

# PRZEKSZTAŁCENIE ZBIORÓW PUBLICZNYCH W REJESTRY STATYSTYCZNE PRZY POMOCY NARZĘDZI ETL



# Plan prezentacji:

---

1. Przedmiot prac
2. Cel prac
3. Schemat
4. Prace wykonywane w Laboratorium danych (OBM-LAB)
5. Prace wykonywane w środowisku Operacyjnej Bazy Mikrodanych (OBM)



# Przedmiot prac

---

1. Powszechny Spis Rolny
  - 20 rejestrów
  - ponad 700 zmiennych

Narodowy Spis Powszechny Ludność i Mieszkań

25 rejestrów  
ponad 1600 zmiennych



Celem prac jest uzyskanie zbioru danych dostatecznie pełnego pod względem podmiotowym oraz przedmiotowym i jednocześnie odpowiadającego wprowadzonym na podstawie ustaw standardom klasyfikacyjnym, nomenklaturom i definicjom podstawowych kategorii

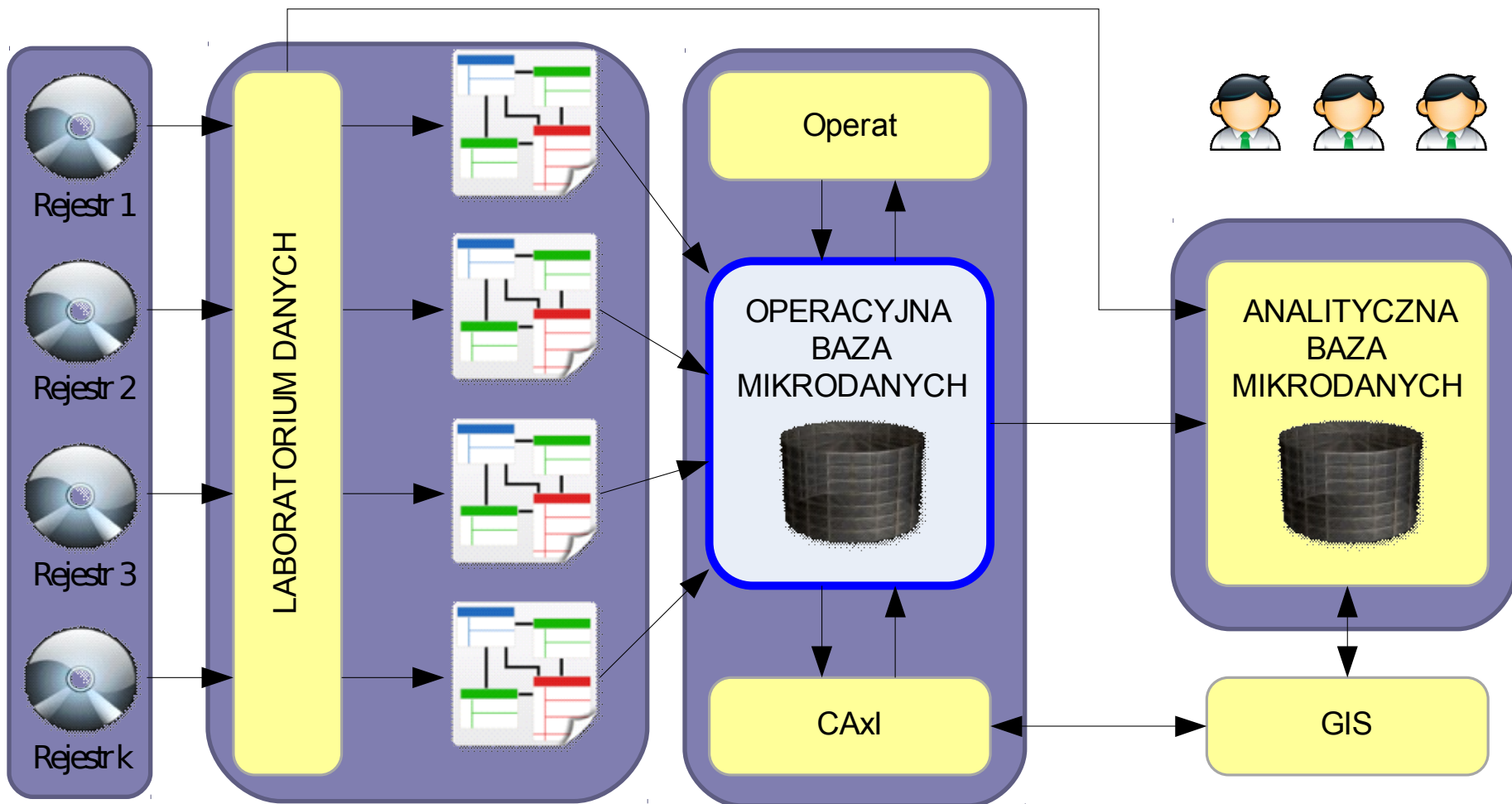
województwo	powiat	gmina	miejsowość	ulica	płeć	obywatelstwo	wykształcenie	miejsce urodzenia
MAZOWQIECKIE	INIWROCLAWS	L?BORK	BIALGARD	KIŁŁĄTAJA	K	KRÓL. NIDERLANDÓW	WYŻSZWE	LONDYN BRIDGE



województwo	powiat	gmina	miejsowość	ulica	płeć	obywatelstwo	wykształcenie	kraj urodzenia
14	1401	1401015	0614506	25438	2	528	1	826



# Schemat



# Laboratorium danych (OBM-LAB)



1. Analiza zakresu podmiotowego i przedmiotowego
2. Czyszczenie danych
3. Kontrola przeprowadzonych procesów
4. Statystyki jakościowe

Województwo	Powiat	Gmina	Miejscowość	Ulica	Obywatelstwo	Miejsce urodzenia
MAZOWQIECKIE	BIAŁOBRZESKI	BIA?OBRZEGI	BEZVZCE	ZAK`TEK	KRÓL. NIDERLANDÓW	LONDYN - ANGLIA
MAZPWOECKIE	BIAŁOBRZ	BAŁOBRZEGI	ZRZE ZCE	ZAKATEK	KRÓLESTWO NIDERLANDÓ	LONDYN – WLK BRYTANIA
ZAZOWIEVCKIE	BIAŁOBRZEGI	BIALBRZEGI	BRZEEZCE	ZATEK	KRÓLESTWO NIDERLANDÓW	LONDYN/CHELSEA
MZAOWIECIE	BIELOBRZEGI	BIALOBAZEGI	BRZEE?CE	ZAK A?TEK	NIDERLANDZKIE	LONDYN BRIDGE



Województwo	Powiat	Gmina	Miejscowość	Cecha	Ulica	Obywatelstwo	Miejsce urodzenia
MAZOWIECKIE	BIAŁOBRZESKI	BIAŁOBRZEGI	BRZEŻCE	UL	ZAKA?TEK	NIDERLANDY	LONDYN



# Czyszczenie danych

---

## Etap oceny jakości danych

Profilowanie - polega na tworzeniu profilu/raportu z jakości posiadanych danych.

1. podstawowe informacje o ilości i jakości danych:
  - liczba wypełnionych rekordów
  - liczba i % pustych rekordów
  - liczba wzorców i unikalnych wpisów
  - frekwencja/częstość występowania wzorców
2. częstość wystąpień wpisów, czyli ilości indywidualnych wpisów.  
**Profil ten sporządza się w celu budowy schematów używanych do czyszczenia zmiennych adresowych.**

Wpis z profilowanej kolumny	Pattern	Przydatność użycia modelu danych
21-150	99-999	Poprawny kod pocztowy
21150kock	9999aaaa	Błędny kod pocztowy



# Czyszczenie danych

---

Etap analizy danych – polegająca na analizie ciągu znaków w celu określenia możliwości rozbicia danego wyrażenia na elementy składowe

Przed parsowaniem	Po parsowaniu	
W kolumnie „ulica” wpisana <b>poprawna</b> nazwa ulicy (zgodnie ze słownikiem TERYT)	Rozpoznanie poprawnego wpisu	
Nazwa ulicy	Przedrostek	Nazwa ulicy
<b>ALEJA SWIERKOWA</b>		<b>ALEJA SWIERKOWA</b>
W kolumnie „ulica” wpisana <b>niepoprawna</b> nazwa ulicy (zgodnie ze słownikiem TERYT)	Rozpoznanie i zapisanie przedrostka do odpowiedniej kolumny	
Nazwa ulicy	Przedrostek	Nazwa ulicy
<b>AL. STANÓW ZJEDNOCZONYCH</b>	<b>ALEJA</b>	<b>STANÓW ZJEDNOCZONYCH</b>

# Czyszczenie danych

Etap czyszczenia schematami - schemat jest to tabela zawierająca dwie kolumny. Z jednej strony znajdują się błędne nazwy, z drugiej odpowiadające im poprawne.

Przykład schematu nazw miejscowości

Rodzaje schematów	Liczba wpisów początkowa	Liczba wpisów końcowa	
SCHEMAT WOJEWÓDZTW		2 225	341
SCHEMAT POWIATÓW	369	6 357	
SCHEMAT GMIN	2 271	70 700	665
SCHEMAT MIEJSCOWOŚCI	58 126	282 430	495
SCHEMAT ULIC	29 060	230 000	461
SCHEMAT KRAJÓW		7066	
SCHEMAT IMION	2 328	24360	289



# Czyszczenie danych

---

## Etap korekty nadmiarowych wpisów

Usuwanie jednej z powtórzonych nazw z kolumn zawierających nazwy ulic i miejscowości

### Przykłady działania transformacji

Przed użyciem transformacji		Po użyciu transformacji	
Nazwa miejscowości	Nazwa ulicy	Nazwa miejscowości	Nazwa ulicy
GDANSK	GDANSK	GDANSK	
1 SIERPNIA	1 SIERPNIA		1 SIERPNIA
ADAMÓWKA	ADAMÓWKA	ADAMÓWKA	ADAMÓWKA



# Czyszczenie danych

---

Etap kontroli i korekty logicznej - polegająca na zamianie różnych kombinacji zapisów na postać standardową w ujęciu kontekstowym.

Przykład dla rejestru:

REJESTR: Bielsko-Biała, Bielsko Biała = TERYT: Bielsko-Biała

REJESTR: Aleja, Al. = TERYT: Aleja

REJESTR: ul. Piękna, al. Piękna, droga Piękna = TERYT: ul. Piękna

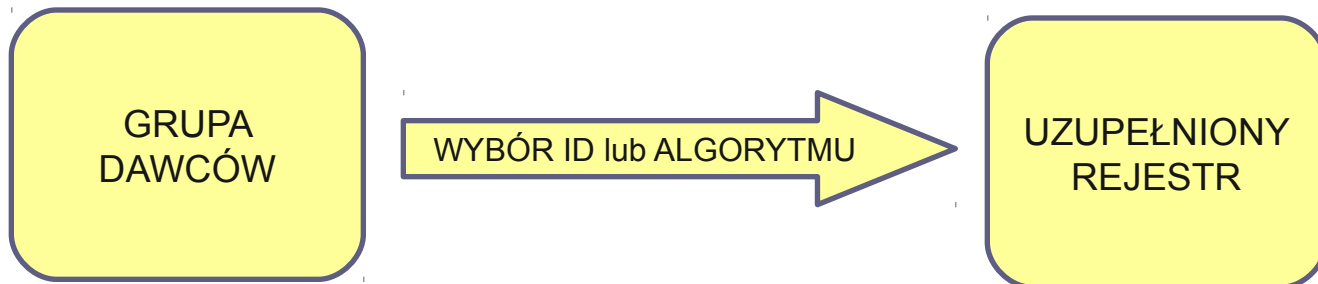


# Czyszczenie danych

---

Etap uzupełnienia identyfikatorów:

- PESEL
- NIP
- REGON
- kody TERYT



# Statystyki

GRUPA ZMIENNYCH	ZMIENNA	ILOŚĆ WYPEŁNIONYCH	BŁĘDNE PRZED CZYSZCZENIEM		ILOŚĆ WYPEŁNIONYCH	BŁĘDNE PO CZYSZCZENIU	
			ILOŚĆ	%		ILOŚĆ	%
ADRES	WOJEWÓDZTWO	9 730 594	470 075	4,83	12 753 861	0	0,00%
	POWIAT	6 067 829	814 607	13,43	12 753 861	52 256	0,47%
	GMINA	6 067 958	1 759 005	28,99	12 753 861	0	0,00%
	MIEJSCOWOŚĆ	17 067 115	3 813 291	22,34	17 070 157	126 566	0,74%
	PRZEDROSTEK				3 177 989	0	0,00%
	ULICA	15 907 593	9 267 692	<b>58,26</b>	13 918 763	618 914	<b>4,45%</b>

GRUPA ZMIENNYCH	ZMIENNA	ILOŚĆ WYPEŁNIONYCH	LICZBA BŁĘDNYCH		ILOŚĆ WYPEŁNIONYCH	LICZBA BŁĘDNYCH	
			ILOŚĆ	%		ILOŚĆ	%
ADRES	WOJEWÓDZTWO	10 743 015	2 744 507	25,55%	15 384 339	703	0,01%
	POWIAT	10 423 133	4 239 673	40,68%	15 287 693	889	0,01%
	GMINA	9 967 908	3 195 702	32,06%	15 251 519	3 721	0,03%
	MIEJSCOWOŚĆ	19 415 748	7 353 878	37,88%	19 143 727	124 639	0,66%
	PRZEDROSTEK	832 574	261 472	31,41%	10 887 303	0	0,00%
	ULICA	15 034 775	6 983 611	<b>46,45%</b>	14 699 866	928 350	<b>6,32%</b>
	ULICA (bez nazw miejscowości)	15 034 775	6 291 754	<b>41,85%</b>	14 699 866	236 493	<b>1,61%</b>



# Operacyjna Baza Mikrodanych

---

1. Ładowanie do STAGE
2. Poprawa jakości danych
  - czyszczenie przy pomocy zdefiniowanych słowników

17

- korekta danych na podstawie reguł logicznych

2100



# Operacyjna Baza Mikrodanych

---

3. Konwersja wartości m.in.:
- płeć
  - stan cywilno-prawny
  - obywatelstwo
  - stopień niepełnosprawności
  - niezdolności do pracy
  - wykształcenia
  - świadczeń rodzinnych

**3220**



# Operacyjna Baza Mikrodanych

---

woj	pow	gmn	miejsc	ulica	pleć	obyw	wyksz	kraj urodzenia
14	1401	1401015	0614506	25438	2	528	1	826

4. Przekształcenie zbiorów do tabel płaskich + deduplikacja i transpozycja
5. Integracja zmiennych z operatem
  - wybrane identyfikatory
  - alternatywne klucze łączenia
6. Wygenerowanie Master Rekordu (MR)
7. Wygenerowanie Golden Rekordu (GR)



DZIĘKUJĘ ZA UWAGĘ

