy danych staje się coraz bardziej widoczny. Również w Polsce mamy do czynienia z rosnącym zainteresowaniem świata akademickiego, jak też podmiotów rynkowych problematyką *Big Data* oraz możliwościami wykorzystywania oprogramowania CAQDAS w projektowaniu i prowadzeniu badań, a także analizie danych jakościowych. Jak dotąd za pomocą programów CAQDAS analizuje się głównie dane tekstowe, takie jak transkrypcje wywiadów, teksty prasowe czy notatki z obserwacji. *Big Data* wnosi jednak nowe rodzaje nieustrukturyzowanych danych dotyczących interakcji (Facebook, Twitter), a także inny niż dotychczas sposób myślenia o samych danych i sposobie ich analizowania.

Nawiązując do wcześniejszych rozważań, jako redaktorzy tego tomu, chcielibyśmy przybliżyć zagadnienie *Big Data* i CAQDAS w praktyce. Oddajemy do rąk czytelników tom „Przeglądu Socjologii Jakościowej" w całości poświęcony tej problematyce. Publikacja zawiera teksty przygotowane przez badaczy i praktyków, których kompetencje w zakresie *Big Data* i pracy z programami CAQDAS oparte są na połączeniu rzetelnej wiedzy i doświadczenia.

Publikację rozpoczyna tekst Mariusza Dzięglewskiego dotyczący korzyści i ograniczeń w wykorzystywaniu oprogramowania CAQDAS w badaniach digitalizacji i odbioru dziedzictwa kulturowego. Autor poddaje refleksji problem łączenia i przenikania się tradycyjnych metod badania i procedur analizy ze wspomaganą komputerowo analizą danych jakościowych. Opisuje on rolę, jaką w projektowaniu badań odgrywa CAQDAS, sposób, w jaki wpływa to oprogramowanie na percepcję problemu badawczego oraz interpretację wyników badania, a także problematykę przenikania się różnych podejść metodologicznych i analitycznych na różnych etapach projektu badawczego: budowania bazy, kodowania danych, analizy, wizualizacji i interpretacji wyników. Analizując możliwości oraz ograniczenia wynikające ze stosowania CAQDAS dla analizy treści dokumentów prawnych i transkrypcji wywiadów pogłębionych, autor poszukuje optymalnego połączenia tradycyjnych i nowoczesnych metod badania oraz analizy danych, które pozwoliłoby na uniknięcie „pułapek" związanych z wykorzystaniem nowych technologii w badaniach społecznych.

Kolejny artykuł – Jakuba Niedbalskiego – ma charakter poglądowy i edukacyjny. Autor stawia sobie za cel zapoznanie czytelników z możliwościami NVivo, narzędzia należącego do rodziny CAQDAS oraz jego faktycznym zastosowaniem w projektach realizowanych zgodnie z założeniami metodologii teorii ugruntowanej. Autor pokazuje, w jaki sposób można wykorzystać narzędzia komputerowego wspomagania analizy danych jakościowych w praktyce badawczej. Na przykładzie konkretnego projektu badawczego przybliża etapy pracy w programie NVivo zgodnie z procedurami metodologii teorii ugruntowanej, wskazując na istniejące udogodnienia i potencjalne trudności związane ze stosowaniem oprogramowania komputerowego jako elementu warsztatu badacza jakościowego.

W artykule dotyczącym „mowy nienawiści" i wykorzystania algorytmów uczenia maszynowego w analizie danych jakościowych Marek Troszyński zajmuje się procesem automatyzacji kodowania (anotacji i tagowania) danych tekstowych pochodzących z forów internetowych w oparciu o znaczenia zawarte w tekście. Wdrożenie tego procesu pozwala na ilościowe analizy korpusów danych tekstowych liczących setki tysięcy tekstów. Autor skupia uwagę na procesie konceptualizacji i operacjonalizacji „mowy nienawiści", przygotowaniu dokładnej instrukcji kodowej oraz treningu zespołu kodującego w celu uzyskania wysokiego współczynnika zgodności między koderami. Następnie przedstawia zastosowane metody kodowania automatycznego, wskazując czynniki, które są kluczowe dla procesu badawczego wykorzystującego uczenie maszynowe.

Problematyka radykalizacji i brutalizacji języka, nadużywania słów nacechowanych negatywnie w dyskursie politycznym i o polityce, a także degradacji znaczenia tych słów stanowi przedmiot zainteresowania Agnieszki Kwiatkowskiej, która przedstawia możliwości zastosowania modeli generatywnych do analizy debat parlamentarnych. W artykule analizuje ona zbiór przemówień sejmowych z lat 1991–2016 odnoszących się do idei hańby, zdrady, niesławy i skandalu. W tym celu wykorzystuje nienadzorowane algorytmy przeszukiwania korpusów tekstów oraz analizy ukrytych tematów, w tym generatywny model tematyczny, metodę ukrytej alokacji Dirichleta i jej rozszerzenie – strukturalny model tematyczny jako metodę ekstrakcji tematów w dużych korpusach danych tekstowych.

W kolejnym artykule: *Dobra zmiana czy Polska w ruinie?* Alicja Zawistowska i Małgorzata Skowrońska przeprowadziły analizę ewolucji znaczeniowej wpisów opatrzonych hashtagami #dobrazmiana i #polskawruinie zamieszczonych w serwisie społecznościowym Twitter. Autorki poddają analizie wpisy, które pojawiły się w latach 2015–2016. Celem tej analizy jest ukazanie dynamiki zabarwienia emocjonalnego obu haseł, a także ukazanie wpływu dominującego na Twitterze stylu komunikacji na wspomnianą ewolucję znaczenia tych wpisów. W artykule przedstawiono również podstawowe problemy metodologiczne związane z zastosowaniem analizy treści w mediach społecznościowych. Publikację kończy artykuł Krzysztofa Tomanka dotyczący metodyki analizy treści w projektach stosujących techniki *TextMining* i oprogramowanie CAQDAS. Autor wskazuje przykładowe dylematy metodologiczne występujące w trakcie pracy z dużymi wolumenami

danych tekstowych pochodzących z różnych źródeł i zapisanych w różnorodnych formatach, zwracając uwagę w szczególności na problem jakości danych nieustrukturyzowanych typu *quan* i *qual*. Na przykładzie własnego projektu przedstawia zastosowanie metody analizy danych wykorzystującej różnorodne narzędzia CAQDAS do (pół-)automatycznej klasyfikacji wypowiedzi pisanych wtedy, gdy mamy do czynienia z danymi o różnorodnej jakości. Próbuje również pokazać, kiedy klasyfikacja (pół-)automatyczna jest przydatna, a kiedy nie ma szans powodzenia oraz momenty, w których badacz jakościowy wykorzystuje wiedzę z innych dziedzin: przetwarzanie języka naturalnego czy uczenie maszynowe w procesie analizy danych.

Wśród zagadnień, którym autorzy poświęcili swoje teksty, na szczególną uwagę zasługują: podjęcie dyskusji nad zgodnością zasad, na jakich funkcjonuje oprogramowanie CAQDAS, z regułami oraz procedurami metodologii badań jakościowych; możliwości zastosowania narzędzi CAQDAS w realizacji projektów badawczych opartych na różnych metodach jakościowych i w ramach różnych podejść analitycznych; zgodności „architektury oprogramowania" z procedurami wybranych metod i technik badawczych; wpływu *Big Data* i nowych technologii na proces badawczy, implementacji nowych algorytmów i technik; wpływu rozwiązań wykorzystywanych w innych dziedzinach nauki na proces analizy i badań opartych na metodach jakościowych.

Bibliografia

Anderson Chris (2008) *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. „Wired", 16 lipca [dostęp 26 kwietnia 2017 r.]. Dostępny w Internecie: ‹http://www.uvm.edu/~cmplxsys/wordpress/wp-content/uploads/reading-group/pdfs/2008/anderson2008.pdf›.

Berelson Bernard (1952) *Content Analysis in Communication Research*. Glencoe, IL: Free Press.

Brent Edward E. (1984) *Qualitative Computing: Approaches and Issues*. „Qualitative Sociology", vol. 7 (1/2), s. 36–60.

Bryda Grzegorz (2014a) *CAQDAS a badania jakościowe w praktyce*. „Przegląd Socjologii Jakościowej", t. 10, nr 2, s. 12–38. Dostępny w Internecie: ‹www.przegladsocjologiijakosciowej.org›.

Bryda Grzegorz (2014b) *Caqdas, Data Mining i odkrywanie wiedzy w danych jakościowych* [w:] Jakub Niedbalski, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo UŁ, s. 13–40.

Bryda Grzegorz, Tomanek Krzysztof (2014) *Od CAQDAS do TextMiningu. Nowe techniki w analizie danych jakościowych* [w:] Jakub Niedbalski, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo UŁ, s. 191–218.

Chang Ray M., Kauffman Robert J., Kwon Young Ok (2013) *Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data*. „Decision Support Systems", vol. 63, s. 67–80.

Chen C. L. Philip, Zhang Chun-Yang (2014) *Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data*. „Information Sciences", vol. 275, s. 314–347.

Hand David, Mannila Heikki, Smyth Padhraic (2005) *Eksploracja danych*. Przełożyła Agnieszka Chądzyńska. Warszawa: WNT.

Hey Tony, Tansley Steward, Tolle Kristin (2009) *Jim Gray on eScience: A Transformed Scientific Method* [w:] Hey Tony, Tansley Steward, Tolle Kristin, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, s. xvii–xxxi.

Ho Yu Chong, Jannasch-Pennell Angel, DiGangi Samuel (2011) *Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability*. „The Qualitative Report", vol. 16, no. 3, s. 730–744.

Holsti Ole R. (1969) *Content Analysis for the Social Sciences and the Humanities*. Reading, MA: Addison-Wesley.

Kitchin Rob (2014) *Big Data, New Epistemologies and Paradigm Shifts*. „Big Data & Society", April-June, s. 1–12.

Krippendorf Klaus (2004) *Content Analysis. An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Larose Daniel T. (2006) *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji*. Przełożyła Anna Wilbik. Warszawa: PWN.

Larose Daniel T. (2008) *Metody i modele eksploracji danych*. Przełożyła Anna Wilbik. Warszawa: PWN.

Masterman Margaret (1970) *The Nature of a Paradigm* [w:] Imre Lakatos, Alan. E. Musgrave, eds., *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press, s. 59–90.

Mayer-Schonberger Viktor, Cukier Kenneth (2014) *Big Data*. Przełożył Michał Gładki. Warszawa: MT Biznes.

Parks Malcolm R. (2014) *Big Data in Communication Research: Its Contents and Discontents*. „Journal of Communication", vol. 64, s. 355–360.

Pigliucci Massimo (2009) *The End of Theory in Science?* „EMBO reports", vol. 10, no. 6, s. 533–662.

Ramsay Stephen (2010) *Reading Machines: Toward an Algorithmic Criticism*. Champaign: University of Illinois Press.

Shah Dhavan V., Cappella Joseph N., Neuman W. Russell (2015) *Big Data, Digital Media, and Computational Social Science: Possibilities and Perils*. „The ANNALS of the American Academy of Political and Social Science", vol. 659, no 1, s. 6–13.

Tashakkori Abbas, Teddlie Charles (2003) *Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks, CA: SAGE Publications.

Weber Robert P. (1990) *Basic Content Analysis*. Newbury Park, CA: Sage.

Wiedemann Gregor (2013) *Opening Up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences*. „Forum Qualitative Sozialforschung", vol. 14, no. 2 [dostęp 26 kwietnia 2017 r.]. Dostępny w Internecie: ‹http://www.qualitative-research.net/index.php/fqs/article/view/1949›.

Cytowanie

Brosz Maciej, Bryda Grzegorz, Siuda Piotr (2017) *Od redaktorów: Big Data* i CAQDAS a procedury badawcze w polu socjologii jakościowej. „Przegląd Socjologii Jakościowej", t. 13, nr 2, s. 6–23 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: ‹www.przegladsocjologiijakosciowej.org›.

Big Data, CAQDAS and research procedure in the field of qualitative research

Abstract: The reality of everyday life is covered by huge amounts of various quantified and digitized data. The quantity of data grows everyday enormously. These data can be processed and treated as research material, also qualitative. The application of Big Data in qualitative research modifies the procedure on every step of research process: from research design up to conclusion. Does implementing Big Data strategy into qualitative research lead to atheoretical approach? What are the consequences of using the complete data sets instead of random sample technique? The purpose of this article is to indicate this changes and their brief characteristics considering the significant role of different kind of software (especially CAQDAS), and so the analysis that can be conducted.

Keywords: big data, CAQDAS, computer-aided qualitative data analysis, data processing, datafication, qualitative data