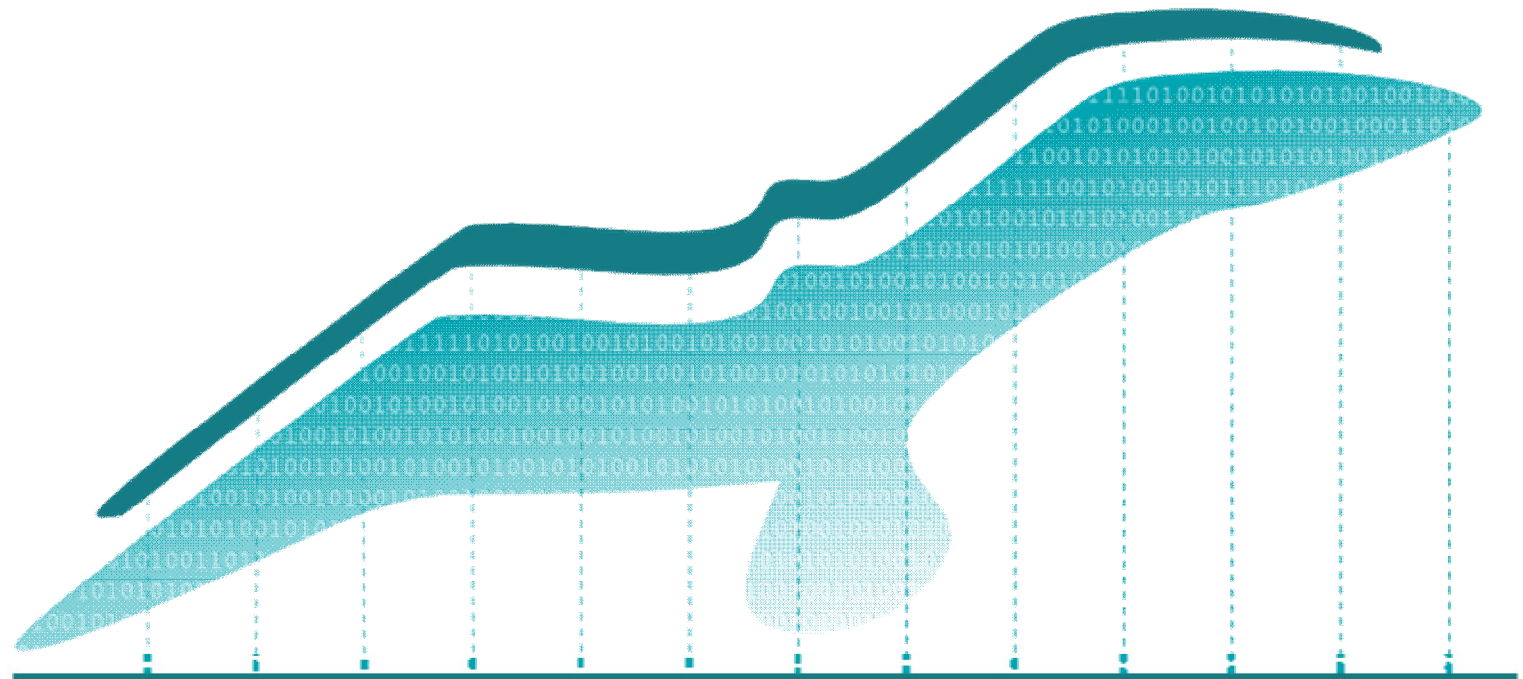


Higiena Danych Adresowych

Rozwiązanie do poprawiania jakości adresowych baz danych

- normalizacja
- standaryzacja
- walidacja
- wzbogacanie





Higiena Danych Adresowych to podnoszenie jakości adresów w bazie danych

| Normalizacja | Standaryzacja |
|--|---|
| <p>Normalizacja adresu oznacza jego podział na pola składowe i obejmuje takie operacje jak:</p> <ul style="list-style-type: none">• oddzielenie kodu pocztowego od nazwy miasta• oddzielenie numeru ulicy od nazwy ulicy• oddzielenie numeru lokalu od numeru budynku• rozdzielenie ulic opisujących róg/skrzyżowanie• oddzielenie dodatkowych informacji opisujących położenie | <p>Standaryzacja oznacza sprowadzenie różnych zapisów tej samej wartości do jednej postaci, zgodnej ze słownikiem. W przypadku adresów standaryzacja obejmuje:</p> <ul style="list-style-type: none">• kody pocztowe• nazwy miejscowości• nazwy ulic (w oparciu o słownik ulic dla danej miejscowości)• numer ulicy (np. 28B zamiast 28 b) |
| Walidacja | Uzupełnianie braków i wzbogacanie |
| <p>Walidacja to sprawdzenie poprawności i logicznej spójności całego adresu. Poszczególne części adresu mogą być poprawne, ale całość nie jest logicznie spójna.</p> <p><i>Przykład: 28-133 Stary Bógpomóż, ul. Gruszek i Jabłuszek 2</i></p> <p>Kod pocztowy dla Pacanowa, miejscowość Stary Bógpomóż, ulica Gruszek i Jabłuszek występują wyłącznie w miejscowości Wólka Kozodawska. Każdy z elementów oddzielnie jest prawidłowy, natomiast złączone w całość tworzą może przyjemny, ale na pewno nie poprawny adres.</p> | <p>Uzupełnianie braków najczęściej dotyczy tych danych, które mogą być odtworzone na podstawie pozostałych elementów:</p> <ul style="list-style-type: none">• kodu pocztowego• nazwy miejscowości <p>Wzbogacanie to przypisywanie do adresu dodatkowych danych, które w adresie nie są niezbędne lub są danymi pochodnymi od adresu. Najczęściej adres jest wzbogacany o następujące dane:</p> <ul style="list-style-type: none">• kod gminy wg rejestru TERYT• symbol nazwy miejscowości wg rejestru TERYT• symbol nazwy ulicy wg rejestru TERTYT• współrzędne geograficzne (długość i szerokość geograficzna) adresu• typ adresu (mieszkalny, biznesowy, mieszkalno-biznesowy) |



Web service Higiena Danych Adresowych

Opracowaliśmy własne rozwiązanie do Higieny Danych Adresowych w trybie całkowicie automatycznym – przy wykorzystaniu web services.

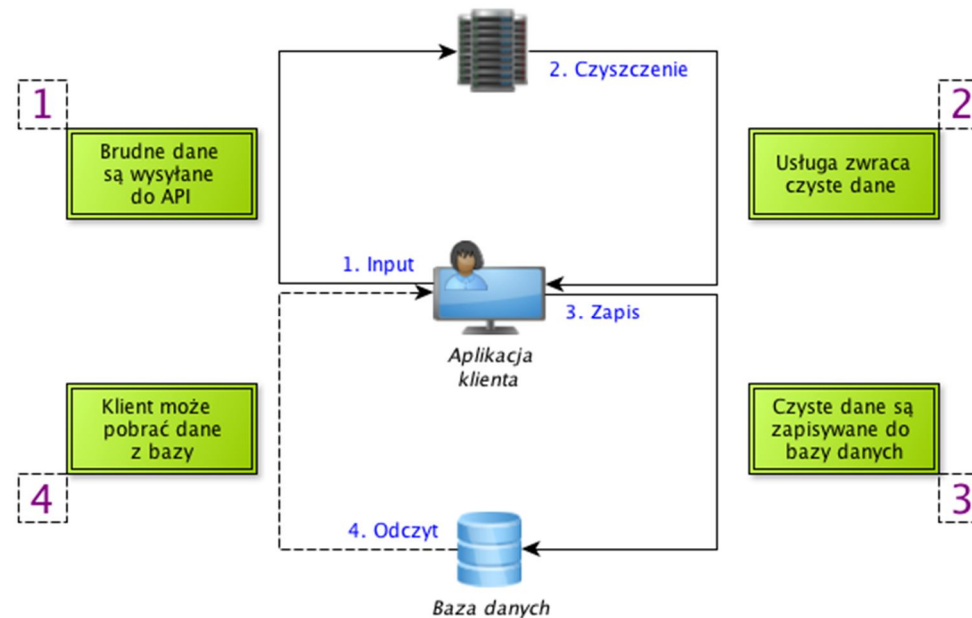
| oznaka niskiej jakości danych | co robi usługa Higieny Danych | przykład |
|---|---|--|
| <ul style="list-style-type: none"> • brak istotnych informacji (np. kodu pocztowego) | <ul style="list-style-type: none"> • uzupełnia kod pocztowy | kod miejscowość GTYNIA 81117 ulica UL. PLK DABKA 21 LOK. 27 (DOM KULTURY) |
| <ul style="list-style-type: none"> • błędy w zapisie nazw miejscowości lub/i nazw ulic | <ul style="list-style-type: none"> • koryguje błędy w nazwach miejscowości lub/i ulic | kod pocztowy 81-107 miejscowość Gdynia typ ulicy ul. nazwa ulicy Dąbka nazwa ulicy GUS płk Stanisława Dąbka |
| <ul style="list-style-type: none"> • wielowariantowość zapisu tego samego adresu | <ul style="list-style-type: none"> • sprowadza różne wersje zapisu do jednej postaci, zgodnej z ustalonym standardem | nr budynku 21 nr lokalu 27 uwagi |
| <ul style="list-style-type: none"> • niespójne kawałki informacji (kod pocztowy odnoszący się do innej miejscowości) | <ul style="list-style-type: none"> • usuwa niespójność poprzez zmianę niepoprawnej wartości | symbol ulicy wg GUS 03654 symbol 0934100 |
| <ul style="list-style-type: none"> • zapisywanie kilku części adresu do jednego pola (braku normalizacji danych) | <ul style="list-style-type: none"> • dzieli adres na elementy składowe i pozwala je zapisać do osobnych pól danych | miejscowości wg GUS kod gminy wg GUS 2262011 współrzędna X 18.537090 współrzędna Y 54.507210 |
| <ul style="list-style-type: none"> • brak dodatkowych informacji, które można uzyskać na podstawie adresu (np. brak nazwy gminy, w której leży dany adres) | <ul style="list-style-type: none"> • dopisuje dodatkowe informacje związane z adresem, np. województwo, powiat, gminę czy współrzędne geograficzne | województwo pomorskie powiat Gdynia gmina Gdynia typ gminy gmina miejska |
| <ul style="list-style-type: none"> • brak wykorzystywania ustalonego standardu zapisu adresu | <ul style="list-style-type: none"> • wprowadza standard zapisu adresu. Standard w usłudze został opracowany przez DataWise. Elastyczność usługi pozwala na zwracanie nazw zarówno wg zapisu DataWise jak również wg zapisu wynikającego z bazy TERYT prowadzonej przez GUS | |



Proponowane rozwiązanie

W pełni automatyczna usługa Higieny Danych Adresowych działająca jako tzw. web service, dostępna na żądanie poprzez sieć informatyczną. Idea działania usługi jest prosta:

- do serwera wysyłany jest adres (dane wejściowe – brudny adres)
- serwer wykonuje proces higieny danych
- w odpowiedzi serwer odsyła adres po procesie higieny danych (zwracane wyniki – czysty adres)





Co jest potrzebne, aby móc korzystać z serwera higieny danych

Usługa działa jako web service, jest to powszechna technologia wymiany danych pomiędzy komputerami, nawet z różnych platform i systemów operacyjnych. Usługa jest dostępna w dwóch najpopularniejszych implementacjach:

- XMLRPC
- SOAP

Korzystanie z usługi jest możliwe wszędzie tam, gdzie technologia (oprogramowanie) umożliwia łączenie się z web service.

W praktyce możliwe jest nawet udostępnienie usługi dla baz danych napisanych w MS Access czy arkuszach kalkulacyjnych MS Excel, i oczywiście w językach programowania (PHP, .NET, VisualStudio, C, itp.);

Cechy funkcjonalne rozwiązania

- działa w czasie rzeczywistym, co oznacza, że wyniki są zwracane w ciągu ułamka sekundy od otrzymania danych wejściowych
- jest dostępna non-stop, przez 24 godziny na dobę
- nie wymaga praktycznie żadnych nakładów na sprzęt (hardware) i oprogramowanie (software)
- możliwe jest zainstalowanie serwera usługi Higieny Danych Adresowych w firmie, co pozwala zwiększyć szybkość i oznacza brak wykorzystywania publicznej sieci Internet do komunikacji klient – serwer (istotne np. w przypadku instytucji finansowych).

Geokodowanie adresów do współrzędnych XY

- web service pozwala na opcjonalne geokodowanie adresów do współrzędnych XY
- geokodowanie adresów wykorzystuje tę samą bazę referencyjną, co AutoMapa



Kiedy i dlaczego stosować

Zastosowania

Higiena Danych Adresowych znajduje zastosowanie w następujących obszarach:

- wprowadzanie nowych danych adresowych do bazy (różnego rodzaju formularze rejestracyjne)
- przy projektach, w których istnieje konieczność integracji różnych baz danych
 - przed wykonywaniem deduplikacji baz danych, aby zwiększyć skuteczność deduplikacji
 - przed wdrażaniem systemów ERP/CRM, aby do systemu były ładowane czyste dane, o wysokiej jakości
- podniesienie jakości adresów w już istniejących bazach danych
- gdy konieczne jest wykorzystywanie informacji o lokalizacji geograficznej (województwo, powiat, gmina, współrzędne geograficzne położenia adresu) aby pokazywać i analizować dane na mapach cyfrowych
- gdy konieczne jest wybieranie z bazy rekordów w jakiejś relacji do podanego adresu (np. identyfikacja najbliższych oddziałów banku dla podanego adresu)

Korzyści

- ułatwione wyszukiwanie i identyfikowanie rekordów
- mniejsze ryzyko powstawania duplikatów w bazie danych ze względu na różne sposoby zapisu adresów
- może być punktem wyjścia do stworzenia jednolitego standardu zapisu danych adresowych w bazie
- pozwala na eliminowanie błędów na etapie wprowadzania danych z formularzy
- uzupełnienie brakujących danych oraz eliminacja błędów, co wprost przekłada się na mniejsze koszty (np. zredukowanie kosztów wysyłania korespondencji pod nieprawidłowe adresy)
- wzbogacenie adresu o dodatkowe dane (np. kod gminy czy współrzędne geograficzne) pozwala wykorzystać te dane do bardzo użytecznych analiz (np. pokazywanie liczby klientów w poszczególnych gminach)
- pozwala na łatwiejszą wymianę informacji pomiędzy różnymi bazami danych (pochodzącymi z różnych systemów)
- zgodność z oficjalnymi rejestrami GUS (TERYT) oraz integracja z kodami pocztowymi Poczty Polskiej



Modele wykorzystywania usługi

| element | dostęp do web service w czasie rzeczywistym | usługa w trybie projektowym | udostępniania środowiska produkcyjnego |
|----------------------------|--|---|--|
| sposób dostarczania usługi | Klient otrzymuje dostęp do web services. Korzystanie z web services odbywa się w czasie rzeczywistym, w momencie kiedy zaistnieje taka potrzeba po stronie Klienta. | Klient przekazuje do DataWise dane w ustalonej postaci (zwykle plików tekstowych). DataWise wykonuje usługę na swoim środowisku produkcyjnym, a następnie zwraca dane w ustalonej postaci (także zwykle plików tekstowych). Procesowanie danych w środowisku produkcyjnym korzysta z tych samych web services. | DataWise udostępnia Klientowi środowisko produkcyjne do wykonywania usługi. Środowisko produkcyjne może być umieszczone na serwerze DataWise (tzw. hostowanie środowiska) lub może być odtworzone na serwerach własnych Klienta (instalacja środowiska). DataWise przeprowadza szkolenie jak korzystać ze środowiska produkcyjnego. |
| sposób rozliczania usługi | Udostępniane konto ma przydzielony określony limit wywołań web service i określony czas na wykorzystanie limitu (zwykle 12 miesięcy). Rozliczanie jest bardzo podobne do tzw. pre-paid w telefonii komórkowej. | Projekt jest indywidualnie wyceniany w zależności od ilości i zakresu danych. | Projekt jest indywidualnie wyceniany w zależności od rodzaju udostępniania (hostowania vs instalacja środowiska). |
| zalety | <ul style="list-style-type: none">• możliwość integracji z systemami informatycznymi• natychmiastowe wyniki (online) | <ul style="list-style-type: none">• minimalny nakład pracy ze strony klienta• bardzo prosty do wykonania• możliwość uwzględnienia dodatkowych, niestandardowych operacji do wykonania (np. własny sposób formatowania danych) | <ul style="list-style-type: none">• polecany sposób do cyklicznych operacji• brak limitów co do liczby procesowanych rekordów |
| wady | <ul style="list-style-type: none">• potrzeba pewnego nakładu pracy na konfigurację systemów informatycznych umożliwiającą łączenie się z naszymi web service | <ul style="list-style-type: none">• nie zapewnia natychmiastowych rezultatów• najdroższy sposób dostępu do usługi | <ul style="list-style-type: none">• wymaga zaangażowania osób ze strony Klienta (pracownik klienta obsługuje środowisko produkcyjne) |



Dodatkowe informacje w Internecie

Dodatkowe informacje o rozwiązaniach DataWise do zarządzania jakością danych można znaleźć na stronie <http://datawise.pl>

Na stronie można skorzystać z formularza pozwalającego na testowanie usługi online <http://datawise.pl/demo/higiena-danych>

Dla użytkowników potrzebujących informacji technicznych, czy przykładów kodu aplikacji pokazujących jak korzystać z usługi przygotowano stronę internetową <http://xmlrpc.higienadanych.pl>

Kontakt

Krzysztof Pędzich

GSM 0-501-725-574

email k.pedzich@datawise.pl

Marek Turlejski

GSM 0-501-099-698

email m.turlejski@datawise.pl