http://dx.doi.org/10.18778/2543-9421.04.10
ISSN 2543-9421

Konwersatorium Wiedzy o Mieście 4 (32), 2019

Iwona Jażdżewska

https://orcid.org/0000-0002-4554-7486
Uniwersytet Łódzki
Wydział Nauk Geograficznych Uniwersytet Łódzki
Instytut Geografii Miast i Turyzmu
iwona.jazdzewska@uni.lodz.pl

DANE DOTYCZĄCE MIAST JAKO PRZEDMIOT BADAŃ GEOGRAFICZNYCH

Śmieci na wejściu – śmieci na wyjściu[1]

Abstrakt: W pracy przedstawiono problemy, z jakimi można się spotkać podczas analizy danych o mieście, od ich pozyskiwania, poprzez czyszczenie, aż po zapis w odpowiednim formacie. Problem ten jest na tyle istotny, że powinien być jednym z podmiotów badań geograficznych. Zwrócono uwagę na potrzebę dyskusji o danych geograficznych i jej prezentacji na łamach czasopism naukowych. Zasygnalizowano możliwości korzystania z nich i udostępniania w repozytoriach otwartych danych w ramach „otwartej nauki".
Słowa kluczowe: geografia miast, eksploracja danych, otwarte dane, ISO, GIScience.

DATA ON CITIES AS THE SUBJECT OF GEOGRAPHICAL RESEARCH

Abstract: Contemporary geographical data on cities come from various sources, and the increase in their number is an avalanche. As the perception of data is changing, so is the way a geographer thinks and works. The abundance of data on cities obtained from various sensors and from the society makes the research problem recognizable in the context of existing data, which makes it necessary to examine it. This paper presents the problems that can be encountered when analysing data for a geographical science research project, from its acquisition, through cleaning, to recording in the appropriate format.
Keywords: urban geography, data mining, open data, ISO standards, GIScience.

1. WPROWADZENIE

Jest oczywiste stwierdzenie, że w geografii miast operuje się różnymi zbiorami danych, które są pozyskane, a następnie odpowiednio opracowane, co pozwala na wyciągnięcie wniosków, testowanie hipotez czy sformułowanie nowych teorii. W przeciwieństwie do danych wykorzystywanych w innych dyscyplinach nauki dane geograficzne powinny mieć dodatkowy atrybut, czyli informację o lokalizacji. Pozyskiwanie danych przez geografów miast może zająć bardzo dużą część pracy badawczej. Ten model pracy w naukach geograficznych ulega zmianie ze względu na coraz większą dostępność gotowych danych geograficznych – w tym *big data* – pochodzących z różnych urządzeń wyposażanych w czujniki, sensory, lasery itd., oraz uzyskiwanych bezpośrednio lub pośrednio od ludzi, np. z ich telefonów komórkowych, mediów społecznościowych, spisów statystycznych. Czasami potrzebne informacje, ale także ich przetwarzanie jest całkowicie zależne od technologii obliczeniowych (Ash, Kitchin, Leszczynski, 2018).

Współczesne dane geograficzne pochodzą z różnych źródeł, często niedostępnych poprzednikom, jak np. dane satelitarne, z sensorów różnych urządzeń pomiarowych, cyfrowe dane geodezyjne i wiele innych (Adamczyk, Będkowski, 2018). Rozwój technologii informatycznych i pomiarowych spowodował, że jest ich bardzo dużo lub mają swoją specyfikę, która wymaga dobrej znajomości nie tylko zagadnień geograficznych, ale również informatycznych i baz danych. Lawinowy wzrost danych nie zawsze idzie w parze z ich dokładnością, możliwością porównania ich w czasie czy połączenia w jedną bazę przestrzenną. Na ich podstawie stawiane są hipotezy badawcze i formułowane wnioski, dlatego konieczna jest ich dobra znajomość, a także umiejętność przetwarzania z użyciem oprogramowania GIS lub innego. Zmienia się

sposób postrzegania danych, a w konsekwencji zmienia się sposób myślenia i styl pracy geografa. Miller i Goodchild (2015) piszą o geografii opartej na danych (*data-driven geography*), która może pojawiać się w odpowiedzi na bogactwo danych georeferencyjnych pozyskiwanych z rozmaitych czujników czy od społeczeństwa (Miller, Goodchild, 2015).

W czasie gdy pobieraniem i archiwizowaniem bardzo dużej liczby danych zajmują się różne jednostki – nie zawsze naukowe – następuje zwrot w modelu pracy geografów, którzy coraz częściej korzystają z danych, a nie są ich producentami. Problem badawczy jest rozpoznawany i rozwiązywany w kontekście istniejących już danych, czyli oparty na indukcyjnej metodzie rozwiązywania problemów badawczych. Polega to na tym, że na podstawie dostępnych informacji wyposażanych w georeferencje poszukuje się wiedzy i nowych idei naukowych w geografii. Jedną z konsekwencji takiego postrzegania danych jest wyodrębnianie się nowej dyscypliny – GIScience, w której podstawową rolę odgrywają cyfrowe dane geograficzne pochodzące z różnych źródeł. Wśród wielu zagadnień, jakimi zajmuje się GIScience znajdują się metody ściśle związane z badaniem danych, czyli ich pozyskiwaniem, eksploracją i przetwarzaniem. Należą do nich m.in.: zbieranie danych i pomiary, modelowanie danych, interoperacyjność informacji geograficznych czy problemy niepewności w geografii związane z danymi (Blaschke, Merschdorf, 2014).

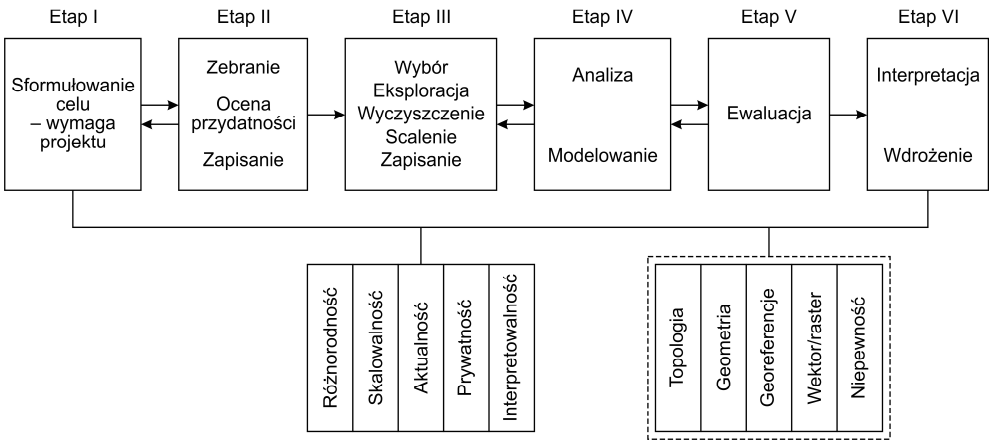
W nauce pojawił się zatem kolejny już paradygmat oparty na przetwarzaniu danych. Rozwój informatyzacji i postęp w pozyskiwaniu danych do badań naukowych jest porównywany do ważnego etapu rozwoju cywilizacyjnego, jakim było wynalezienie prasy drukarskiej (Bell, 2009).

Skoro dane są tak ważne, a od ich specyfiki zależą wyniki prac badawczych z zakresu geografii miast i nie tylko, to powinny być one również przedmiotem badań oraz niezbędnym etapem badania zjawisk i procesów przestrzennych. Naukowcy zajmujący się studiowaniem danych mogą mieć wykształcenie w dowolnej dziedzinie, ale badaniem danych przestrzennych dotyczących miast mogą zajmować się badacze związani z geografią społeczno-ekonomiczną, geografią fizyczną, gospodarką przestrzenną, urbanistyką, architekturą, geodezją i innymi pokrewnymi dziedzinami, w których dominują analizy zjawisk w przestrzeni geograficznej miasta.

2. ANALIZA DANYCH GEOGRAFICZNYCH

Dane geograficzne – w tym dane o miastach – w przeciwieństwie do innych danych wykorzystywanych w badaniach empirycznych mają postać graficzną i atrybutową (tabelaryczną) lub opisową. Dane graficzne są dwu- lub trzywymiarowe, co odróżnia je od danych atrybutowych. Systemy Informacji Geograficznej (GIS z ang.) pozwalają na równoczesne wykonywanie analiz z ich udziałem. Ich jakość była i jest przedmiotem dociekań wielu badaczy zajmujących się GIS, gdyż determinuje ona w dużym stopniu wyniki badań przestrzennych (Guo, Goodchild, Annoni, 2020; Shi, Fisher, Goodchild, 2002).

Z doświadczenia Dasu i Johnsona wynika, że „eksploracja i czyszczenie danych stanowią 80% wysiłku, który determinuje 80% wartości analizy danych" (Dasu, Johnson, 2003, s. ix), w Systemach Informacji Geograficznej wydatki na gromadzenie



Rys. 1. Etapy procesu „wydobywania" wiedzy z danych oraz związane z nim problemy
Źródło: Szeliga (2017, s. 3) zmodyfikowane o dane geograficzne (graficzne)

danych mogą stanowić 60–85% kosztów całkowitych projektu (Longley, Goodchild, Maguire, Rhind, 2006). Wysiłek ten wymaga wiedzy oraz umiejętności stawiania hipotez badawczych obejmujących dane (Sadiq i in., 2018).

Analiza danych dotyczących miast może być częścią procesu „wydobywania" wiedzy z danych (rys. 1) według metodyki CRISP-DM (Szeliga, 2017), który powinien zostać poddany modyfikacji ze względu na specyfikę danych geograficznych. W pracy omówione będą dwa pierwsze etapy tego procesu.

2.1. ETAP PIERWSZY

Etap pierwszy analizy danych obejmuje poszukiwanie i pozyskiwanie danych geograficznych dotyczących miast z różnych źródeł jako dane pierwotne lub wtórne, i sprawdzenie pod kątem ich przydatności do realizacji zamierzonych badań. Ważne jest kto lub jaka organizacja zbierała dane i je udostępnia. Czy nie były one wcześniej przetworzone lub przefiltrowane. Może bowiem okazać się, że znalezione dane nie zawierają niezbędnych informacji potrzebnych do rozpoczęcia procesu badawczego. Może to wynikać z błędnie pozyskanych zbiorów, np. dla innego okresu, zbyt małej liczby atrybutów, źle dobranej skali przestrzennej (zbyt dokładnej lub zgeneralizowanej). Ważny jest standard danych i ich format, najlepiej aby był znany i akceptowany, choć może pojawić się nowe interesujące źródło niespełniające dotychczasowych wymagań, ale zawierające cenne informacje, które warto sprawdzić pod kątem przydatności do badań.

Na pierwszym etapie zbieramy różnorodne dane, niezależnie od tego, czy w tym momencie potrafimy je zintegrować. Jeśli zbiór danych jest bardzo duży, to pobieramy małą próbę danych i sprawdzamy ich metadane, strukturę (np. liczbę zmiennych i skalę pomiarową, skalę przestrzenną, rozdzielczość, geometrię), po czym decydujemy o pobraniu całości lub potrzebnego do badań fragmentu zbioru. Musimy pamiętać, że niektóre dane, np. rastrowe, wymagają dużo miejsca w pamięci operacyjnej lub działania „w chmurze". Dane społeczne, do których od niedawna mają dostęp geografowie (Shelton, 2017), należą do danych „wrażliwych". Wymagają specjalnego traktowania, aby nie naruszyć prywatności osób, a także aby publikowane wyniki nie pozostały w sprzeczności z prawami człowieka oraz nie spowodowały kategoryzacji ludzi i miejsc, np. wynikające z korelacji, a nie z faktycznego ich zachowania (Miller, Goodchild, 2015). Szczególnie w geografii społecznej miast jest to niezwykle ważne, gdyż wskutek nieprawidłowej interpretacji danych i wyników badań można pewne fragmenty miasta lub konkretne lokalizacje naznaczyć nieodpowiednio.

Po wstępnym zaakceptowaniu danych należy zdecydować o sposobie ich zapisu i przejrzeć pod kątem ewentualnych błędów, wstrzymując się od ich poprawy do czasu oceny ich wiarygodności i jakości. Nie wszystkie zebrane informacje są jednakowo wartościowe. Pozyskane ze stron rządowych zawierają zazwyczaj dokładny opis procesu ich pozyskiwania i specyfikację techniczną. Również instytucje czy organizacje, takie jak np. Open StreetMap (OSM)[2], podają takie informacje. Należy jednak pamiętać, że są one wpisywane przez wolontariuszy i mogą być niekompletne albo obarczone błędem natury ludzkiej. W przypadku OSM powinno się podkreślić, że ta niekompletność jest zróżnicowana przestrzennie, co dodatkowo utrudnia ich wykorzystanie. Nieznane organizacje zbierające i udostępniające dane mogą mieć zarówno bardzo dobre jakościowo dane, jak i „śmieci", których czyszczenie nie jest warte poświęcania czasu. Ocena jakości danych powinna być rzetelnie przeprowadzona i przedstawiona, gdyż problemy, z jakimi spotyka się badacz w pierwszej fazie pozyskiwania danych, mogą mieć wpływ na cały proces wnioskowania.

2.2. ETAP DRUGI

Na drugim etapie analizy danych – który można nazwać wstępną obróbką danych – oceniamy zgromadzone dane pod kątem ich aktualności, liczebności i przydatności. Przeszukujemy bazy danych i określamy, które włączymy do badań. Trudny moment pojawia się, kiedy trzeba pozbyć się mało wartościowych lub nieprzydatnych zbiorów dnych czy ich części, a zostawić te, które według wiedzy badacza mogą okazać się przydatne. Zbyt duża liczba danych wymaga czasu na ich opracowanie i może nadmiernie wydłużyć proces badań, dlatego powinno się wybierać tylko niezbędne do rozwiązania problemu badawczego, a pozostałe usuwać. Czasami tylko część zbioru danych może okazać się konieczna do realizacji projektu badawczego. Na przykład pozyskane dane z Ewidencji Gruntów i Budynków (EGiB)[3] zawierają wiele informacji, z których tylko część będzie wykorzystywana. Podczas eksploracji danych może okazać się również, że

w pozyskanej bazie brakuje informacji np. o fragmencie miasta. Z taką sytuacją można spotkać się, gdy źródłem danych są portale lub organizacje społecznościowe, takie jak m.in. OpenStreetMap. W tym przypadku należy podjąć decyzję o możliwości uzupełnienia danych źródłowych lub, czy mimo braków, włączyć je do zbioru danych. Problem z nadmiarem lub niedostatkiem informacji może dotyczyć zarówno danych atrybutowych, jak i graficznych. Identyfikacja źródeł błędów powinna według Bieleckiej (2006) zawierać kolejne etapy, takie jak: szacowanie wielkości błędów, modelowanie rozprzestrzeniania się błędów, strategia zarządzania błędami, strategia redukcji błędów.

Po wstępnym wybraniu z dostępnych baz danych tych, które będą wykorzystywane w procesie badawczym, kolejnym etapem eksploracji danych jest ich czyszczenie. Jest to niezwykle ważny etap (choć polega na mało ciekawym, a czasami żmudnym zajęciu) – analizy danych wykorzystujących duże zbiory do procesu badawczego. Problem z jakością danych może przekładać się na uzyskanie sfałszowanych wyników, a w konsekwencji powodować, że badacze widzieli w nich coś, czego w rzeczywistości nie było (Ryza, Laserson, Owen, Wills, 2016). Wyniki uzyskane na podstawie źle oczyszczonych danych mogą nie mieć sensu i wymuszać zakończenie badań lub, co gorsze, prowadzić do niewiarygodnych wniosków i podejmowanych na ich podstawie decyzji. W trakcie przeglądania danych można spotkać się z różnymi problemami wymagającymi interwencji, np. zbędnymi polami, rekordami z brakującymi wartościami, danymi odstającymi, danymi w złym formacie czy po prostu z danymi niezgodnymi z zasadami zdrowego rozsądku (Larose, 2013). Dane geograficzne reprezentujące świat rzeczywisty są często przedstawiane w postaci modeli, bowiem nie jest możliwe precyzyjne ich opisanie, wynikające z niepewności (Longley, Goodchild, Maguire, Rhind, 2006). Dlatego związane są z nimi problemy, z którymi badacze borykają się w trakcie przygotowania danych do pracy (tab. 1). Zostaną one mówione pokrótce, gdyż jest to zagadnienie bardzo szerokie, a prezentowane w literaturze związanej z GIScience i Data Minning (Cai, Xie, 2007; Shi, Fisher, Goodchild, 2002).

Część procesu eksploracji i czyszczenia danych dotyczy zarówno danych atrybutowych, jak i graficznych (kolumna 1 tab. 1).

Dane odstające mają duży wpływ na wnioskowanie, w szczególności wówczas, gdy zamierzone procedury stosują metody czułe na ich występowanie, np. średnia arytmetyczna i statystyki ją wykorzystujące. Dotyczą one zarówno danych atrybutowych, jak i geometrycznych (Longley, Goodchild, Maguire, Rhind, 2006). Występowanie wartości odstających może wynikać ze specyfiki zjawiska lub może być spowodowane błędnym zapisem w zbiorze danych, np. przy zapisie wieku respondenta widnieje 250 zamiast 25 lat. Można je wskazać za pomocą kilku metod, m.in. histogramu, wykresu rozrzutu, a także metodą Tukeya (Foreman, 2019) lub poprzez wizualizację na mapie (Jażdżewska, 2018).

Tab. 1. Problemy z przygotowaniem zbiorów danych do analizy w GIS

Problemy z danymi	Typ danych	
	atrybutowe	graficzne
Dane odstające	T	T
Dane błędnie zapisane	T	T
Dane niekompletne	T	T
Dane niedokładne	T	T
Dane nieaktualne	T	T
Duplikaty danych	T	T
Niespójność formatu danych	T	T
Źle dobrana próba	T	T
Brak metadanych	T	T
Brak atrybutu georeferencji	T	T
Skala	T	T
Wybór jednostki przestrzennej	T	T
Niepewność i nieokreśloność danych	T	T
Niejednoznaczność nazw geograficznych	T	T
Model wektorowy *versus* rastrowy	N	T
Błędy w kalibracji zobrazowań lotniczych	N	T
Błędna generalizacja kształtu	N	T
Błędna digitalizacja (wektoryzacja)	N	T
Błędy w topologii danych	N	T
Błąd określenia położenia geograficznego	N	T

Źródło: opracowanie własne (T – tak, N – nie).

Problem **błędnie zapisanych danych** dotyczy zarówno danych numerycznych, tekstowych, jak i geometrycznych. Niedoświadczony operator może wprowadzić niewłaściwe dane, z kolei dane pozyskiwane z portali społecznościowych mogą mieć różny zapis w zależności od języka, precyzji formułowania zapisu czy po prostu błędów ortograficznych. Problemy pojawiają się w błędnym zapisie nazw geograficznych, takich jak: nazwy ulic, miast, państw czy regionów. Ich wychwycenie i poprawienie stanowi nieraz duże wyzwanie i jest czasochłonne. Z kolei dane wektorowe mogą być niestarannie lub błędnie wprowadzone (np. brak węzłów, niedociągnięcia, duplikaty).

O niekompletności danych mówimy wówczas, gdy atrybuty mają pojedyncze brakujące pola,

a w danych graficznych brakuje elementów (np. kilku budynków lub ulic w mieście). Można spróbować uzupełnić dane atrybutowe poprzez: wpisanie wartości z innego wiarygodnego źródła lub – jeśli nie jest możliwe uzyskanie prawidłowej wartości – oszacowanie jej (m.in. predykcja). Jeśli brak wartości zakłóca wyniki, to obiekty – rekordy, w których nie ma wartości, należy wyłączyć z analizy lub wstawić wartości zastępcze, np. wartość stałą ustaloną przez analityka, wartość średnią lub modalną, wartość wygenerowaną losowo z obserwowanego rozkładu zmiennych (Larose, 2013). W przypadku danych graficznych należałoby je uzupełnić z innych źródeł.

Dane niedokładne, czyli mało precyzyjne, mogą wystąpić z różnych powodów, choćby na skutek zamiany skali danych z ilorazowej na porządkową (wartości są zastępowane przedziałami lub opisami) (Osowski, 2013) lub usterek wykorzystywanego oprzyrządowania czy nieprecyzyjnego zapisu danych, np. mniejszej liczby miejsc dziesiętnych (Hand, 2005). W naukach geograficznych można spotkać się z tym problemem zarówno dla danych atrybutowych, jak i geometrycznych, co jest związane m.in. z charakterem danych ciągłych i dyskretnych. Błędy pomiaru mogą wynikać z braku uwzględnienia odpowiedniego odwzorowania lub nieporawnie skalibrowanego odbiornika GPS (Bielecka, 2006). W przypadku danych przestrzennych niedokładność może wynikać z próby wykorzystania danych przygotowanych do (mniejszej) skali przstrzennej. Niekiedy poprawa precyzji wyników wymagałaby powtórzenia eksperymentu, co np. przy szybko zmieniającym się środowisku geograficznym może przynieść zarówno dobre efekty, jak i nowy zestaw danych nieporównywalny z wyjściowym.

Dane nieaktualne lub brak aktualnych danych może wystąpić z kilku powodów: błędu zapisu, nieaktualizowania danych na bieżąco, błędów w metadanych czy niemożności ich pozyskania, np. liczby ludności w mieście w trakcie konfliktu zbrojnego. Problem aktualizacji danych obejmuje różne aspekty: od danych geodezyjnych, poprzez dane ekonomiczne, do społecznych. Nie wystarczy sporządzić bazę danych geograficznych na dany moment, bowiem środowisko geograficzne jest tak zmienne, że musi być ona na bieżąco aktualizowana oraz w metadanych podawana informacja o dacie aktualizacji. Zachowanie historii zmian danych jest również jednym z ważnych aspektów baz, aby móc je wykorzystać do badań porównawczych. Dlatego powinno się budować systemy wersjonowania baz, w celu ich fizycznego odtworzenia lub odczytania stanu historycznego (Bach, Stańczak, Werner, 2009).

Mając do dyspozycji dane z różnych źródeł możemy znaleźć w pozyskanych zbiorach **duplikaty danych**, np. siatki ulic, granice jednostek czy pojedyncze atrybuty. Takie duplikaty mogą występować zarówno w jednym zbiorze, jak i w kilku. W przypadku spotkania duplikatów w różnych zbiorach danych należy też pamiętać, że mogą one różnić się, nawet znacząco, zasięgiem, przebiegiem linii czy miejscem wstawienia punktu. Warto wówczas sprawdzić ich wiarygodność, kompletność, dostępność oraz jakość metadanych i wybrać najlepsze zbiory lub połączyć je usuwając duplikaty.

Niespójność formatu danych jest poważną bolączką naukowców pracujących z wieloma bazami. Wynika ona z różnorodności formatów informatycznych wykorzystywanych przy opracowywaniu danych, braku standardów lub ich niestosowania. Problemy te dotyczą zarówno danych tworzonych przez pojedynczych badaczy, jak i danych komercyjnych czy opracowywanych przez instytucje rządowe. Powoduje to problemy z wymianą danych i ich implementacją do projektów (Pachół, Zieliński, 2003). Pewnym rozwiązaniem są normy ISO, ale nie wszyscy się do nich stosują.

Źle dobrana próba danych nie da wiarygodnych i istotnych statystycznie wyników, czyli nie pozwoli na ich uogólnienie na całą zbiorowość. Próba może być pobierana zarówno spośród obiektów dyskretnych, np.: budynki, ludzie, przedsiębiorcy, jak i z przestrzeni geograficznej (Jażdżewska, 2013). Opis sposobu pobierania próby powinien być dostępny i oceniony pod kątem jej reprezentatywności w metadanych.

Dane udostępniane użytkownikom systemów informatycznych powinny zawierać **metadane**, czyli „dane o danych". Brak metadanych lub ich zły opis jest istotnym problemem w analizie danych. Metadane pojawiły się, gdy zaczęto gromadzić dane lub informacje i wystąpiła potrzeba wyszukania określonej pozycji. Można wyróżnić trzy rodzaje metadanych, tj. metadane wyszukiwania, rozpoznania i stosowania. Są one niezbędne do sprawnego zarządzania bazami danych (Iwaniak, 2005). Opisują zbiory danych, tak aby umożliwić ich wyszukanie, czyli ich lokalizację, oraz relacje z innymi bazami danych (Nahotko, 2013). W odniesieniu do danych geograficznych i GIS metadane mogą dotyczyć m.in.: projektu GIS, arkusza mapy, warstwy *shape*, zdjęcia lotniczego, danych atrybutowych i innych. Brak metadanych lub ich niekompletność powoduje, że mogą być one niewiarygodne, mogą naruszać

prawa autorskie lub właściciela danych, przez co ich stosowanie wiąże się z dużym ryzykiem. Na jakość metadanych w Polsce miała wpływ dyrektywa INSPIRE (Infrastructure for Spatial Information in Europe) z 2007 r., w której dane przestrzenne modelujące środowisko określono w jego 34 aspektach (tematach) i opisano metadanymi (Gaździcki, 2008).

Dane atrybutowe lub graficzne w pracy geografa miast muszą mieć odniesienie przestrzenne, **brak atrybutu georeferencji** powoduje, że są bezużyteczne, dlatego kontrola danych powinna uwzględniać fakt, czy zawierają one informację o lokalizacji oraz sposób jej określenia.

Pewne problemy z eksploracją danych geograficznych dotyczą jedynie danych graficznych (kolumna 2 w tab. 1), które reprezentują różne aspekty środowiska geograficznego, wykorzystywane są w Systemach Informacji Geograficznej i opisują czas, miejsce i atrybuty (Longley, Goodchild, Maguire, Rhind, 2006). Dobrze skonstruowane bazy danych przestrzennych, pozwolą na wartościowe analizy zjawisk zachodzących w przyrodzie. Bazy danych przestrzennych składają się przynajmniej z dwóch komponentów danych atrybutowych lub graficznych (wektorowych lub rastrowych), połączonych ze sobą w systemie informatycznym (Urbański, 1997). Problemy pojawiające się dla danych graficznych zostaną przedstawione dla danych wektorowych lub rastrowych.

Już na wstępie procesu wyboru i analizy danych może pojawić się problem z wyborem odpowiedniego modelu danych graficznych: **wektorowy czy rastrowy**. Każdy z nich ma swoje wady i zalety, o których należy mieć odpowiednią wiedzę. Ponadto część danych wektorowych powstaje z wektoryzacji danych rastrowych i odwrotnie, a jakość tego procesu ma duży wpływ na jakość danych. Ponadto część metod analizy danych przestrzennych jest odpowiednia dla jednego z tych modeli, dlatego decyzja o jego wyborze jest bardzo ważna (Urbański, 1997, 2008; Werner, 2004).

Wybór **jednostki przestrzennej** w naukach geograficznych jest jedną z podstawowych kwestii badawczych. Zbyt duże jednostki dają bardziej zgeneralizowane wyniki, zbyt małe nie pozwalają na uogólnienia. Źle dobrane jednostki mogą dawać fałszywe wyniki, niezgodne z faktycznym obrazem zjawiska w przestrzeni. Analiza korelacji i regresji w odniesieniu do różnych atrybutów i jednostek przestrzennych wykazała, że agregacja danych ma wpływ na wyniki (Openshaw, Taylor, 1979; Openshaw, 1984). Potwierdziła to w swojej pracy Nalej (2019), która wykazała, że problem zmiennych jednostek odniesienia (Modifiable Areal Unit Problem – MAUP) ma wpływ na wyniki badań pokrycia terenu w zależności od skali danych, rodzaju oraz wielkości jednostek przestrzennych zastosowanych w analizach (Nalej, 2019). Wybór jednostek jest związany też z kosztami projektu, bowiem więcej danych więcej kosztuje i wymaga lepszego sprzętu komputerowego. Ostateczna decyzja o wyborze jednostek przestrzennych należy do badacza lub zespołu.

Niepewność i nieokreśloność danych geograficznych jest związana ze specyfiką tych danych. Może wynikać z trudności z wydzielaniem granic regionów, gdyż granice między nimi bywają nieostre (Kraak, Ormeling, 1998), m.in. map użytkowania ziemi, rozdzielczości przestrzennej map rastrowych, np. gdy piksel reprezentuje więcej niż jeden typ pokrycia terenu (Longley, Goodchild, Maguire, Rhind, 2006). W przypadku atrybutów można również spotkać się z niejednoznacznymi lub zmiennymi w czasie definicjami, np. zbieranymi przez urzędy statystyczne informacjami o bezrobociu w różnych okresach, a także koniecznością wyboru, np. funkcji budynku, w przypadku gdy pełni on ich kilka. Pojawia się też problem z **niejednoznacznością nazw geograficznych** w zależności od uwarunkowań kulturowych lub języków, w jakich one funkcjonują (Longley, Goodchild, Maguire, Rhind, 2006).

Słowo **skala** ma wiele znaczeń w języku polskim[4], w przypadku danych geograficznych można ją odnieść do: stosunku odległości na mapie do odległości w terenie (skala mapy), w statystyce do skali pomiarowej, poziomu szczegółowości danych. W pierwszym przypadku mapy małoskalowe obejmują swoim zasięgiem duży obszar, ale są mniej szczegółowe, w przypadku map wielkoskalowych jest odwrotnie. Kiedy mamy do czynienia z danymi rastrowymi, pojawia się pojęcie rozdzielczości przestrzennej – przy wyższej rozdzielczości piksel jest mniejszy, a informacje bardziej dokładne, przy niższej jest odwrotnie. Przeszukiwanie i pozyskiwanie danych powinno więc uwzględniać niezbędny do badań poziom szczegółowości danych, gdyż mało szczegółowe mogą być mało wartościowe, a zbyt szczegółowe spowolnić analizę.

Ogromne zasoby map i planów historycznych mają postać rastra i wymagają znajomości metody kalibracji (nadawania georeferencji, wpasowania przestrzennego lub rejestracji w układzie współrzędnych), która polega na usunięciu zniekształceń spowodowanych skanowaniem i zniekształceniami mapy papierowej oraz zdefiniowaniu układu geo-

dezyjnego (Jaskulski, Łukasiewicz, Nalej, 2013). **Błędna kalibracja danych rastrowych** skutkuje utratą precyzji, kartometryczności i możliwości włączenia ich do szerszego zbioru danych (Graf, Kaniecki, Medyńska-Gulij, 2008).

Dane wektorowe wymagają jak najlepszej precyzji zapisu. Od jej jakości zależą też interpretacje wyników badań. Można spotkać się z kilkoma problemami podczas **generalizacji kształtu danych**. Zmiana skali mapy na mniejszą powoduje konieczność jej generalizacji, np. wybór obiektów, ich uproszczenie, agregacja i inne, co w konsekwencji prowadzi do utraty informacji zawartej na mapie. Jej proces nie jest przypadkowy, ale określony pewnymi zasadami (Iwaniak, Paluszyński, Żyszkowska, 1998).

Podczas zmiany formatu z rastrowego na wektorowy pojawia się bardzo dużo **błędów wektoryzacji/digitalizacji danych**, np. niedociągnięcia i przeciągnięcia, wiszące segmenty, duplikaty, które należy skorygować. Powodują one nie tylko błędy w obliczeniach, ale również uniemożliwiają poprawną analizę przestrzenną. Często łączą się z **błędami w topologii obiektów**, czyli relacjami geometrycznymi między obiektami. Dane geometrycznie niemające topologii lub niepoprawnie określoną nie pozwalają na późniejszą analizę sieciową czy sąsiedztwa. Można ją sprawdzić za pomocą testów i skorygować.

Są różne sposoby **określenia położenia geograficznego**, zapewne najdokładniejszy jest zapis za pomocą długości i szerokości geograficznej, ale przy pozyskiwaniu danych przestrzennych z różnych źródeł można spotkać się z różnymi zapisami lokalizacji, np. z państwowymi lub lokalnymi systemami odniesień przestrzennych, numerem arkusza, kodem pocztowym, adresem, jednostką administracyjną czy nazwą geograficzną.

Coraz więcej danych można pozyskać z różnych źródeł: rządowych, społecznościowych, naukowych, komercyjnych i wielu innych. Liczba danych wzrasta tak szybko, że pamięć komputera nie jest w stanie ich przetworzyć i potrzebne są nowe narzędzia oraz technologie do ich przetwarzania, nazwano je *big data* (Mayer-Schönberger, Cukier, 2017).

3. WNIOSKI

Eksploracja danych na potrzeby badań geografii miast, i innych dyscyplin naukowych, zajmuje dużą część pracy badawczej i powinna być postrzegana jako jeden w ważniejszych etapów pracy lub jako oddzielny problem badawczy. Od uzyskanych w jej końcowej fazie danych będzie bowiem zależeć cały dalszy proces wnioskowania i odkrywania wiedzy. Z tego powodu metodologia eksploracji danych ma już określoną rangę w badaniach naukowych. Również w naukach geograficznych jest ona potrzebna ze względu na liczbę danych dostępnych do analiz. Można w niej korzystać z doświadczeń technik wyszukiwania danych (*data mining*) w zakresie danych atrybutowych (Cai, Xie, 2007), trudniej jest z danymi graficznymi. Błędy pojawiają się zarówno na etapie badań terenowych, jak i ich przetwarzania (Wolski, 2012) oraz w zbiorach danych pozyskanych ze źródeł zewnętrznych. Napotkane problemy z danymi są niestety często niekomentowane przez badaczy i nieopisywane w artykułach. Wynika to często z ograniczenia liczby znaków, które autor musi uwzględnić składając artykuł do druku lub w przypadku braku określonych wymagań redaktorów czasopism. W ostatnich latach pojawiły się nowe czasopisma naukowe (*Data Journal*), na łamach których można podzielić się z innymi doświadczeniem z eksploracją danych i przystosowaniem ich do dalszych badań. Wśród nich pojawiły się np. *Data Science Journal* czy *Geoscience Data Journal*, w których dane geograficzne mogą być formalnie naukowo recenzowane i publikowane. Dzięki publikacji w tego typu czasopismach można opisywać sposoby tworzenia przez siebie baz danych, dzielić się swoimi doświadczeniami z innymi, podejmować współpracę lub zyskać wzrost liczby cytowań.

Pojawienie się licznych publikacji z zakresu pozyskiwania i przygotowywania danych do dalszej analitycznej pracy świadczy o problemach, z jakimi badacze spotykają się, a także o randze i jakości danych.

Innym ważnym zagadnieniem jest dzielenie się danymi z innymi badaczami, czyli ich udostępnianie w Internecie w ramach „otwartej nauki". Szczególnie geografowie mogą pozyskać w ten sposób wiele danych, a nie przygotowywać ich niepotrzebnie od nowa w każdym ośrodku badawczym. Zapewne zdarzają się sytuacje, kiedy naukowiec rezygnuje z pracy, a zebrane przez niego dane pozostają w szufladzie lub na dysku tak długo, aż zostaną usunięte. Wiele takich wartościowych danych już zniknęło albo nadal leżą w szufladzie.

Dane można by udostępniać w różny sposób, np. na stronach internetowych instytucji, wydawców artykułów lub specjalnie do tego przeznaczonych repozytoriach danych cyfrowych (Assante, Candela, Castelli, Tani, 2016). W Polsce funkcjonuje

bezpłatne, przeznaczone dla naukowców ze wszystkich dziedzin nauki, nowe Repozytorium Otwartych Danych – RepOD[5], które uruchomiło Interdyscyplinarne Centrum Modelowania Matemtycznego i Komputerowego na Uniwersytecie Warszawskim, a także jest dostępne – wspierane przez Europejską Organizację Badań Jądrowych CERN – repozytorum Zenodo[6] przeznaczone dla tzw. małych danych. Dane w repozytoriach mają odniesienie DOI, co ułatwia ich cytowanie.

Problem z danymi pozyskanymi przez naukowców pracujących w instytucjach publicznych jest taki sam jak z danymi zbieranymi przez inne podmioty państwowe. Przez wiele lat domagano się, aby były one bezpłatne, gdyż są finansowane przez podatników. Z badań z roku 2015 Jachimczyka wynika, że część polskich instytutów badawczych udostępnia swoje zasoby. Co trzeci zasób pochodzi od trzech instytutów: Państwowego Instytutu Geologicznnego (PIG), Ośrodka Przetwarzania Informacji (OPI) i Instytutu Badawczego Leśnictwa (Jachimczyk, 2015). Nie jest to jednak sytuacja występująca lub choćby znana w uczelniach państwowych. Może przyszedł czas, aby naukowcy udostępniali zgromadzone dane innym. Należałoby wszcząć dyskusję na tą kwestią. Wymaga ona od udostępniającego nie tylko trudnej decyzji o podzieleniu się swoimi zasobami, ale również czasu na sprawdzenie ich jakości i przygotowania metadanych. Umożliwiłoby to podejmowanie studiów porównawczych w dziedzinie geografii miast w różnym zakresie, przekładając się na upowszechnianie nowej wiedzy.

Podsumowując można potwierdzić przytoczoną we wstępie sugestię, że dane dotyczące miast powinny być przedmiotem badań i niezbędnym etapem badania zjawisk oraz procesów przestrzennych zachodzących w nich. Praca nad nimi powinna być udostępniana szerokiemu gronu naukowców w postaci zbiorów danych lub opisu ich eksploracji. Wymaga to współdziałania klasycznych metod zbierania danych z nowymi źródłami dostępnymi w sieci internetowej, a w konsekwencji współpracy geografów miast ze specjalistami GIS (GIScience). Może to dać nowy impuls w rozwoju geografii miast, a także GIScience.

Na koniec refleksja nad faktem „produkcji" coraz większej liczby danych, której nie ma możliwości ani ludzkich, ani technicznych opracować. Musimy unikać „dyktatury" danych: dane bowiem powinny jedynie wspierać, a nie zastępować podejmowanie decyzji przez inteligentnych i sceptycznych ludzi (Miller, Goodchild, 2015).

PRZYPISY

[1] *Garbage in, garbage out* (akronim: GIGO) to angielski zwrot mówiący o tym, że nawet, gdy program lub procedura informatyczna przetwarzania danych były poprawne, to jeśli będą wprowadzone błędne dane, zostaną uzyskane błędne wyniki (Hand, 2005).

[2] https://www.openstreetmap.org/

[3] Ewidencja gruntów i budynków, http://www.gugik.gov.pl/projekty/zsin-faza-i/dane-egib

[4] https://sjp.pwn.pl/sjp/skala;2575516.html (9.01.2020).

[5] https://repod.pon.edu.pl/pl

[6] https://zenodo.org/

BIBLIOGRAFIA

Adamczyk, J., Będkowski, K. (2018). Źródła numerycznych danych geoprzestrzennych. W: A. Obidziński (red.), *Inwentaryzacja i waloryzacja przyrodnicza. Metody naziemne i geomatyczne* (s. 17–27). Warszawa: Wyd. SGGW.

Ash, J., Kitchin, R., Leszczynski, A. (2018). Digital turn, digital geographies? *Progress in Human Geography*, *42* (1), s. 25–43; https://doi.org/10.1177/0309132516664800

Assante, M., Candela, L., Castelli, D., Tani, A. (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal*, *15* (6); https://doi.org/10.5334/dsj-2016-006

Bach, M., Stańczak, M., Werner, A. (2009). Wpływ przyjętego modelu wersjonowania danych na efektywność relacyjnej bazy danych. *Studia Informatica*, *30* (2B), s. 253–263.

Bell, G. (2009). Foreword. W: T. Hey, S. Tansley, K. Tolle (red.), *The fourth paradigm. Data-Intensive scientific discovery* (s. xi–xvi). Redmond, Washington: Microsoft Research.

Bielecka, E. (2006). *Sytemy Informacji Geograficznej. Teoria i zastosowania.* Warszawa: Wyd. PJWSTK.

Blaschke, T., Merschdorf, H. (2014). Geographic information science as a multidisciplinary and multiparadigmatic field. *Cartography and Geographic Information Science*, *41* (3), s. 196–213; https://doi.org/10.1080/15230406.2014.905755

Cai, C., Xie, K. (2007). Measuring data quality of geoscience datasets using data mining techniques. *Data Science Journal*, *6*, S738–S742; https://doi.org/10.2481/dsj.6.S738

Dasu, T., Johnson, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons.

Foreman, J.W. (2019). *Mistrz analizy danych: od danych do wiedzy.* Tłum. K. Matuk. Gliwice: Helion.

Gaździcki, J. (2008). Implementacja dyrektywy INSPIRE w Polsce: stan aktualny, problemy i wyzwania. *Roczniki Geomatyki – Annals of Geomatics*, *6* (3), s. 23–32.

Graf, R., Kaniecki, A., Medyńska-Gulij, B. (2008). Dawne mapy jako źródło informacji o wodach śródlądowych i stopniu ich antropogenicznych przeobrażeń. *Badania Fizjograficzne nad Polską Zachodnią, Seria A – Geografia Fizyczna*, *59*, s. 11–27.

Guo, H., Goodchild, M.F., Annoni, A. (red.) (2020). *Manual of digital Earth*. Singapore: Springer Open; International Society for Digital Earth.

Hand, D.J. (2005). *Eksploracja danych*. Tłum. A. Chądzyńska. Warszawa: Wydawnictwa Naukowo-Techniczne.

Iwaniak, A. (2005). Metodyka opracowania i stosowania metadanych w Polsce. *Roczniki Geomatyki – Annals of Geomatics*, *3* (3), s. 47–58.

Iwaniak, A., Paluszyński, W., Żyszkowska, W. (1998). Generalizacja map numerycznych – koncepcje i narzędzia. Cz. 1. *Polski Przegląd Kartograficzny*, *30* (2), s. 78–88.

Jachimczyk, A. (2015). Otwarte dane badawcze. Casus polskich instytutów badawczych. *Zagadnienia Naukoznawstwa*, *2206*, s. 409–424.

Jaskulski, M., Łukasiewicz, G., Nalej, M. (2013). Porównanie metod transformacji map historycznych. *Roczniki Geomatyki – Annals of Geomatics*, *11* (4), s. 41–56.

Jażdżewska, I. (2013). *Statystyka dla geografów*. Łódź: Wyd. Uniwerstetu Łódzkiego.

Jażdżewska, I. (2018). The use of centrographic measures in analysing the dispersion of historic factories, villas and palaces in Lodz (Poland). *Folia Geographica*, *60* (1), s. 50–61.

Kraak, M.J., Ormeling, F. (1998). *Kartografia – wizualizacja danych przestrzennych*. Tłum. W. Żyszkowska. Warszawa: Wydawnictwo Naukowe PWN.

Larose, D.T. (2013). *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji danych.* Tłum. A. Wilbik. Warszawa: Wydawnictwo Naukowe PWN.

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (2006). *GIS. Teoria i praktyka*. Warszawa: Wydawnictwo Naukowe PWN.

Mayer-Schönberger, V., Cukier, K. (2017). *Big data: rewolucja, która zmieni nasze myślenie, pracę i życie, efektywna analiza danych.* Tłum. M. Głatki. Warszawa: Wyd. MT Biznes.

Miller, H.J., Goodchild, M.F. (2015). Data-driven geography. *GeoJournal*, *80*, s. 449–461; https://doi.org/10.1007/s10708-014-9602-6

Nahotko, M. (2013). Współdziałanie metadanych w systemach informacyjnych. *Zagadnienia Informacji Naukowej*, *51* (1), s. 61–83.

Nalej, M. (2019). *Problem zmiennych jednostek odniesienia (MAUP) w badaniach pokrycia terenu. Przykład Łódzkiego Obszaru Metropolitalnego* (University of Lodz). Pobrano z http://dspace.uni.lodz.pl/xmlui/bitstream/handle/11089/26386/nalej-streszczenie.pdf?sequence=3&isAllowed=y

Openshaw, S. (1984). Modifiable Areal Unit Problem. W: *International encyclopedia of human geography*; https://doi.org/10.1016/b978-008044910-4.00475-2

Openshaw, S., Taylor, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. W: N. Wrigley (ed.), *Statistical applications in the spatial sciences* (s. 127–144). London: Pion.

Osowski, S. (2013). *Metody i narzędzia eksploracji danych*. Legionowo: Wyd. BTC.

Pachół, P., Zieliński, J. (2003). Wymiana danych wchodzących w skład krajowego systemu informacji o terenie. *Roczniki Geomatyki – Annals of Geomatics*, *1* (1), s. 38–52.

Ryza, S., Laserson, U., Owen, S., Wills, J. (2016). *Spark: zaawansowana analiza danych*. Tłum. A. Watrak. Gliwice: Helion.

Sadiq, S., Srivastava, D., Dasu, T., Dong, X.L., Freire, J., Ilyas, I., … Zhou, X. (2018). Data quality: The role of empiricism. *ACM SIGMOD Record*, *46* (4), s. 35–43; https://doi.org/10.1145/3186549.3186559

Shelton, T. (2017). Spatialities of data: mapping social media 'beyond the geotag'. *GeoJournal*, *82*, s. 721–734. https://doi.org/10.1007/s10708-016-9713-3

Shi, W., Fisher, P.F., Goodchild, M.F. (red.) (2002). *Spatial data quality*. London: Taylor & Francis.

Szeliga, M. (2017). *Data Science i uczenie maszynowe*. Warszawa: Wydawnictwo Naukowe PWN.

Urbański, J. (1997). *Zrozumieć GIS. Analiza informacji przestrzennej*. Warszawa: Państwowe Wydawnictwo Naukowe.

Urbański, J. (2008). *GIS w badaniach przyrodniczych*. Pobrano z https://kiw.ug.edu.pl/pl/ebooki/644-gis-w-badaniach-przyrodniczych.html

Werner, P. (2004). *Wprowadzenie do systemów geoinformacyjnych*. Warszawa: Wyd. Jark.

Wolski, J. (2012). Błędy i niepewność w procesie tworzenia map numerycznych. *Prace Komisji Krajobrazu Kulturowego*, *16*, s. 15–32.

Artykuł wpłynął:
2 września 2019
Zaakceptowano do druku:
27 października 2019