POLSKIE TOWARZYSTWO INFORMACJI PRZESTRZENNEJ
ROCZNIKI GEOMATYKI 2016 ❍ Tom XIV ❍ Zeszyt 2(72): 201–211

Jakość danych OpenStreetMap – analiza informacji o budynkach na terenie Siedlecczyzny

OpenStreetMap building data quality: the Siedleckie county study

Joanna Nowak Da Costa, Elżbieta Bielecka, Beata Całka

Wojskowa Akademia Techniczna, Wydział Inżynierii Lądowej i Geodezji

Słowa kluczowe: OpenStreetMap, dane o budynkach, wolontariackie dane przestrzenne VGI, kompletność, jakość danych
Keywords: OpenStreetMap, building data, Volunteered Geographic Information (VGI), completeness, data quality

Wprowadzenie

Zagadnienie jakości danych przestrzennych od wielu lat jest przedmiotem szerokiego zainteresowania nie tylko producentów i dystrybutorów danych, ale także użytkowników i naukowców. Znaczenie jakości danych w prowadzeniu działalności biznesowej i nauce jest dobrze rozpoznane i szeroko opisywane od co najmniej ćwierćwiecza (Redman, 1996; Loshin 2006; Xia i in., 2011; Bielecka, 2015). Z punktu widzenia dostawcy lub dystrybutora danych ocena jakości jest jednym z kluczowych elementów produkcji i zawsze jest analizowana w kontekście zgodności ze specyfikacjami technicznymi. Problem ten jest dość szeroko naświetlony przez międzynarodową organizację normalizacyjną, która kwestiom oceny i raportowania jakości danych poświęciła kilka norm, w tym normy: ISO 8000-Data Quality i ISO 19 157:2003-Geographic information – Data quality. W dokumentach normatywnych ISO jakość jest definiowana jako całościowy zespół cech i charakterystyk zbiorów danych i usług, które wpływają na możliwość zaspokojenia obecnych i przyszłych wymagań użytkowników (ISO19157:2003). Cechy i charakterystyki wspomniane w normie, w odniesieniu do zbiorów danych przestrzennych, są określane za pomocą kilkunastu wskaźników ilościowych i jakościowych (Bielecka, 2010). Przy czym, do najczęściej stosowanych i obowiązujących między innymi także dla zbiorów danych INSPIRE należą: kompletność (brak i nadmiar obiektów), spójność logiczna (pojęciowa, topologiczna, dziedziny, formatu), dokładność położenia, dokładność czasowa i dokładność tematyczna (np. poprawność klasyfikacji lub poprawność określenia atrybutów jakościowych) oraz pochodzenie (ang. *lineage*). Wszystkie ilościowe elementy jakości są oceniane w kontekście zgodności ze specyfikacjami technicznymi danych, a wyniki oceny raportowane w metadanych. Naukowcy zwracają także uwagę na dostępność danych, która jest często kluczowym elementem jakości oraz autorytet dystrybutora gwarantujący lepszą jakość (Xia i in., 2011).

W odniesieniu do danych zbieranych dobrowolnie i bezpłatnie przez bardzo dużą liczbę wolontariuszy, określanych w języku angielskim jako *volunteered geographic information* (VGI) lub *crowdsourcing geodata*, stosowanie wyżej wymienionych wskaźników staje się problematyczne. Powodem tego jest brak szczegółowych specyfikacji technicznych, a jedynie podanie zasad i wskazówek dostarczania danych oraz częsty brak formalnej weryfikacji wszystkich wprowadzanych danych. Wolontariuszom zostawia się zazwyczaj dużo swobody w zakresie dokładności wprowadzanych danych oraz szczegółowości ich charakterystyki opisowej. Weryfikacja danych jest wykonywana z reguły przez innych użytkowników, potencjalnie lepiej znających dany teren lub chcących wykorzystać dane społecznościowe do realizacji swoich zadań.

Najczęściej badanym i ocenianym pod kątem jakości danych jest OpenStreetMap (OSM). Od co najmniej dekady wielu naukowców na całym świecie analizuje jakość danych OSM skupiając się głównie na kompletności i dokładności położenia dróg.

W artykule przedstawiono uwarunkowania heterogenicznej charakterystyki OpenStreetMap, zwracając uwagę na aspekt niedoskonałości ustaleń semantycznych i założeń jakościowych tej oddolnej inicjatywy. Przeanalizowano także kompletność i dokładność lokalizacji wybranych klas obiektów OSM w stosunku do krajowych danych urzędowych zgromadzonych w bazie danych obiektów topograficznych (BDOT10k). Analizy wykonano dla powiatu siedleckiego i miasta Siedlce. Opracowanie to dopełnia dotychczasowe rezultaty badawcze w zakresie analiz ilościowych jakości OSM, szczególnie w odniesieniu do terytorium Polski.

OpenStreetMap

Istota i założenia ogólne OSM

W 2004 roku Steve Coast, zniecierpliwiony permanentnym brakiem otwartego dostępu do danych przestrzennych w Wielkiej Brytanii, zainicjował projekt wolnej, otwartej i edytowalnej mapy OpenStreetMap (OSM). Projekt, którego misją jest dostarczanie zarówno gotowych map, jak i surowych geodanych „wszystkim, którzy tego potrzebują", jest tworzony przez wolontariuszy na całym świecie. Mimo swojej nazwy OpenStreetMap nie jest tylko mapą dróg. Drogi stanowią 28% wszystkich obiektów, a najliczniej reprezentowane są obiekty o znaczniku (tag) 'building' (OSM, 2015a).

System OSM bazuje na idei otwartego serwisu społecznościowego i wykorzystuje technologię wiki, co w praktyce oznacza, że każdy w dowolnym momencie może dodać lub edytować dowolny obiekt bazy. W bazie przechowywana jest historia edycji każdego obiektu (nawet już nieistniejącego), dzięki czemu można wycofać skutki pomyłek lub wandalizmu. OSM ma własną infrastrukturę do przechowywania, udostępniania, przeszukiwania i wizualizacji danych, która nie jest zgodna z standardami OGC. Niezgodność tę Haklay i Weber (2008) tłumaczą niemożnością pogodzenia idei pracy „w stylu wiki", z powszechnie obowiązującymi standardami z zakresu informacji geograficznej. Dane OSM są przechowywane w relacyjnej bazie PostgreSQL, bez rozszerzeń przestrzennych, w układzie współrzędnych WGS84. Do reprezentacji geometrii używa się podstawowych typów geometrycznych (ang. *primitive types*), które w połączeniu z dość dowolnym schematem etykietowania pozwalają opisać praktycznie każdy obiekt geograficzny, włącznie z jego topologią. Na etykietę (tag), zwaną też znacznikiem, składa się para klucz=wartość, którą można utożsamić z atrybutem,

przykładowo building=yes oznacza budynek (ten obiekt jest budynkiem), *building=house* oznacza dom jednorodzinny.

Najczęstszym typem reprezentacji geometrycznej budynków i budowli w bazie OSM jest wielobok obrazujący ich obrys. Użytkownicy-dostawcy danych OSM do pozyskania danych o budynkach wykorzystują różne techniki oraz źródła danych, między innymi wektoryzację ortoobrazów, pomiar ręcznym odbiornikiem GPS podczas spaceru (lub nawet przejazdu rowerem) dookoła budynku, szkic lub pomiar z poziomu najbliższej drogi, zaimportowanie upublicznionych danych oficjalnych, itp. (OSM, 2015b; Arsanjani i in., 2015; Ramm i in., 2010). Zależność dokładności geometrycznej od metod i technik pozyskiwania danych wyjaśnia heterogeniczną dokładność geometryczną danych w bazie OSM. Ponadto, na efekt różnorodności technik i urządzeń do pozyskiwania obiektów, nakłada się efekt interpretacji. Prawidłowa interpretacja obrysu budowli szczególnie na podstawie obrazów satelitarnych lub lotniczych wymaga doświadczenia, najlepiej podpartego odpowiednim szkoleniem. W 2007 roku Goodchild (2007), ambasador idei i terminu VGI, wyraził opinię, iż dobrze rozróżnialne obiekty geograficzne są mniej wymagające jeśli chodzi o szkolenie i doświadczenie obserwatorów-wolontariuszy. Traktując tę opinię jako hipotezę badawczą potwierdzimy, że budynki i wybrane budowle należą do dobrze rozróżnialnych obiektów geograficznych, czego dowodem jest niewielki średni kwadratowy błąd położenia budynków OSM względem danych referencyjnych wynoszący dla Siedlec 0,49 m i 1,23 m dla powiatu siedleckiego.

Metodyczne aspekty oceny jakości danych OSM

Jakość danych OpenStreetMap, a w szczególności takie jej elementy ilościowe jak kompletność oraz dokładność położenia, wzbudza szerokie zainteresowanie potencjalnych użytkowników na świecie. Wolontariacki sposób zbierania danych, bez szczegółowych wytycznych technicznych uniemożliwia zastosowanie wprost zasad oceny danych geograficznych zawartych w normie ISO19157, która odwołuje się do porównania danych ze specyfikacją techniczną. Goodchild (2012) wymienił trzy alternatywne podejścia do oceny jakości danych geograficznych pozyskiwanych w ramach projektów takich jak OpenStreetMap:

1) ujęcie wolontariackie (ang. *crowd-sourcing approach*) – bazujące na założeniu, że błędne dane zostaną wykryte i poprawione przez użytkowników,
2) ujęcie społeczne (ang. *social approach*) – zakładające minimalną kontrolę poprawności danych przez administratorów,
3) ujęcie geomatyczne (ang. *geographical approach*) – zakładające wykorzystanie programów typu GIS do kontroli jakości danych przez sprawdzenie poprawności topologii i reguł logicznych.

Takie podejścia do oceny jakości danych VGI nabierają popularności, choć nadal powszechnie stosowana jest ocena zewnętrzna, wymagająca dostępu do innych, najczęściej dokładniejszych i bardziej wiarygodnych danych. Podejście takie umożliwia ocenę kompletności danych, która jest dość istotna z punktu widzenia potencjalnych użytkowników. Dobór danych referencyjnych stanowi jednak problematyczną kwestię. Obawę o to jaki wybrać zbiór referencyjny dla celów kontroli jakości baz danych przestrzennych zasilanych przez ochotników nie-kartografów wyrażali również Goodchild i Li (2012) oraz Goodchild i Glennon (2010) i Haklay (2010). A zdaniem Cichocińskiego (2013) metoda oceny jakości wykorzystująca dane zewnętrzne jest skomplikowana, nie zawsze wiarygodna i kłóci się z ideą OSM.

Do najczęściej analizowanych danych OSM należy sieć drogowa. Badania takie przeprowadzono dla wielu państw europejskich, a także Brazylii, USA, Iranu i innych. Do oceny wykorzystywano na ogół dane referencyjne, które stanowiły bazy urzędowe. Uzyskane wyniki dotyczące kompletności pokrycia drogami mapy OSM oraz geometrycznej dokładności względnej są zbliżone. Haklay (2010), Zielstra i Zipfl (2010), Esmail i inni (2013) oraz Zielstra i inni (2013) zauważyli, że kompletność danych o drogach jest bardzo zróżnicowana i znacznie większa w dużych niż małych miastach, najmniejsza zaś na obszarach rolniczych. Podobnie oceniona została dokładność: wysoka dla dróg głównych w dużych miastach, niska – dla dróg lokalnych na terenach rolnych i leśnych. Fakt ten tłumaczony jest większą aktywnością wolontariuszy zamieszkałych w dużych miastach niż w małych miastach i wsiach, zaś większe doświadczenie wpływa na znaczne zmniejszenie liczby popełnianych błędów (Ma i in., 2015). Według często cytowanego Haklay'a, w centrum Londynu uzyskano wyniki średniej różnicy położenia 3,2-4,8 m, podczas gdy w dzielnicach obrzeżnych spadała do 6,8-8,3 m (z maksymalnymi 20-metrowymi odchyłkami) pomiędzy drogami zarejestrowanymi w bazie OSM a danymi Ordnance Survey.

Analiza jakości danych o budynkach była przeprowadzona jedynie dla niewielkiego obszaru miasta Cramlington oraz terenów wiejskich Clara Valey, w hrabstwie Nortumberland, w Wielkiej Brytanii (Al-Bakri, Fairbairn, 2010) i obejmowała jedynie określenie dokładności odwzorowania kształtu i lokalizacji 40 budynków. Ostatnio, na dużym terenie Niemiec (prawie 15% całego terytorium), przebadano znaczącą liczbę budynków, zarówno na terenie silnie i mniej zurbanizowanym. Hecht, Kunze i Hahmann (2013) skupili się głównie na porównaniu położenia budynków OSM względem cyfrowych danych państwowych rezygnując z porównań na poziomie semantycznym.

Zważywszy na fakt, że większość dotychczasowych analiz dotyczących jakości i przydatności OSM dotyczy dróg w dużych miastach, przedstawione poniżej wyniki badań mają charakter nowatorski, obejmują bowiem analizę danych o budynkach dla usytuowanego peryferyjnie obszaru małego miasta oraz okalających go terenów powiatu ziemskiego.

Eksperyment badawczy

Charakterystyka obszaru i wykorzystanych danych

Obszar testowy stanowi powiat siedlecki oraz miasto Siedlce (powiat grodzki). Jest to mało zurbanizowany i dość słabo zaludniony obszar, leżący w środkowo-wschodniej Polsce. Siedlce należą do średnich miast w Polsce zarówno pod względem demografii (48 lokata w Polsce i 4 w województwie mazowieckim), jak i pod względem rozwoju gospodarczego, edukacji i kultury.

Na testowane dane składały się obiekty OSM mające znacznik 'budynek' (tag:building). Dane w formie plików shape (format Esri) zostały pozyskane z serwisu OSM Geofabrik (www.geofabrik.de). Aktualność badanych danych OSM to 28 maja 2015 roku. Analizowany zestaw danych zawiera obiekty powierzchniowe w liczbie 24 000, z czego większość leży w mieście Siedlce (21 434). Jako zbiór referencyjny wykorzystano dane zgromadzone w bazie danych obiektów topograficznych BDOT10k, o aktualności na rok 2013 (16/8/2013). Obiekty bazy OSM o znaczniku 'building' porównywano z obiektami należącymi do kategorii klas obiektów 'budynki, budowle i urządzenia' (BU) BDOT10k. W szczególności

skupiono się na klasie obiektów 'budynki' (BUBD). Dodatkowo wykorzystano nieliczne obiekty z klas 'inne urządzenia techniczne' (BUIT) i 'wysoka budowla techniczna' (BUWT) wyrażone zarysem podstawy lub posiadające punktową reprezentację geometryczną. Ogólną charakterystykę danych OSM w zestawieniu z danymi BODT10k zawiera tabela.

Tabela. Charakterystyka analizowanych danych OSM i danych referencyjnych BDOT10k

Cechy	Obiekty OpenStreetMap o znaczniku 'building'	Obiekty klas obiektów 'budynki, budowle i urządzenia'	
		obiekty klasy 'budynki' (BUBD)	obiekty klas BUIT i BUWT
Definicja	brak	Obiekty budowlane, trwale związane z gruntem, posiadające fundamenty, wydzielone z przestrzeni za pomocą przegród budowlanych (…).	Pozostałe urządzenia techniczne, istotne z topograficznego punktu widzenia, nieuwzglednione w innych klasach z obiektami technicznymi i niebędące budynkami. Wysokie budowle o różnym przeznaczeniu.
Reprezentacja geometryczna	Poligon/ rekomendowany obrys podstawy budynku	Poligon / zarys podstawowy lub maksymalny zasięg	Poligon lub punkt / zarys podstawy lub środek podstawy
Sposób pozyskania danych	Pomiary z ręcznych odbiorników GPS, zdjęć lotniczych oraz innych dostępnych źródeł danych.	Ewidencja gruntów i budynków, wektoryzacja ortofotomapy, pomiar terenowy.	Pomiar geodezyjny, ewidencja gruntów i budynków lub inne rejestry państwowe.
Dokładność, szczegółowość	Heterogeniczna dokładność i szczegółowość, zależna od techniki pozyskania danych, uszczegółowienia obrysu obiektu oraz umiejętności obserwatora.	Szczegółowość i dokładność odpowiadająca skali 1:10 000.	
Kontrola jakości	Kodeks postępowania dostarczyciela danych (odpowiedzialność honorowa, moralna). Możliwość weryfikacji danych geometrycznych i opisowych przez pomiar innego użytkownika. Możliwość wykorzystania istniejących narzędzi do kontroli jakości danych metodami wewnętrznymi.	Reguły pomiarowe i nadzór technologiczny pomiarów oraz system kontroli danych przekazywanych do zasobu BDOT10k (kontrole topologii i geometrii, kontrole semantyczne, syntaktyczne i atrybutowe, itp.).	
Aktualność	Zróżnicowana, zależna od aktywności wolontariuszy	Aktualizacja bieżąca, niezwłocznie po uzyskaniu nowych danych ze zbiorów zasilających	
Układ współrzędnych	WGS84	PL 1992	

Metoda badań

Zestawione w tabeli cechy charakteryzujące oba zbiory unaoczniają podstawowe źródła rozbieżności pomiędzy analizowanymi zbiorami, a mianowicie: model pojęciowy, reguły pomiarowe i nadzór technologiczny oraz system kontroli danych przekazywanych i przechowywanych w zasobach baz OSM i BDOT10k. Mając na względzie wymienione różnice w obu zbiorach prace badawcze obejmowały: analizę semantyczną, identyfikację odpowiadających sobie obiektów w obu zbiorach, analizę dokładności geometrycznej oraz kompletności obiektów i atrybutów.

Badania semantyki wykonano analizując definicje budynków oraz definicje i opisy atrybutów budynków w obu zbiorach. Poza Polską Klasyfikacją Obiektów Budowlanych (PKOB, 1999), stosowaną do określenia funkcji ogólnej budynku w BDOT10k, przeanalizowano fora dyskusyjne społeczności internetowej OSM (OSM, 2015c) wyjaśniające znaczenie poszczególnych znaczników przypisanych do budynków.

Analizę dokładności przeprowadzono na podstawie ręcznego pomiaru odpowiadających sobie narożników budynków OSM względem BDOT10k. Do pomiaru wybrano 661 budynków, z czego 316 jest położonych na terenie powiatu ziemskiego i 345 powiatu grodzkiego. Budynki te zlokalizowane są na terenie każdej z gmin. Wyniki przedstawiono w postaci średniego błędu kwadratowego (RMSE). Kompletność obiektów i atrybutów wykonano przez porównanie budynków i ich atrybutów z OSM z danymi z BDOT10k. Odpowiadające sobie budynki w obu zbiorach dopasowano automatycznie, wykorzystując zapytanie przestrzenne umożliwiające przyporządkowanie centroidów budynków OSM wnętrzu wieloboku określającego położenie budynków w BDOT10k.

Wyniki

Analiza kompatybilności semantycznej

Podstawowa trudność w harmonizacji semantycznej OSM z innymi danymi wynika z braku ogólnie przyjętych reguł definiowania obiektów i zdanie się na intuicyjne rozumienie terminu 'budynek' przez wszystkich wolontariuszy. Nie ma też rygorystycznych zaleceń odnośnie przypisywania atrybutów budynkom, stąd aż 67,8% budynków na terenie powiatu i 23,6% w mieście ma jedynie etykietę o wartości 'yes' co oznacza, że obiekt jest budynkiem, budowlą lub urządzeniem. Lista dostępnych wartości znaczników wykorzystywanych do oznaczenia budynków obejmuje 53 pozycje, z których większość opisuje funkcje, a niektóre sposoby użytkowania. Są one zgrupowane w cztery kategorie: budynki mieszkalne (do których zaliczone są także hotele), budynki komercyjne, budynki publiczne oraz inne. Zastosowana klasyfikacja jest niepełna, nierozłączna, niejednoznaczna i reprezentuje różne poziomy hierarchii. Z tego też powodu harmonizację znaczeniową wykonano ręcznie, poszukując odpowiedników poszczególnych znaczników w wartościach atrybutu 'funkcja ogólna'. Z analizy tej wynika, że nie można jednoznacznie przyporządkować sobie odpowiednich atrybutów, na przykład budynek typu 'residential' odpowiada czterem klasom budynków mieszkaniowych (jednorodzinne, o dwóch mieszkaniach, o trzech i więcej mieszkaniach oraz zbiorowego zamieszkania), jednocześnie istnieje etykieta 'house' oznaczająca dom jednorodzinny, 'detached' odpowiadające w polskiej klasyfikacji także domowi jednorodzinnemu oraz 'terrace' – domy jednorodzinne w zabudowie szeregowej. Jednoznaczna odpowiedniość

zachodzi pomiędzy budynkami typu: church – świątynia, garage – garaż, industrial – budynki przemysłowe. Najczęściej pomiędzy cechami budynków i budowli zachodzi relacja jeden do wielu, przykładowo 'school' i 'education' odpowiadają klasie BUBD15 oznaczającej budynki szkół i instytucji badawczych; 'electricity' obejmuje dwie klasy: transformator lub zespół transformatorów, itp. Do budynków, użytkownicy OSM zaliczyli także wiaty lub zadaszenia, klasyfikowane w BDOT jako obiekty inne o znaczeniu orientacyjnym (OIOR11). Warto także zaznaczyć, iż niektóre budowle i urządzenia zaklasyfikowane w Polskiej Klasyfikacji Obiektów Budowlanych do budynków i budowli mogą mieć w OSM znacznik 'man-made' zamiast 'building'. Wolontariusze najczęściej zaliczają tutaj wysokie budowle techniczne (BUWT).

Analiza dokładności położenia

Maksymalne przesunięcie narożników budynków OSM względem odpowiadających sobie narożników budynków w BDOT10k wynosi 9,4 m na terenie powiatu siedleckiego oraz 6,1 m na terenie miasta Siedlce. Odpowiednio, średni błąd kwadratowy położenia budynku w stosunku do BDOT10k wynosi 1,23 m dla powiatu i 0,49 m dla miasta. Na 661 analizowanych budynków jedynie w 7 przypadkach zaobserwowano błędy, które można zaklasyfikować jako błędy grube (rys. 1).

Analiza kompletności

Kompletność budynków w bazie OSM dla dwóch analizowanych powiatów jest krańcowo różna. Dla powiatu ziemskiego do bazy wprowadzono zaledwie 3,29% budynków, z czego aż 67,8% nie ma określonej funkcji (tag:yes), 30% to, według wolontariuszy OSM, budynki mieszkalne określane jako 'residential' lub 'house'. W Siedlcach (powiat grodzki) zaobserwowano nadmiar budynków OSM w stosunku do BDOT10k, wynoszący 20,2%. W bazie OSM znajduje się 5599 budynków o powierzchni mniejszej niż 40 m^2 (z czego jedna czwarta nie osiąga nawet 10 m^2), które zgodnie z rozporządzeniem w sprawie bazy danych obiektów topograficznych oraz bazy danych ogólnogeograficznych, a także standardowych opracowań kartograficznych (MSWiA, 2012) są pominięte lub zagregowane. Dla przykładu: twórcy danych OSM wprowadzali jako osobne budynki pojedyncze, przylegające do siebie garaże, które w bazie BDOT10k stanowią jeden obiekt. Sytuację taką pokazano na rysunku 2a i 2b. Kreatorzy danych OSM często wydzielali jako osobne budynki fragmenty budynku różniące się wysokością lub rodzajem zadaszenia (np. niższe wejście do budynku). Takie różnice w modelowaniu zilustrowano na rysunku 2c i 2d.

Różnice w kompletności budynków na terenie poszczególnych gmin analizowanego obszaru pokazano w formie kartogramu na rysunku 3. Analiza tej mapy wskazuje tendencję zmniejszenia kompletności wraz z oddalaniem się od miasta Siedlce. Dla gminy wiejskiej Siedlce wynosi ona niespełna 15%, podczas kiedy dla pozostałych gmin waha się od 3,15 do 0%. W kilku gminach twórcy danych OSM wprowadzili pojedyncze budynki, kościół, szkołę lub urząd (po jednym budynku w gminach Skórzec, Domanice i Przesmyki, 3 budynki w Suchożebrach).

Poszczególne budynki i budowle w OpenStreetMap są opisywane atrybutami bardzo skromnie. Procent budynków nieopisanych żadnym z atrybutów dla powiatu ziemskiego wynosi prawie 68%, a dla miasta 23,6%. Niespełna 0,5% budynków dla powiatu grodzkiego i 0,89% dla powiatu ziemskiego ma przypisane nazwy. W większości są to urzędy administracji publicznej, kościoły, hotele, restauracje, sklepy, a zatem z punktu widzenia użytkow-

nika ważne obiekty publiczne. Kontrola spójności pomiędzy funkcją budynku a jego nazwą wykazała wiele niezgodności, przykładowo wpisanie funkcji w miejscu nazwy ('sklep spożywczy', 'biblioteka'). Wolontariusze ciekawie zaklasyfikowali halę widowiskową w Siedlcach jako obiekt edukacyjny, a bank i kilka sklepów spożywczych jako budynki mieszkalne.

Dyskusja i wnioski

Zagadnienia ontologiczne w zakresie danych przestrzennych (w tym zagadnienie definicji budynku oraz sposobów jego użytkowania) w szerszym, międzynarodowym kontekście obejmuje uwarunkowania kulturowe, historyczne, geograficzne, a nawet socjologiczne. Listę dopełniają uwarunkowania formalno-pragmatyczne, będące często konsekwencją krajowych lub regionalnych działań o charakterze prawno-organizacyjnym (np. prawo budowlane). W efekcie, niezależnie stworzone bazy geodanych odnoszące się do innego, subiektywnego postrzegania rzeczywistości, wykorzystujące inne reprezentacje geometryczne, schematy modelowe, taksonomie, a nawet języki prowadzą do problemów komunikacyjnych w procesie współużytkowania danych (Nowak i in., 2005).

Kwestia harmonizacji semantycznej cech opisujących budynki w OSM i BDOT10k wydaje się obecnie nierozwiązywalna ze względu na różnice w modelu pojęciowym zbiorów, w tym naturalne aczkolwiek często nieścisłe, subiektywne i niewyczerpujące definicje dotyczące zastosowania znacznika 'building' i związanych z nim kluczy 'use' dla oznaczania obiektów bazy OSM. Sami członkowie społeczności OSM zauważają w tym pewną niekonsekwencję i źródło problemów co do jednoznacznego zapisania danych w bazie OSM oraz ich późniejszego wykorzystania ze zrozumieniem, na co wskazuje ciągle niezamknięta dyskusja na temat znacznika 'building' i klucza 'use' wśród członków społeczności (OSM, 2015c). Ewoluująca w sposób oddolny lista etykiet, bazująca na naturalnych i potocznych pojęciach, nie jest wolna od niedociągnięć. Koegzystencja synonimów oraz homonimów jest źródłem niepewności przy wprowadzaniu danych do bazy OSM, jak i przy korzystaniu z niej. Dodatkowo prezentacja etykiet w postaci prostej listy słownikowej w wiki-serwisie społecznościowym OSM nie ułatwia wykrycia ewentualnych mankamentów.

Podobnie wytyczne OSM w zakresie definicji obiektów o znaczniku 'building', jak i sposób ich pomiaru, można określić mianem lakonicznych i nieustandaryzowanych, co również skonstatowali nasi poprzednicy, niemieccy badacze Hecht, Kunze i Hahmann (2013). Podczas gdy w Polsce, ustalenia co do definicji i klasyfikacji budynków, budowli i urządzeń, sposobu pomiaru oraz ich reprezentacji kartograficznej są bardzo szczegółowe i zatwierdzone w postaci aktów prawnych.

W przypadku obiektów sieci drogowej, standardy ogólnoświatowe dotyczące ich topologii, geometrii i kategoryzacji, umożliwiają szczegółową wewnętrzną ocenę (i ewentualną poprawę) jakości zbiorów danych między innymi przez kontrolę spójności topologicznej lub tematycznej. Takie sprzyjające okoliczności na razie nie występują w przypadku budynków i budowli ani na poziomie światowym, ani w wytycznych OSM. Brak ujednolicenia sposobów pozyskiwania oraz modelowania danych klasy 'budynki/budowle' w bazie danych OSM skutecznie zawęża możliwość oceny jej jakości do niewiele ponad badania kompletności atrybutów, w sensie ich wypełnienia informacją. I chociaż wiedza o kompletności wartości atrybutów i istnieniu potencjalnych, uzasadnionych braków nierzadko jest pożądana przez użytkowników baz danych przestrzennych, to nie zastępuje dokładności położenia, dokładności tematycznej, aktualności lub kompletności obiektów.

Uzyskane wyniki, dotyczące zarówno dokładności geometrycznej, jak i kompletności potwierdzają ogólne wnioski wynikające z analiz dokładności danych OSM w innych krajach mówiące, że na terenie miast dane są dokładniejsze i pełniejsze. Z kolei zróżnicowanie dokładności położenia budynków w OSM wynika przede wszystkim ze znacząco różnych technologii pozyskiwania danych, na przykład kameralna wektoryzacja obrysu budynku na ortofotomapie lub szkic z ulicy, a także z wykorzystania odbiorników GPS różnych klas.

Bardzo wysoka dokładność położenia budynków zarówno w mieście Siedlce (0,49 m) i trzy razy mniejsza – lecz również wysoka – dokładność na terenie powiatu (1,23 m) w zasadzie przekracza dokładność popularnych odbiorników GPS oraz dokładność wektoryzacji na ogólnodostępnych ortofotomapach. Może to wskazywać na pozyskanie części danych w formie cyfrowej z baz danych o wysokiej szczegółowości i dokładności. Wniosek ten potwierdziła analiza metod pozyskania danych o budynkach z terenu Siedlec, z której wynika, że w 2011 roku za pozwoleniem Starosty Powiatu Siedleckiego nastąpiło zaimportowanie części danych o budynkach zgromadzonych w państwowym zasobie geodezyjnym i kartograficznym w Siedlcach. Dane w postaci pliku DXF zostały przekonwertowane autorskim skryptem do postaci akceptowalnej przez OSM, z zastosowaniem kameralnej edycji (bez kontroli terenowej) oparte o doświadczenie autora skryptu, a zarazem członka społeczności OSM (Load building footprints (not outlines\!) in Siedlce based on the Siedlce County's official data, some editing, and guessing of the data format details). Wyniki analizy dokładnościowej potwierdzają także hipotezę o lepszej jakości danych dla obszarów zabudowanych niż rolno-leśnych.

Poprawna identyfikacja i obrys budynku na ortoobrazie pozwala zaliczyć je, za Goodchildem (2007) do obiektów łatwo rozpoznawalnych. Problem pojawia się jednak z określeniem funkcji budynku, która tylko w nielicznych przypadkach jest możliwa do jednoznacznego określenia na podstawie obrazu satelitarnego lub zdjęcia lotniczego. Podobnie nie łatwo jest określić funkcję budynku obserwując go z zewnątrz w terenie (wysoki wieżowiec może być apartamentowcem lub biurowcem lub mieścić w sobie muzeum sztuki nowoczesnej). Dodatkowo w różnych krajach, różne tradycje budowlane uwarunkowane historycznie oraz prawnie mogą zaburzać ocenę wizualną, szczególne nielokalnemu obserwatorowi. Zdaniem autorów, to są powody braków atrybutowych w budynkach bazy OpenStreetMap.

Mimo niedogodności jakościowych, baza danych OSM, włącznie z danymi o budynkach i budowlach, ma wielki i ciągle rosnący potencjał informacyjny. W szczególności w krajach o słabym pokryciu mapowym na poziomie krajowym (ascetyczne pokrycie mapowe, przestarzałe dane, itp.) jak na przykład Brazylia, z danymi z bazy OSM wiąże się wielkie oczekiwania, nawet jako źródłem oficjalnych danych przestrzennych (Camboim, 2015). Według Westrope i innych (2014) cała infrastruktura OSM, do której zaliczylibyśmy wolontariuszy-dostarczycieli danych, użytkowników, narzędzia i podejścia zaprojektowane dla wspomagania centralnej bazy danych i mapy OSM, i dane, jest silnym, prężnym i łatwo adaptowalnym organizmem. Można się spodziewać, że wytyczne dla zapewnienia wysokiej jakości danych w OSM ulegną doprecyzowaniu i unormowaniu. Niemniej powstaje pytanie czy wzmożona formalizacja, wymagająca większego nakładu czasu i intelektu wolontariuszy (dla zapoznania się, zrozumienia i wdrożenia wytycznych), nie odbije się negatywnie na zaangażowaniu wolontariuszy.

Literatura

Al.-Bakri M., Fairbairn D., 2010: Assessing the accuracy of "crowdsourced" data and its integration with official spatial data sets. Accuracy 2010 Symposium, July 20-23, Leicester, UK.

Arsanjani J-J., Zipf A., Mooney P., Helbich M. (eds), 2015: OpenStreetMap in GIScience: Experiences, Research, and Applications. Lecture Notes in Geoinformation and Cartography.

Bielecka E., 2010: Zasady oceny jakości danych przestrzennych. Ocena jakości danych gromadzonych w TBD. *Roczniki Geomatyki* t. 8, z. 4(40): 53-66, PTIP, Warszawa.

Bielecka E., 2015: Geographical data sets fitness of use evaluation. *Geodetski Vestnik* vol. 59, No. 2: 335-348, DOI: 10.15292/geodetski-vestnik.2015.02.335-348.

Cichociński P., 2012: Ocena przydatności OpenStreetMap jako źródła danych dla analiz sieciowych. *Roczniki Geomatyki* t. 10, z. 7: 15-24, PTIP, Warszawa.

Esmail R., Naeseri F., Esmail A., 2013: Quality assessment of Volunteered Geographic Information. *American Journal for Geographic Information System* vol. 2(2):19-26. DOI: 10.5923/j.ajgis.20130202.01.

Goodchild M.F., 2007: Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research* vol.2: 24-32.

Goodchild M.F., 2008: Spatial Accuracy 2.0. 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences.

Goodchild M.F., Glennon J.A., 2010: Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth* 3(3): 231-241.

Goodchild M.F., Li L., 2012: Assuring the quality of Volunteered Geographic Information. *Spatial statistics* 1:110-120.

Haklay M., Weber P., 2008: OpenStreetMap – User-generated Street Map. *IEEE Pervasive Computing* vol. 7: 12-18.

Haklay M., 2010: How good is volunteered geographical information? Comparative study of OpenStreetMap and Ordnance Survey Dataset. *Environmental&Planning B: Planning and Design* vol. 37 (4): 682-703.

Hecht R., Kunze C., Hahmann S., 2013: Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. ISPRS *International Journal of Geo-Information* vol. 2: 1066-1091.

ISO 19 157:2003 – informacja geograficzna – jakość danych.

Loshin D., 2010: Monitoring Data Quality Performance: Using Data Quality Metrics. White Paper Informatica. http://it.ojp.gov/docdownloader.aspx?ddid=999

Ma D., Sandberg M., Jiang B., 2015: Characterizing the Heterogeneity of the OpenStreetMap Data and Community. *ISPRS Int. J. Geo-Inf*. 4: 535-550, DOI:10.3390/ijgi4020535.

MSWiA, 2012: Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 17 listopada 2011 r. w sprawie bazy danych obiektów topograficznych oraz bazy danych obiektów ogólnogeograficznych, a także standardowych opracowań kartograficznych. Dz.U. 2011 nr 279 poz. 1642.

Nowak J., Nogueras-Iso J., Peedell S., 2005: Issues of multilinguality on creating a European SDI – the perspective for spatial data interoperability. [In:] K. Fullerton (ed.), Abstract book of 11th EC-GI&GIS Workshop – ESDI: Setting the Framework, European Commission , DG Joint Research Centre: 47-48.

OSM, 2015a: Tags statistics – OpenStreetMap Wiki. (dostęp 22.08.2015 r.) http://taginfo.openstreetmap.org/tags

OSM, 2015b: Buildings – OpenStreetMap Wiki. (dostęp 22.08.2015 r.) http://wiki.openstreetmap.org/wiki/Buildings

OSM, 2015c. Open Discussion – Key:Building – OpenStreetMap Wiki. (dostęp 22.08.2015 r.) http://wiki.openstreetmap.org/wiki/Talk:Key:building

Camboim S.P., Bravo J.V.M., Sluter C.R., 2015: An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. *ISPRS Int. J. Geo-Inf. (IJGI)* vol. 4: 1366-1388. DOI:10.3390/ijgi4031366.

POKB, 1999: Rozporządzenie Rady Ministrów z dnia 30.12.1999 r. w sprawie Polskiej Klasyfikacji Obiektów Budowlanych (PKOB). Dz.U. 1999, nr 112, poz. 1316 i Dz.U. 2002, nr 18, poz. 170.

Ramm F., Topf J., Chilton S., 2010: OpenStreetMap: Using and Enhancing the Free Map of the World; UIT Cambridge: Cambridge, UK.

Redman T.C. (ed.). 1996: Data Quality for the Information Age. Boston, MA: Artech House.

Westrope C., Banick R., Levine M., 2014: Groundtruthing OpenStreetMap building damage assessment. *Procedia Eng.* vol.78: 29-39.

Xia J., Myers R.L., Wilhiote S.K., 2011: Multiple open access availability and citation impact. *Journal of Information Science* 37 (1): 19-28.

Zielstra D., Hochmair H.H., Neis P., 2013: Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transactions in GIS* 17(3): 315-334.

Zielstra D., Zipf A., 2010: A comparative study of proprietary geodata and volunteered geographic information for Germany. [In:] Painho M., Santos M.Y., Pundt H., (eds.), AGILE 2010: 13th AGILE International Conference on Geographic Information Science. Springer Verlag, Guimaraes, Portugal, 2010.

Streszczenie

Jakość danych OpenStreetMap (OSM), a w szczególności takie jej elementy ilościowe jak kompletność oraz dokładność położenia, wzbudza szerokie zainteresowanie naukowców na świecie. W artykule przedstawiono okoliczności powszechnie obserwowanej heterogenicznej charakterystyki OSM, zwracając uwagę na aspekt niedoskonałości ustaleń semantycznych i założeń jakościowych inicjatywy oddolnej jaką jest OpenStreetMap. Część praktyczną badań stanowi ocena kompletności i dokładności lokalizacji danych o budynkach i budowlach OSM w stosunku do krajowych danych urzędowych, bazy danych obiektów topograficznych BDOT10k. Analizy zostały przeprowadzone dla peryferyjnie położonego powiatu siedleckiego i miasta Siedlce. Opracowanie dopełnia dotychczasowe rezultaty badawcze w zakresie analiz ilościowych jakości OSM, a otrzymane wyniki potwierdzają zróżnicowaną jakość danych o budynkach, w sensie ich kompletności i wypełnienia wartościami ich atrybutów oraz dokładności lokalizacji, także na terenie Polski. Niemniej jednak, wyniki analizy dokładności geometrycznej są zaskakująco dobre. W dyskusji autorzy zwracają uwagę na fakt, że mimo niedoskonałości danych wolnych i otwartych są one powszechnie wykorzystywane przez użytkowników.

Abstract

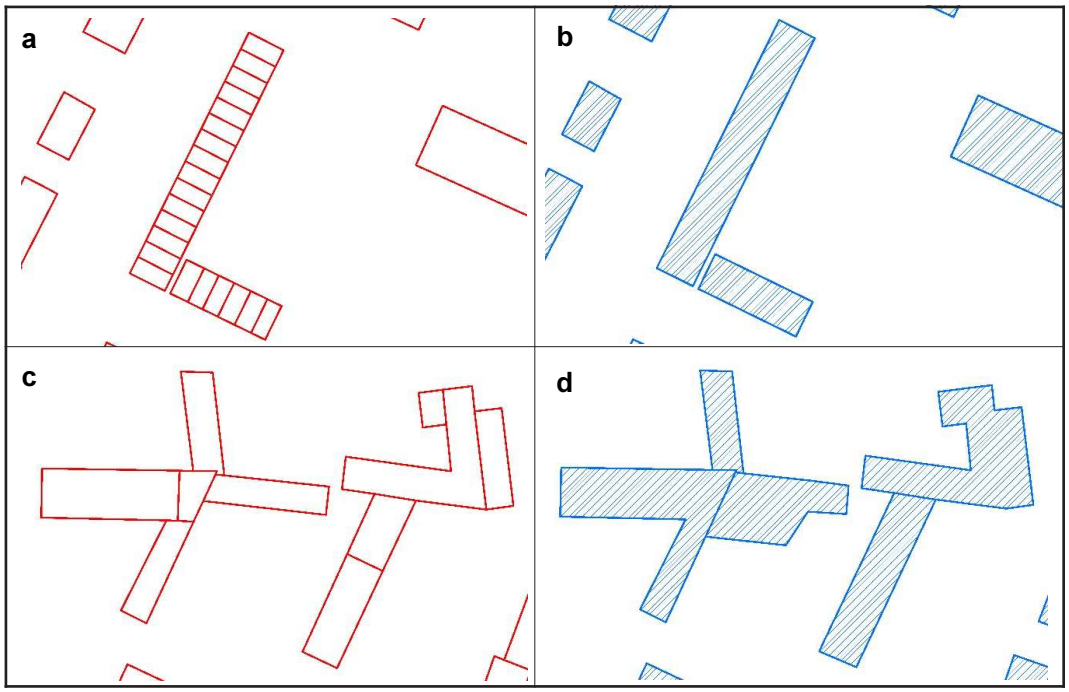
Researchers all over the world are interested in OpenStreetMap data and its quality including completeness and geometric accuracy. This article looks into the commonly observed heterogeneous characteristics of OpenStreetMap geospatial data and draws attention to the vague semantic and quality foundations of this important grass-roots initiative. The experiment is an assessment of the completeness and positional accuracy of OSM building data compared to the national data: the Database of Topographic Objects in Poland (BDOT10k). The analysis was performed for the county and city of Siedlce. This study complements previous research results in the quantitative analysis of OpenStreetMap data quality. The results confirm the variable quality of OSM data in terms of completeness and updating of building information found in their attribute's, and the positional accuracy of building corners even for the Polish territory. Nevertheless, the analysis did find that the positional accuracy of the OpenStreetMap building data was very good in comparison to the BDOT10K database. The authors draw attention to the fact that Free and Open geospatial data, despite its imperfections, is widely adopted by users including public administrations.

dr inż. Joanna Nowak Da Costa
joanna.nowakdc@wat.edu.pl

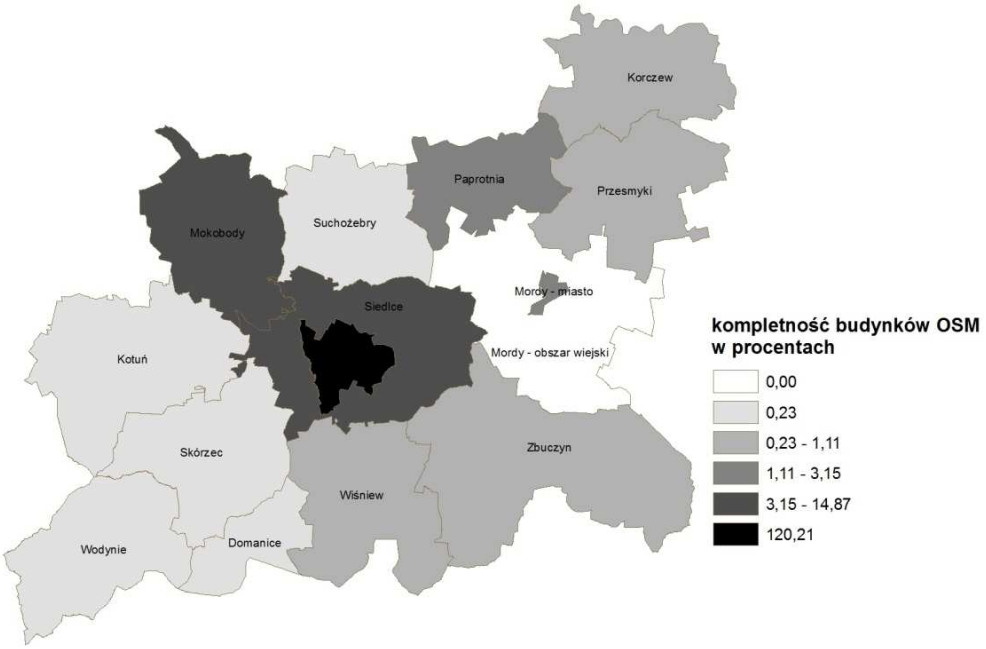
dr hab. inż. Elżbieta Bielecka, prof. WAT
elzbieta.bielecka@wat.edu.pl

dr inż. Beata Całka
beata.calka@wat.edu.pl

Rysunek 1. Błędna identyfikacja i lokalizacja budynków w bazie OSM (kolor czerwony); kolorem zielonym zaznaczono budynki pochodzące z BDOT10k



Rysunek 2. Różny sposób modelowania obiektów: a, c – w OSM, b, d – w BDOT10k; górna para obrazów ilustruje garaże o przyległych ścianach, dolna para – budynki o zróżnicowanej wysokości i zadaszeniu



Rysunek 3. Kompletność budynków OSM w stosunku do BDOT10k