WPŁYW JAKOŚCI DANYCH MAPOWYCH NA UŻYTECZNOŚĆ TELEINFORMATYCZNYCH SYSTEMÓW INFORMACJI LOGISTYCZNEJ

IMPACT OF MAP DATA QUALITY ON THE USABILITY OF ICT LOGISTICS INFORMATION SYSTEMS

Ewa KALBARCZYK-GUZEK
ewa.kalbarczyk@wat.edu.pl
https://orcid.org/0000-0001-5948-8483

Wojskowa Akademia Techniczna
Wydział Bezpieczeństwa, Logistyki i Zarządzania
Instytut Logistyki

Streszczenie: *W artykule omówiono zagrożenia jakie niesie za sobą nieprawidłowa jakość danych mapowych w organizacjach zajmujących się logistyką w kontekście procesów jakie obsługują. Zwrócono uwagę na problematykę błędnych danych adresowych, zarówno na poziomie strategicznym jak i operacyjnym. Wyszczególniono problemy, z jakimi borykają się firmy logistyczne posiadające bazy z błędnie zgeokodowanymi kontrahentami. Opisano dostępne na rynku metody czyszczenia danych oferowane przez wyspecjalizowane przedsiębiorstwa rynkowe.*

Abstract: *The article discusses the risks of incorrect map data quality in logistics organizations in the context of the processes they support. Attention was paid to the problem of erroneous address data, both at the strategic and operational levels. Listed are the problems faced by logistics companies with databases with incorrectly geocoded contractors. Data cleaning methods available on the market, offered by specialized market companies, have been described.*

Słowa kluczowe: *geolokalizacja, łańcuch dostaw, czyszczenie danych, lokalizacja centrum dystrybucji*

Keywords: *geolocation, supply chain, data cleaning, distribution center location*

WSTĘP

Problemy związane z lokalizacją oraz przetwarzaniem danych mapowych znane są w literaturze bez mała od trzech stuleci. Prekursorem teorii lokalizacji był J.H. Thunen, który określił najbardziej opłacalny ekonomicznie układ stref rolniczych wokół miasta, będącego dla produktów rolniczych rynkiem zbytu. W tym celu stworzył graficzny model przedstawiający rozmieszczenie różnych typów gospodarki rolnej wokół centralnie usytuowanego rynku (Dziekoński, Matusewicz, 2014).

Współczesne przedsiębiorstwa z branży logistycznej, niezależnie od swojej struktury organizacyjnej, działają w rzeczywistości nasyconej ogromną liczbą danych, które są gromadzone i przetwarzane w wykorzystywanych systemach informatycznych (Grzelak, Zdunek, 2017). Dane adresowe służą do codziennej pracy operacyjnej, a także stanowią nieodzowny filar do podejmowania decyzji o charakterze strategicznym, mającym wpływ na dalszy kierunek rozwoju przedsiębiorstwa, kształtowania jego pozycji konkurencyjnej i świadczonej ofercie rynkowej.

W przypadku pojawiania się sytuacji nieprzewidzianych stanowią podstawę do skutecznego reagowania na ich siłę oddziaływania i wprowadzania planów awaryjnych. Dlatego też na rynku, wśród firm z sektora logistycznego istnieje potrzeba zapewnienia najwyższej jakości przetwarzania danych adresowych i ich wizualizacji przy pomocy map elektronicznych.

Celem artykułu jest przedstawienie zagrożeń, jakie niesie za sobą użycie złych danych mapowych w narzędziach informatycznych wspierających logistykę, zarówno na poziomie strategicznym, jak i operacyjnym oraz zaprezentowanie możliwości przeciwdziałania takim sytuacjom. Odniesienie do poziomu strategicznego zostało przedstawione w kontekście określenia i wizualizacji na mapach elektronicznych lokalizacji nowych centrów logistycznych, podziale obszaru na regiony oraz przyporządkowania do nich klientów i zasobów (w rozumieniu pracowników oraz sprzętu, maszyn). Poziom operacyjny oscylował wokół działań służących do optymalizacji, planowania i rozliczania tras pojazdów dystrybucyjnych.

Do realizacji niniejszego celu posłużono się następującymi metodami i narzędziami badawczymi: analizą dostępnej literatury, prezentacją badania przeprowadzonego przez dr Roberta Wojtachnika, polegającego na uzyskaniu automatycznej geolokalizacji dla bazy 31 000 kontrahentów oraz przeprowadzonymi rozmowami z przedsiębiorcami branży logistycznej, którzy ze względu na rodzaj prowadzonej działalności zetknęli się z koniecznością wykonania czyszczenia baz adresowych i potwierdzili, iż nieuporządkowanie w tym aspekcie wpływa na obniżenie jakości świadczonych usług oraz dewaluację wyniku finansowego osiąganego przez firmy.

1. ZASTOSOWANIE MAP CYFROWYCH W NOWOCZESNYCH NARZĘDZIACH INFORMATYCZNYCH DLA LOGISTYKI

W nowoczesnych systemach dedykowanych logistyce dane adresowe nieodzownie związane są z procesem realizacji dostaw. Wizualizacja danych na mapach wymaga wskazania w procesie geolokalizacji dla każdego adresu współrzędnych geograficznych (długości i szerokości na kuli ziemskiej). Dokładność tego procesu wywiera istotny wpływ na poprawność prezentowania danych na mapie, precyzję uzyskiwanych obliczeń oraz podejmowanie optymalnych decyzji przez planistów transportu.

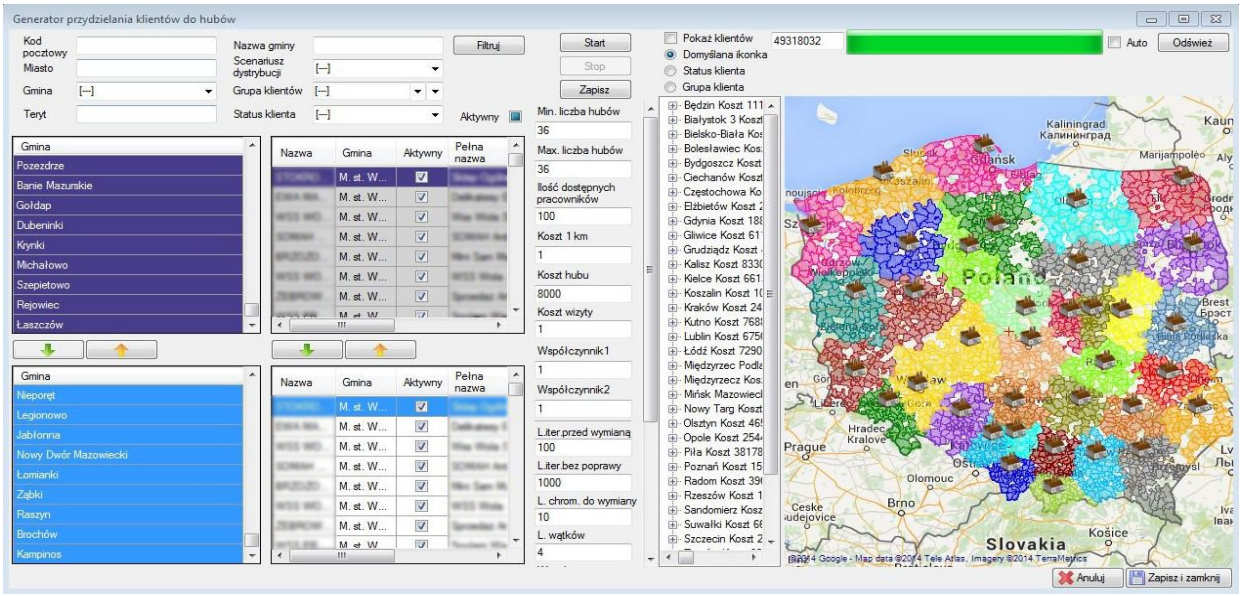
Na poziomie strategicznym w informatycznych systemach klasy TMS (ang. Transport Management Systems) prawidłowo wpisany, a potem naniesiony na mapę zgeokodowany adres, posiadający współrzędne na kuli ziemskiej, stanowi podstawę do określenia lokalizacji centrów logistycznych i podziału wybranego obszaru na regiony (rys.1).

Rys.1. Podział obszaru Polski na regiony, wyznaczenie położenia centrów logistycznych
Źródło: system TMS.

Podczas wyznaczania granic poszczególnych regionów i lokalizacji centrów logistycznych działają skomplikowane, wielokryterialne algorytmy obliczeniowe, uwzględniające:

- wielkość transportowanych strumieni towarowych;
- strukturę populacji;
- położenie geograficzne obecnych klientów;
- korytarze transportowe;
- częstotliwość dostaw do klientów;
- umiejscowienie obiektów przedsiębiorstwa, które optymalizuje wydajność łańcucha wartości (Vos, Akkermans, 1996);
- odległość między lokalizacją każdego kontrahenta w obszarze, a centrum logistycznym;
- występowanie w okolicy innych centrów logistycznych.

Po przeprowadzeniu podziału na regiony uruchamiany jest algorytm, który zajmuje się przydzieleniem klientów do poszczególnych regionów i centrów logistycznych. Wizualizacja wyniku przydziału klientów do regionów zobrazowana została na rysunku 2.



Rys. 2. Generator przydzielania klientów
Źródło: system TMS.

Problematyka lokalizacji i wizualizacji centrów dystrybucji doczekała się wielu opracowań teoretycznych wraz z towarzyszącymi im rozwiązaniami informatycznymi (Mirchandani, Francis, 1990), które cieszą się obecnie niesłabnącą popularnością. Zaprezentowana powyżej wizualizacja jest jedną z nich. Rozważania na temat lokalizacji nowego centrum logistycznego to zagadnienie, przed którym, prędzej czy później, stanie każdy rozwijający się operator logistyczny, jest jednym z najważniejszych elementów udanej inwestycji i powodzenia całego przedsięwzięcia gospodarczego (Walerjańczyk, 2010). Określając obiekt mianem centrum logistycznego należy mieć na uwadze wszelkiego rodzaju usługi związane z magazynowaniem, przemieszczaniem się towaru od nadawcy do odbiorcy oraz infrastrukturę umożliwiającą wykonywanie tych operacji (co składa się na centrum magazynowe), ale również przewozy intermodalne i dodatkowe usługi oferowane niezależnym przedsiębiorstwom (Kasperska-Moroń, Krzyżaniak, 2009). Według Lalonde i Delaney, koszty transportu związane z przepływem towarów od miejsc produkcji do miejsca przeznaczenia stanowią około 50% całkowitych kosztów w łańcuchu logistycznym, zaś około 30% całkowitych kosztów logistycznych generują bezpośrednie koszty magazynowania (LaLonde, Pohlen, 1996). Wynika z tego, jak istotną rolę pełnią prawidłowe dane adresowe i właściwa ich wizualizacja na mapie.

Mimo różnic w podejściu do wyznaczania nowej lokalizacji, podstawowe założenia modeli lokalizacji zawsze obejmują w kontekście analiz: przestrzeń, klientów i producentów, których lokalizacje w danej przestrzeni są znane i obiektów, których lokalizacje muszą być ustalone według określonej funkcji celu. Zwieńczeniem procesu poszukiwania nowej lokalizacji jest wizualizacja osiągniętego wyniku na mapie.

Na poziomie operacyjnym firmy zajmujące się dystrybucją towarów, w swojej codziennej pracy wykorzystują systemy klasy TMS, które oferują wielowymiarową wizualizację danych

i wyznaczanie tras przy zastosowaniu cyfrowych map dla biznesu. Możliwość prezentacji danych na mapie wiąże się z koniecznością wskazania współrzędnych geograficznych podczas procesu geolokalizacji. Prawidłowa geolokalizacja adresu stanowi o bezpieczeństwie realizowanego procesu planistycznego w następujących obszarach:

- możliwości realizacji trasy (wykorzystanie właściwie dobranych tonażowo pojazdów do realizacji trasy);
- terminowości realizowanych dostaw;
- możliwości przeprowadzenia globalnego planowania i optymalizacji dużej ilości tras, z wyznaczeniem kolejności odwiedzanych punktów;
- obliczaniu długości trasy;
- obliczaniu kosztów i przychodów wynikających z realizacji trasy;
- wyznaczaniu stref zakazanych, czyli miejsc do których kierowcy nie mogą jeździć.

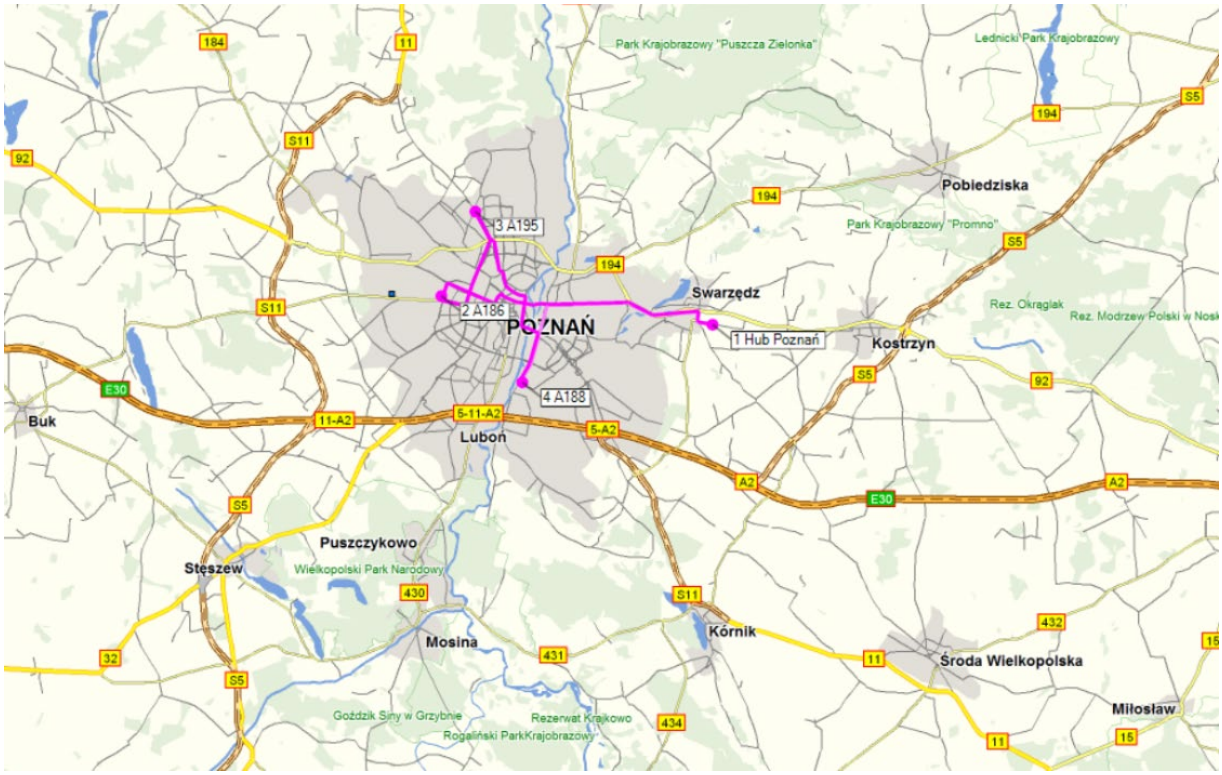
Mapa cyfrowa wykorzystywana w biznesie jest mapą elektroniczną, której działanie opiera się na połączeniu elementów graficznych z przypisanymi im w formie elektronicznej informacjami. Bazuje ona na zebranych i przetworzonych do postaci cyfrowej danych kartograficznych (http://emapa.pl/mapy-cyfrowe/mapy-cyfrowe-1). Przewagi map cyfrowych nad mapami drukowanymi są następujące:

- mapy cyfrowe obejmują dowolny obszar;
- są skalowalne – można je dowolnie przybliżać i oddalać;
- są dokładniejsze - nie ma map drukowanych w takiej skali, jak największe przybliżenia map cyfrowych;
- są aktualniejsze - mogą być zdalnie aktualizowane przy znacznie mniejszych nakładach kosztów i pracy.

Mapy cyfrowe współpracują z nowoczesnymi systemami wspierającymi logistykę (przykład danych mapowych na rysunku 3). Są obecne w modułach:

- do planowania i optymalizacji, gdzie wizualizowane są zaplanowane przejazdy, kontrahenci oraz zlecenia do realizacji, trasy są na bieżąco planowane i korygowane, wyliczane są odległości między punktami;
- do monitorowania realizacji tras, gdzie na mapach widoczne są pojazdy dystrybuujące towar wraz z informacją o ewentualnych spóźnieniach;
- do monitorowania nieprawidłowych zachowań kierowców, np. wjazdów samochodów do stref zakazanych, czyli miejsc w których odbywa się handel paletami, sprzedaż nadwyżek paliwowych itp.;

- do kontroli wykonania tras, gdzie porównywane są plany tras wraz z ich rzeczywistym wykonaniem;
- do rozliczeń z przewoźnikami i klientami, gdzie na podstawie informacji o pokonanym dystansie przygotowywane są zestawienia do faktur;
- do naliczania wynagrodzeń kierowcom, jeśli otrzymują premie za ilość przejechanych kilometrów;
- do wyznaczania lokalizacji centrów dystrybucyjnych.



Rys.3. Przykład danych mapowych z wizualizacją zaplanowanej trasy
Źródło: System TMS.

Mapa drogowa Polski w postaci elektronicznej zawiera bazę dróg, w skład której wchodzi ponad 700 tys. km dróg wraz z atrybutami bazodanowymi:

- ID - unikalny identyfikator odcinka;
- nazwa ulicy - Oficjalna nazwa ulicy;
- numer międzynarodowy - międzynarodowy numer drogi na odcinku;
- numer drogi - numer drogi na odcinku (krajowy lub wojewódzki);
- kategoria - kategoryzacja dróg;
- typ - rodzaj odcinka drogowego: autostrada, jednojezdniówka, dwujezdniówka itp.;
- podtyp - rodzaj odcinka w zależności od poziomu drogi (wiadukt, most, tunel);
- rodzaj nawierzchni - rodzaj nawierzchni na odcinku (grunt, asfalt);
- kierunek - informacja o kierunkowości na odcinku.

Mapy cyfrowe przystosowane są do integracji z dowolnymi systemami informatycznymi, umożliwiając wykonywanie na nich szeregu operacji: nanoszenie obiektów, wyszukiwanie ich na mapie, wyszukiwanie informacji o obiektach, wytyczanie tras do wskazanych punktów czy wyszukiwanie najbliższych obiektów.

2. PROBLEMY FIRM WYNIKAJĄCE Z POSIADANIA BŁĘDNYCH DANYCH ADRESOWYCH

Współczesne firmy działające nie tylko w branży logistycznej stają przed wyzwaniem utrzymywania coraz większej ilości danych dotyczących swoich klientów. Dane te podlegają systematycznym przeglądom oraz, w razie konieczności, korektom. W celu poprawnego przeprowadzenia procesu geokodowania i prezentacji danych na mapie należy wskazać dla każdego z adresów współrzędne geograficzne (długość i szerokość geograficzną), wynikające z prawidłowo zapisanych danych adresowych. Dane adresowe często zawierają błędy w postaci zdublowanych lub niekompletnych wpisów, przez co są niespójne. Najczęstszymi generatorami tego stanu rzeczy są:

- niepoprawnie przypisywane kody pocztowe do miast;
- problemy z niewłaściwym sposobem zapisu adresów mniejszych miejscowości, gdzie dla prawidłowego zadziałania algorytmu geokodującego, w pole ulica wystarczy wpisać nazwę miasta;
- błędnie wpisywane nazwy miejscowości np. Poznań / poznań/ Poznan, wprowadzane ręcznie przez użytkowników lub importowane z systemów zewnętrznych;
- wpisywanie miejscowości w innym języku np. Warsaw (ang.);
- duplikatory rekordów – nie wiadomo, czy to jest ta sama osoba, czy też różne, np.: Jan Kowalski, 00-001 Warszawa oraz Jan Kowalski 00-001 Kraków;
- brak standardu zapisywania danych, np.:

Imię i nazwisko	Kod	Miejscowość	Ulica
Jan Kowalski	00-101	Warszawa	Gwiazdowa 1

lub

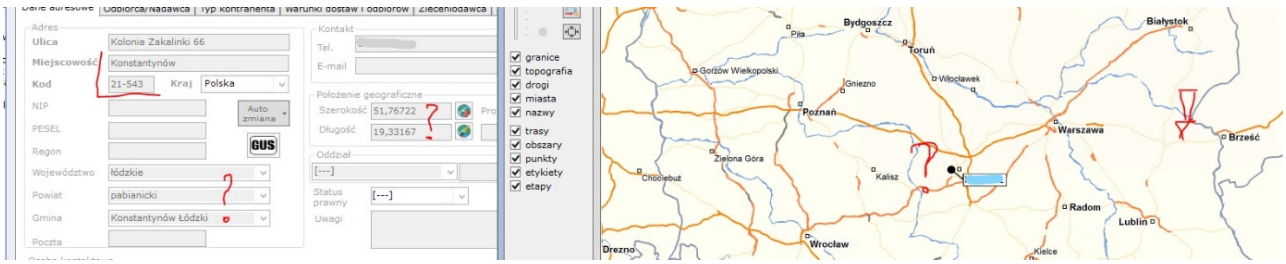
Imię	Nazwisko	Kod	Miejscowość	Ulica
Jan	Kowalski	00101	W-wa	gwiazdowa1

- błędy ludzkie – literówki, małe litery, skróty, złe kody pocztowe, źle przypisane poczty do kodów pocztowych;
- błędne lub brakujące dane na mapie (np. brak nr domu, ulicy);
- nadpisywanie błędnymi danymi danych prawidłowych.

Od jakości posiadanych danych zależy jakość realizowanych procesów w logistyce, w tym planowania tras (odbiory, przerzuty międzymagazynowe, nadania), spływu należności (klient nie płaci za wykonaną usługę, gdyż nie otrzymał na właściwy adres faktury), efektywności działań marketingowych (przekazywanie pod właściwe adresy materiałów i ulotek reklamowych), planowania odwiedzin handlowców u swoich klientów oraz świadczenia usług serwisowych w terenie.

Statystyki podane przez Halo Business Intelligence są w tym aspekcie zatrważające i wskazują, że aż 92% badanych firm przyznaje, że posiadane przez nich dane teleadresowe są niedokładne, zaś 66% badanych organizacji wierzy, że niepoprawne dane mają negatywny wpływ na ich działalność (https://halobi.com/blog/infographic-data-quality-in-bi-the-costs-and-benefits/). Firmy Lemonly.com i Software AG przeprowadziły obliczenia, na podstawie których określono, że koszt biznesowy wynikający z niskiej jakości danych może sięgać nawet 10%-25% przychodów firm (https://lemonly.com/work/the-cost-of-bad-data). Raporty amerykańskiego Data Warehousing Institute dowodzą, iż problemy z jakością danych w USA kosztują przedsiębiorstwa 600 mld USD rocznie. Koszty czyszczenia danych mogą stanowić nawet 80% budżetu przeznaczonego na wdrożenie hurtowni danych. Ponad 50 % projektów CRM (ang. Customer Relationship Managment) zakończyło się niepowodzeniem z powodu złej jakości danych.

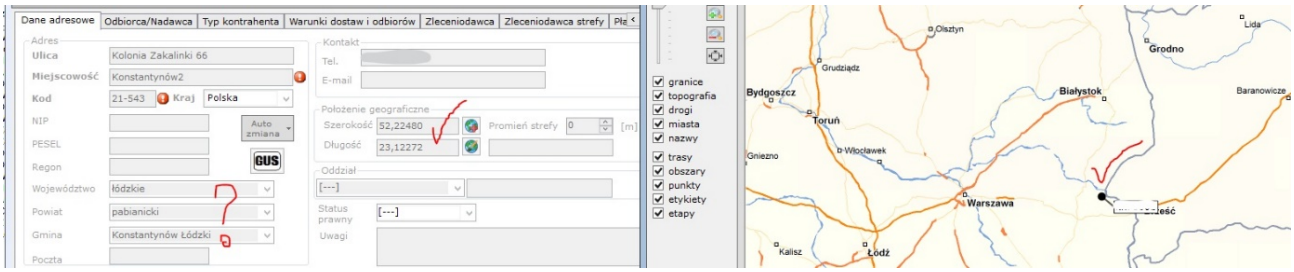
Oprócz błędów w danych adresowych, zdarzają się również usterki samego silnika geokodującego adresy na mapach elektronicznych, co oznacza, że punkty wyświetlane są w nieprawidłowych miejscach. Przykładem takiej sytuacji jest wskazany na rysunku 4 adres: Kolonia Zakalinki 66, 21-543 i miasto Konstantynów, które położone jest w województwie lubelskim, mapa natomiast wykonuje geolokalizację na Konstantynów w województwie łódzkim, nie biorąc w ogóle pod uwagę kodu pocztowego.



Rys. 4. Błędna geolokalizacja danych na mapie, pomimo podania dobrych danych adresowych
Źródło: System TMS.

Dla silnika geokodującego najważniejszym determinantem w ustaleniu współrzędnych powinien być kod pocztowy i miejscowość, zaś w dalszej części ulica. Niestety we wspomnianym przypadku główny determinant silnika mapowego – kod pocztowy – nie został wzięty pod uwagę i punkt został naniesiony na mapie na podstawie miejscowości Konstantynów, zlokalizowanej w województwie łódzkim, co było fałszem. Wszelkie rozliczenia z klientem, które opierały się o odległości drogowe

były błędne i wymagały przeprowadzenia skomplikowanych korekt. Tego typu przypadki narażają firmy logistyczne na straty. Poniżej, na rysunku 5 zaprezentowano prawidłowe zadziałanie silnika geolokalizacyjnego. Nawet w przypadku, gdy wpisana miejscowość jest błędna Konstantynów 2, system prawidłowo określił położenie geolokalizacyjne kontrahenta z Konstantynowa, biorąc pod uwagę jako nadrzędny determinant kod pocztowy.



Rys. 5. Poprawna geolokalizacja danych kontrahenta na mapie.
Źródło: System TMS.

Na podstawie kodu pocztowego system prawidłowo umieścił na mapie kontrahenta, natomiast nie zamienił danych kartotekowych związanych z województwem, powiatem i gminą. Nie miało jednak to wpływu na proces geokodowania. Tego rodzaju błędy, zgłaszane do firm mapowych są usuwane na bieżąco, dzięki czemu jakość danych zaszytych na mapie stale się poprawia.

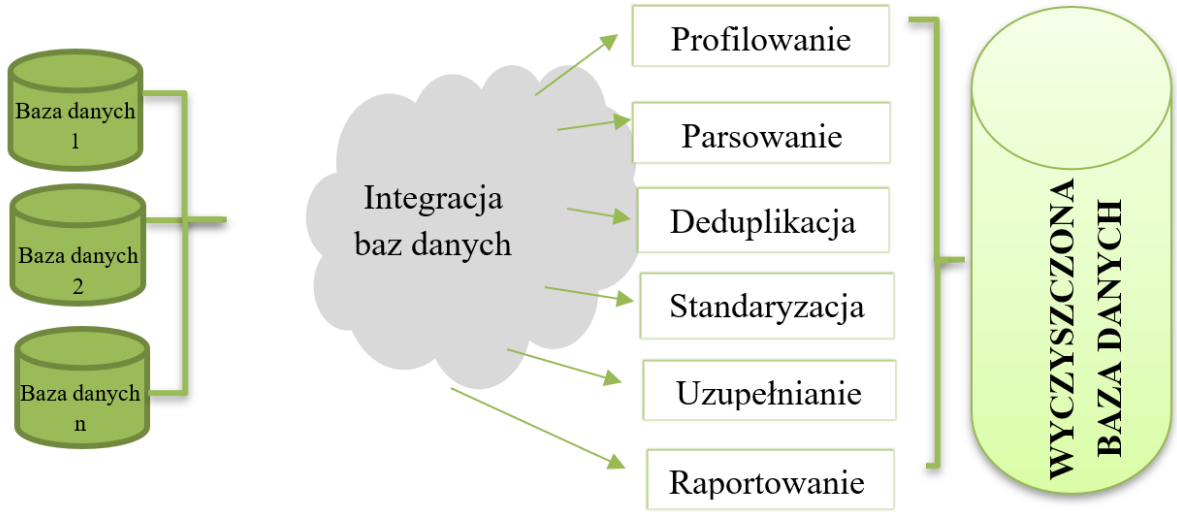
3. METODY POPRAWY JAKOŚCI DANYCH ADRESOWYCH

W związku z negatywnymi konsekwencjami błędnej lokalizacji kontrahentów, na rynku zauważalne jest duże zainteresowanie usługami poprawy jakości baz zawierających dane adresowe. Wyspecjalizowane przedsiębiorstwa świadczą usługi w zakresie czyszczenia danych, realizując następujące etapy:

- profilowanie – czyli podsumowanie stanu danych na dzień wykonywania analizy, identyfikacja źródeł danych, krótkie ich opisanie, wyłowienie oczywistych niespójności i błędów;
- parsowanie - rozbicie jednego złożonego pola na wiele pól w oparciu o znaczenie danych i kontekst, na przykład imię i nazwisko, kod i miejscowość itp.;
- standaryzację - określenie standardów formatowania danych, zamiana wielu różnych wystąpień tej samej wartości zmiennej jedną wartością, utworzenie jednolitego zapisu danych wg przyjętych kryteriów np. „Warszawa” i „W-wa” zostaną zidentyfikowane jako ta sama wartość i zastąpione jedną, zdefiniowaną wartością. W procesie standaryzacji dane podlegające obróbce są łączone ze słownikiem referencyjnym przy pomocy złożonych algorytmów oraz słowników pomocniczych, w których zapisane są spotykane błędne zapisy, bądź alternatywne nazwy (np. w nazwach miejscowości czy ulic w celu ujednolicenia ich zapisu). Algorytm oprócz łączenia takich samych nazw sprawdza także możliwe literówki, czeskie błędy, skróty które mogły być zastosowane (http://standaryzacjadanych.hoga.pl/default.asp);

- deduplikację, która pozwala na wykrycie powtórzonych rekordów, ich konsolidację, a tym samym eliminację zdublowanych danych, np. wyszukiwanie wielokrotnych wpisów tego samego klienta w bazie, nawet gdy dane są zapisane na różne sposoby, łączenie baz z wielu źródeł i ich ujednolicenie polegające na stworzeniu rekordu klienta obejmującego wszystkie informacje z różnych źródeł (warto zaznaczyć, że łączenie informacji wcześniej rozproszonych pozwala na zmniejszenie ilości wymaganej pamięci do przechowywania i przetwarzania danych);
- przygotowanie raportu końcowego;
- wprowadzenie automatyzacji (parsowane, standaryzacja, deduplikacja);
- przeprowadzenie geokodowania i uzupełnianie danych (http://dataquality.pl/na-czym-polega-i-jak-przeprowadzic-projekt-czyszczenia-danych/) – proces uzupełniania danych jest stosowany w przypadku wykrycia niekompletnych danych adresowych i wtedy dane podlegają rozszerzeniu lub uzupełnieniu takimi danymi jak: brakujące kody pocztowe, miejscowości, ulice, numery budynków.

Oferowane przez wspomniane przedsiębiorstwa usługi czyszczenia danych niejednokrotnie stanowią doraźne i jednorazowe wsparcie, jednak pozwalają na zaoszczędzenie czasu i są znacznie tańsze niż korekta ręczna. Jednakże ze względu na fakt, że przedsiębiorstwa rozszerzają współpracę, pozyskując nowych kontrahentów, istnieje potrzeba zaimplementowania mechanizmów wykonujących czyszczenie baz danych nieustannie, automatycznie, w trybie ciągłym dla każdego pozyskanego adresu, by nie zatracić osiągniętego wcześniej efektu (rys.6).



Rys. 6. Proces przeprowadzania czyszczenia baz danych
Źródło: Opracowanie własne na podstawie http://dataquality.pl (dostęp 18.12.2019).

Wprowadzenie takich mechanizmów do systemów, których działanie oparte jest o dane adresowe, powinno stać się standardem.

Jednym z proponowanych rozwiązań, które sprawiają, iż proces czyszczenia danych jest permanentnie rozwijany to wprowadzenie słowników, które podpowiadają możliwe do wyboru dane

adresowe. Przykładem może być formularz, w którym przeprowadza się rejestrację kontrahenta. Może on zawierać podpowiedzi, które po wpisaniu fragmentu danych automatycznie uzupełniają pozostałą część wyrazu. Inną metodą jest zawężenie możliwości wyboru miast po wybraniu kodu pocztowego. Często stosowaną praktyką biznesową jest wprowadzenie właścicieli danych tzw. data stewardów, których rolą jest dbanie o jakość danych, w odniesieniu do kluczowych elementów danych występujących w ramach konkretnej struktury operacyjnej przedsiębiorstwa. Są oni osobami powołanymi do nadzorowania procesu wprowadzania danych (Strauss, Baczyk, Tess, Smits, 2014). Zdobywają informacje o definicji i znaczeniu danych dla właścicieli danych, opracowują i odpowiadają za proces przewarzania danych oraz za audyt jakości i bezpieczeństwa.

4. IMPLEMENTACJA PROCESU GEOLOKALIZACJI Z CZYSZCZENIEM DANYCH - PRZEDSTAWIENIE WYNIKU BADANIA

W kontekście naprawy i czyszczenia danych adresowych istniejących w systemach informatycznych firm logistycznych, dr Robert Wojtachnik z Politechniki Warszawskiej przeprowadził interesujące badanie, którego przedmiotem był proces geolokalizacji bazy 31 000 istniejących kontrahentów przedsiębiorstwa, zaś celem wykonanie geolokalizacji adresów z dokładnością do numeru budynku (Wojtachnik, 2016). Umożliwiło to poprawne naniesienie na mapy kontrahentów i zapewniło o prawidłowości rozliczeń finansowych, które były uzależnione od odległości drogowych. W przedmiotowym badaniu postawiono następujące pytania badawcze:

- jakie czynniki wpływają na wynik geolokalizacji,
- czy istnieją metody poprawy danych, które zmniejszą ilość błędów geolokalizacji.

Przeprowadzone badanie przyniosło odpowiedź, iż najczęstszymi źródłami błędów w danych adresowych okazały się:

- błędy ludzkie,
- import danych bez polskich znaków z systemów zewnętrznych,
- błędne i brakujące dane na mapie (odpowiedzialność dostawcy mapowego),
- rozbieżność między danymi poczty polskiej a danymi mapowymi (np. w obrębie zakresu kodów pocztowych),
- literówki w danych kartotekowych oraz na mapie,
- nieprawdziwe / pomylone dane,
- zapis małych miejscowości pod kodem pocztowym większych miejscowości,
- nadpisywanie prawidłowych danych błędnymi.

Na podstawie przeprowadzonego badania dr Robert Wojtachnik zaproponował stworzenie nowego algorytmu czyszczenia danych adresowych, który z powodzeniem może być stosowany w praktyce

biznesowej. Algorytm czyszczący podzielono na dwie kategorie. Pierwszy odpowiedzialny był za naprawę błędów w miejscowości i kodzie pocztowym, drugi zaś za naprawę ulicy i numeru budynku. Dzięki przeprowadzeniu takiej naprawy silniki mapowe radziły sobie z nadaniem adresowi współrzędnych geograficznych, odpowiadających stanowi rzeczywistemu. Algorytmy czyszczące uruchamiały się zgodnie z kolejnością odpowiadającą nadanej numeracji:

Błędy dotyczące miejscowości i kodu pocztowego:

1. Zmiana nazw miast na poprawne według kodu pocztowego – w przypadku zaistnienia błędu kodu pocztowego na podstawie bazy miast i kodów pocztowych podmieniane jest miasto w adresie na prawidłowe.
2. W przypadku stwierdzenia niezgodności kodu pocztowego z miastem:
   a) pobranie kodu pocztowego dla miasta wg mapy,
   b) podstawienie 20 pasujących miast według oryginalnego kodu pocztowego.
4. Zamiana nazwy miasta zawierającej ciąg (myślnik między białymi znakami) na pojedynczy myślnik, pojedynczą spację lub usunięcie tego ciągu.
7. Dodanie do nazwy ulicy nazwy miasta.
8. Usunięcie z nazwy ulicy nazwy miasta.

Błędy popełnione w nazwie ulicy i numerze budynku:

3. Zamiana nazwy ulicy na nazwę miasta:
   a) podstawienie adresu ze zmienioną nazwy miasta (oryginalną nazwą ulicy),
   b) pobranie 10 pasujących kodów pocztowych wg oryginalnej nazwy miasta.
5. Zamiana w nazwie ulicy prefixów dotyczących ulicy, placu, alei, osiedla na ul., pl., al., oś.
6. Usunięcie z nazwy ulicy prefixu dotyczącego ulicy, placu, alei, osiedla:
   a) podstawienie adresu ze zmienioną nazwą miasta (nazwa wyciągnięta z nazwy ulicy) i ulicy (z usuniętą nazwą miasta),
   b) pobranie 10 pasujących kodów pocztowych wg nazwy miasta wyciągniętej z nazwy ulicy.
9. Błędna nazwa ulicy (. Lub ul.):
   a) podstawienie adresu bez ulicy.
   b) pobranie 10 pasujących kodów pocztowych wg nazwy miasta z poprzedniej funkcji (wyciągniętej z nazwy ulicy).
10. Usunięcie z nazwy ulicy ciągu po numerze domu.
11. Usunięcie z nazwy ulicy ciągu ul. Oraz ciągu po numerze domu.
12. Pobranie wariantów nazw ulic z tabeli zawierającej takie mapowania (np. Zamiana jana pawła na jp).

14. Zamiana nazwy ulicy zawierającej ciągi 3-go,5-go, i,ii,sl,wlkp i pochodnych tych ciągów na pełne nazwy.

W wyniku przeprowadzonego badania i po zaimplementowaniu procesu geolokalizacji z zastosowaniem czyszczenia danych do systemu TMS osiągnięto cel polegający na zgeokodowaniu 31 000 z dokładnością do numeru budynku w 96,3%. Przed wdrożeniem algorytmu uzyskano wynik 2240 błędnych adresów (7,2% wszystkich adresów). Wdrożenie algorytmu zaowocowało poprawą jakości danych i uzyskaniem 1145 błędnych adresów (3,7% wszystkich adresów). Uzyskano poprawę bezbłędności na poziomie 3,5 p.p. (co stanowi 51,2% wszystkich błędów). Nie udało się zgeolokalizować 3 adresów (0,13% błędnych adresów). Wśród powodów błędnej geolokalizacji 1145 adresów można wskazać błędy mapy oraz błędy w procesie wprowadzania danych co uniemożliwia jednoznacznie stwierdzić o jaki adres chodziło.

Przeprowadzone badanie jest nadzieją na poprawę jakości pracy w działach, które wykorzystują dane mapowe do codziennych operacji, gdyż daje możliwość przewidywania przez system logistyczny geokolalizacji błędnie wprowadzonych adresów, co tym samym automatyzuje proces i eliminuje pracę manualną użytkowników. Jak wynika z badania, proces poprawy jakości danych adresowych w systemach logistycznych powinien być oparty o cztery mechanizmy:

- weryfikację w procesie wprowadzania danych, czyli mechanizmy kontrolujące i podpowiadające użytkownikowi dane adresowe na etapie wprowadzania;
- wdrożenie procesu geolokalizacji z czyszczeniem danych;
- wdrożenie procedur zarządzania jakością danych opartych o zdefiniowanie właścicieli danych i wprowadzenie stanowiska Data Stewartów oraz Chief Dara Officer;
- zbieranie / weryfikowanie współrzędnych adresów z urządzeń mobilnych kierowców.

PODSUMOWANIE

Przeprowadzone analizy miały na celu uzasadnienie potrzeby nieustannej poprawy jakości adresowych baz danych. Zaproponowane metody spełniają kryterium dostępności oraz równej atrakcyjności dla największej liczby potencjalnych klientów w branży logistycznej. Zła jakość danych, ich niekompletność oraz niepoprawność może być przyczyną niechęci użytkowników do korzystania z systemów informatycznych, które wykorzystują tego typu wybrakowane dane. Informacje powstałe na podstawie złych danych prowadzą do wyciągania błędnych wniosków i podejmowania nieefektywnych decyzji biznesowych, prowadzących do strat w przedsiębiorstwie oraz osłabienia i obniżenia pozycji rynkowej przedsiębiorstwa. Po wykonaniu czyszczenia danych należy nieustannie dbać o to, żeby dane nie uległy ponownemu „zanieczyszczeniu". Stąd dobrą praktyką jest wykorzystywanie narzędzi, które mają zaszyte automatyczne mechanizmy czuwające nad zapewnianiem najwyższych standardów jakości danych. Z przeprowadzonych obserwacji

wynika, iż proces poprawy jakości danych adresowych oparty o cztery mechanizmy: weryfikację, wdrożenie procesu geolokalizacji z czyszczeniem danych, wdrożenie procedur zarządzania jakością danych oraz zbieranie / weryfikowanie współrzędnych adresów z urządzeń mobilnych kierowców spełnia swoją rolę. Wdrożenie ich w przedsiębiorstwach przetwarzających dużą ilość danych, wydaje się być koniecznością, gdyż wpływa na uzyskanie przewagi konkurencyjnej i przyczynia się do wzrostu wydajności realizowanych procesów.

LITERATURA:

[1] DZIEKOŃSKI, K., & MATUSEWICZ, E. (2014). Koncepcja lokalizacji nowego centrum logistycznego w Polsce. *Logistyka*. (3). 21-27.

[2] GRZELAK, M., & ZDUNEK, P. (2017). Process Optimization of Order Fulfillment. *Systemy Logistyczne Wojsk*, (46), 68-77.

[3] http://dataquality.pl/na-czym-polega-i-jak-przeprowadzic-projekt-czyszczenia-danych/ (13.12.2019)

[4] http://emapa.pl/mapy-cyfrowe/mapy-cyfrowe-1 (20.12.2019)

[5] http://standaryzacjadanych.hoga.pl/default.asp (13.12.2019)

[6] https://halobi.com/blog/infographic-data-quality-in-bi-the-costs-and-benefits/(20.12.2019)

[7] https://lemonly.com/work/the-cost-of-bad-data (19.12.2019)

[8] KISPERSKA-MOROŃ, D., & KRZYŻANIAK, S. (2009). Logistyka.

[9] LALONDE, B. J., & POHLEN, T. L. (1996). Issues in supply chain costing. *The International Journal of Logistics Management*, *7*(1), 1-12.

[10] MIRCHANDANI, P. B., & FRANCIS, R. L. (1990). *Discrete location theory*.

[11] STRAUSS, D., BACZYK, B., TESS, P. G., & SMITS, J. (2014). How to establish a CDO Office in Your Organization. MIT 2014 CDOIQ Symposium, summary from Session 9D on July.

[12] VOS, B., & AKKERMANS, H. (1996). Capturing the dynamics of facility allocation. *International Journal of Operations & Production Management*.

[13] WALERJAŃCZYK W. (2010). TransComp – XIV International Conference Computer Systems Aided Science, Industry and Transport. Zakopane.

[14] WOJTACHNIK, R. (2016). Metoda poprawy jakości danych adresowych w systemach logistycznych. *Logistyka*, (2), 53-56.