POZNAN UNIVERSITY OF TECHNOLOGY

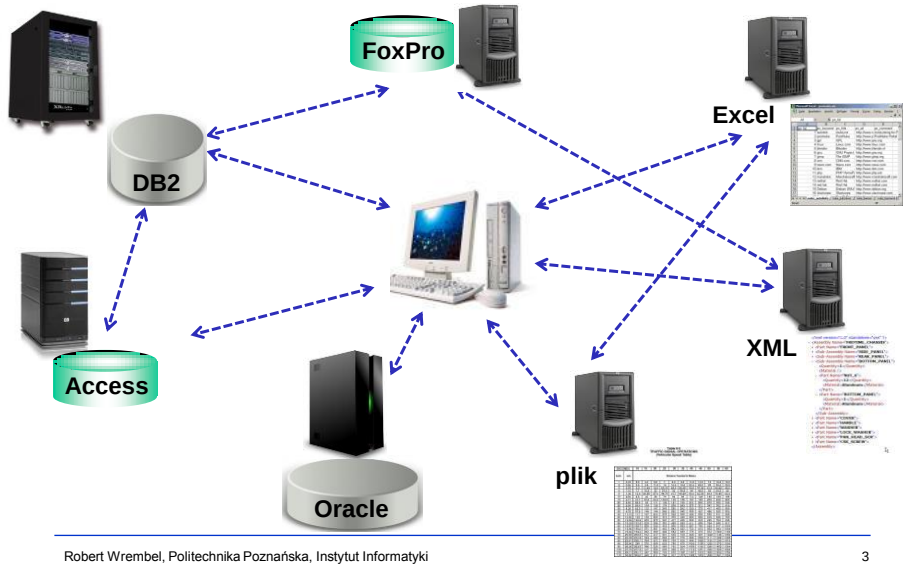
Architektury i technologie integracji danych

Robert Wrembel
Politechnika Poznańska
Instytut Informatyki
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel

Problematyka i architektury integracji danych

- Problemy dostępu do heterogenicznych źródeł danych
- Podstawowe architektury integracyjne
  - P2P
  - Systemy mediacyjne
  - Sfederowane BD
  - WEB Services/SOA
  - Systemy hurtowni danych

Problematyka integracji danych



Problematyka integracji danych

- Charakterystyka systemów źródłowych
  - rozproszenie
  - heterogeniczność
- Cele integracji
  - systemy rozproszone (rozproszone BD)
  - systemy transakcyjne
  - systemy analityczne (BI)

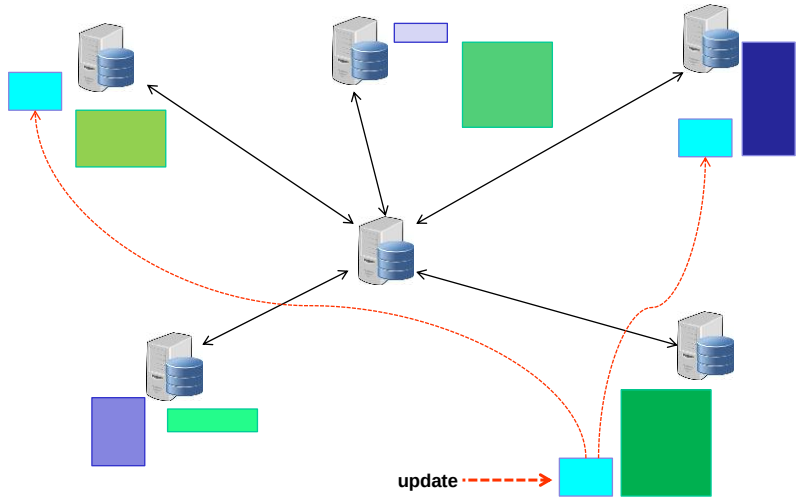
Rozproszone BD

- Przetwarzanie równoległe
- Load balancing
- Redukowanie ruchu sieciowego ⇨ dane "blisko" konsumenta
- Wzrost bezpieczeństwa danych na skutek awarii węzła

Rozproszone BD

- Mechanizmy
  - partycje
  - repliki
  - odświeżanie replik
    - pełne / przyrostowe
    - wykrywanie zmian
    - moment odświeżenia
  - protokół 2PC
  - algorytmy alokowania zasobów w węzłach
    - statyczne /dynamiczne
  - optymalizacja zapytań
    - redukcja planów wzgl. partycji
    - algorytmy łączenia
    - wybór replik

Rozproszone BD



Heterogeniczność źródeł

- **Różni producenci**
  - różne technologie implementacyjne
- **Różna funkcjonalność**
  - bazy danych / nie bazy danych
  - dialekty SQL
  - sposoby dostępu i przetwarzania danych
- **Różne modele danych**
  - hierarchiczne, sieciowe
  - relacyjne
  - obiektowe
  - obiektowo-relacyjne
  - wielowymiarowe
  - semistructured
  - grafowe

Heterogeniczność źródeł

- Różne modele danych w źródłach (relacyjny, obiektowy, semistructured, ...)
- Różne typy danych
  - smallint, int, biging, decimal (SQLServer)
  - smallint, int, biging, float, real, double (DB2)
  - number, binary_integer (Oracle)
  - znakowe typy danych o stałej i zmiennej długości
- Różne (sprzeczne) ograniczenia integralnościowe

Heterogeniczność źródeł

- Inna reprezentacja tych samych danych
  - Pracownicy{NIP, imię, nazwisko, adres_koresp}
  - Prac{NIP, imię_nazw, ulica, dom, kod, miasto}
  - Pojazdy (zawiera samochody osobowe + dostawcze)
  - Samochody_Osobowe, Samochody_Dostawcze
- Homonimy
  - Produkty.kod – oznacza kod produktu
  - Klienci.kod – oznacza kod pocztowy
- Synonimy
  - Pacjenci.pesel
  - Pacjenci.pacjentID (z wartością peselu)

Heterogeniczność źródeł

- Konflikty na poziomie danych
  - Zduplikowane dane
  - Brakujące i błędne dane
  - Błędy wprowadzania wartości
- Różne ziarno agregacji
  - sprzedaż dzienna
  - sprzedaż tygodniowa
- Różne jednostki miary
  - cena {PLN, EUR, USD}
  - waga {kg, dkg}

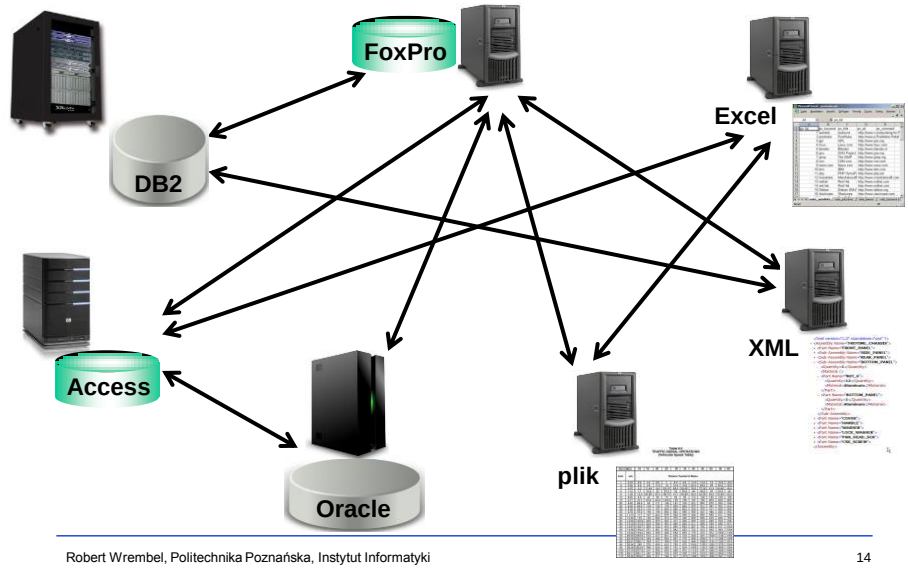
Heterogeniczność źródeł

Atrybut	niedozwolona wartość	data_ur=30.13.1970	
	brakująca wartość	NIP=null	
	błąd literowy	Pozań	
	oznaczenia symboliczne	LC3X	czerwony metalik
	skróty	Pozn., P-ń, PŃ	
	wielkość liter	Poznań, poznań, POZNAŃ	
	różne oznaczenia symboliczne	płeć: {0, 1}, {K, M}, {kobieta, mężczyzna}	
	format	20-03-2008, 03/20/08	
	wartości złożone	R. Wrembel, 25.06.68, Szamotuły	
	kolejność wartości	{R.Wrembel} {Wrembel R.}	
	jednoznaczność	Wenecja (Włochy), Wenecja (Polska)	

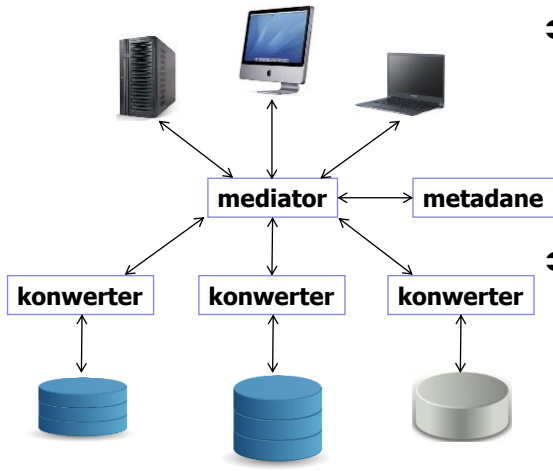
Heterogeniczność źródeł

Rekord	niespełniona zależność pomiędzy wartościami atrybutów	cena_netto=100 cena_brutto=190	cena_brutto=cena_netto* 1,22
		ulica='Piotrowo' kod=62-300	
	naruszenie unikalności	{Robert Wrembel, WL8539024} {Bartosz Bębel, WL8539024}	nr dowodu osobistego powinien być unikalny
	wskazanie do nieistniejącego rekordu	{Robert Wrembel, Z20}	zespół Z20 nie istnieje
	duplikaty	{Robert Wrembel} {Wrembel Robert} {R. Wrembel}	
	sprzeczne wartości	{R. Wrembel, Szamotuły} {R. Wrembel, Poznań}	

P2P



System mediacyjny



- **Wady**
  - czas dostępu do danych
  - niedostępność źródeł
  - konwersja zapytań i danych
  - brak modyfikacji
- **Zalety**
  - brak redundancji danych
  - dostęp do danych aktualnych

Przetwarzanie zapytań

- Wybór źródeł danych
- Dekompozycja zapytania ze względu na źródła
- Optymalizacja zapytania (jak w RBD)
  - statyczna lub dynamiczna
- Przesyłanie zapytań do źródeł danych (wrapper'ów)
- Wykonywanie zapytań w źródłach danych
- Przesyłanie wyników do mediatora (ewentualnie do innych wrapper'ów celem wykonania dalszych etapów zapytania)
- Łączenie wyników uzyskanych z poziomu poszczególnych źródeł na poziomie mediatora
- Prezentacja wyników użytkownikowi

Metadane

- Opisują
  - schemat globalny dostępny dla użytkownika
  - źródła danych
    - zawartość
    - funkcjonalność
  - odwzorowania schematu globalnego schematy udostępniane przez wrapper'y

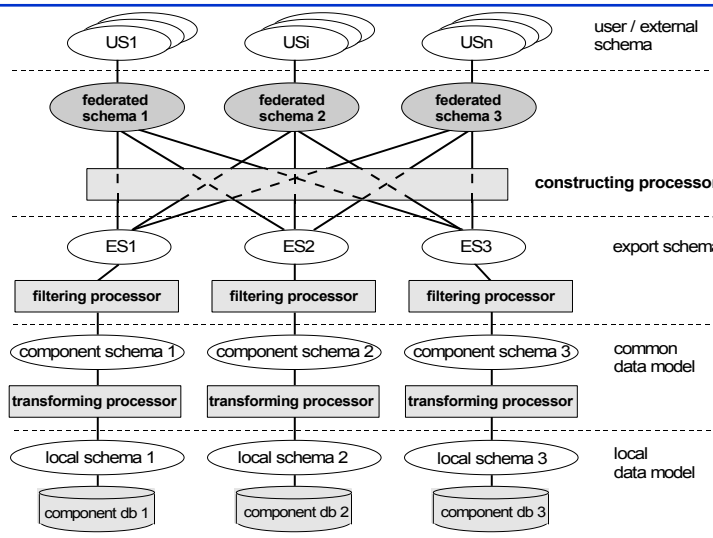
Dołączanie źródeł

- Silna integracja
  - na początku funkcjonowania systemu mediacyjnego źródła są analizowane i jest budowany schemat globalny
  - pojawienie się nowych źródeł wymaga wprowadzenia zmian w schemacie globalnym, które w istotny sposób wpływają na sposób działania mediatora w zakresie dekompozycji i optymalizacji zapytań
- Słaba integracja
  - każde źródło udostępnia swój lokalny schemat w modelu schematu globalnego
  - schematy lokalne są integrowane w globalnym

Rodzaje architektur

- Scentralizowana
  - źródła danych realizują podstawowe operacje
  - optymalizacja zapytania, wybór danych, łączenie, sortowanie, konwersja wyników są realizowane przez mediator
  - wrapper'y nie komunikują się między sobą
- Zdecentralizowana
  - źródła danych realizują złożone operacje takie selekcja danych, łączenie, sortowanie
  - optymalizacją zapytania zajmuje się mediator, ale pozostałe operacje mogą być przejęte przez warstwę źródeł danych
  - wrapper'y mają możliwość komunikowania się między sobą (np. wykonanie rozproszonego połączenia)

Sfederowane BD



WEB Services

- Application integration technology
  - allows programs written in different languages on different platforms to communicate with each other based on an agreed way of communication
  - software service
- Exchange data between different applications and different platforms
- Program-to-program communications model, built on existing and emerging standards such as HTTP, XML, SOAP, WSDL, and UDDI
- Exchanged data encoded as XML

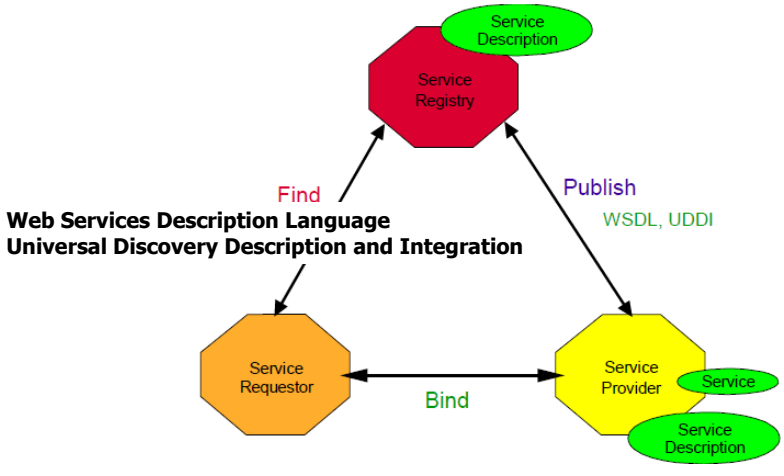
WEB Services

- Transport protocol
  - XML-based Simple Object Access Protocol (SOAP) to let applications exchange data over HTTP
- Web services provide a way to describe their interfaces so that other applications can communicate with Web services
  - this description is usually provided in an XML document called a Web Services Description Language (WSDL) document
- Web services are registered in a directory for the purpose of finding them
  - registration is done by means of Universal Discovery Description and Integration (UDDI)

WEB Services

- Example
- A purchasing application
  - automatically obtain price information from multiple vendors
  - select a vendor
  - submit the order
  - track the shipment until it is received
- A vendor application
  - exposing its services on the Web
  - check the customer's credit
  - charge the customer's account
  - set up the shipment with a shipping company

WEB Services



Web Services Conceptual Architecture (IBM)

WEB Services

- **Service provider**
  - **makes available a software (service)**
  - **provides the service description** (data types, operations, binding information and network location)
  - **publishes the description to a service registry or requestor**
- **Service requestor**
  - **finds and retrieves the service description**
  - **uses the service description to bind with the service provider and to invoke or interact with the Web service implementation**
- **Service registry**
  - **searchable registry of service descriptions where service providers publish their service descriptions**
  - **service requestors find services and obtain binding information for services during development (for static binding) or during execution (for dynamic binding)**

SOAP

- **Simple Object Access Protocol → Communication protocol for Web services**
- **SOAP is a specification that defines**
  - **the XML format for messages, i.e., how to represent data**
  - **the format of an HTTP message that contains a SOAP message**
- **Independent of application programming languages**

SOAP

- In practice, SOAP messages are created and parsed by various toolkits that translate function calls from some kind of language to a SOAP message.
  - Microsoft SOAP Toolkit translates COM function calls to SOAP
  - Apache Toolkit translates JAVA function calls to SOAP
  - types of function calls and the data types of the parameters depend on a SOAP implementation

WSDL

- Web Services Description Language
- WSDL is an XML document that describes a set of SOAP messages and how the messages are exchanged
  - in practice generated by toolkits based on existing program interfaces
  - parsed by software
- WSDL describes Web services and is used to locate them

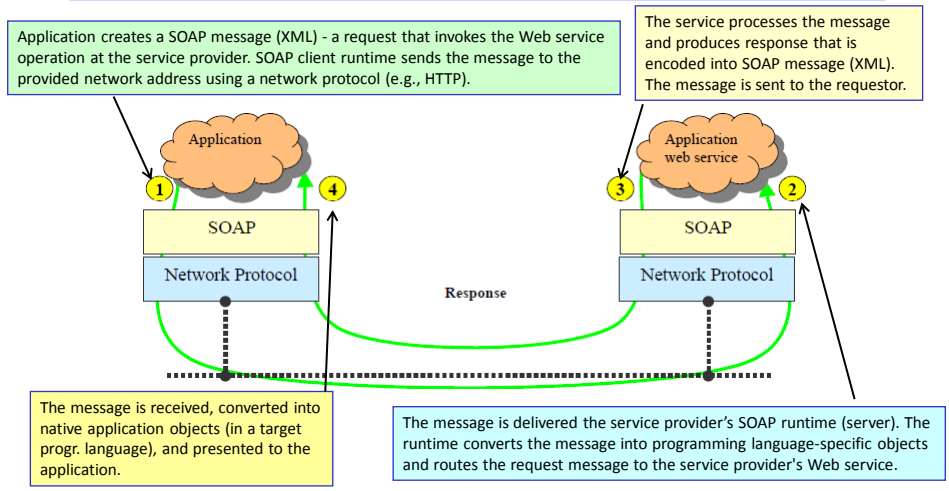
WSDL

- The service interface includes
  - WSDL:binding - describes among others a protocol, data format, security for a particular service interface (WSDL:portType)
  - WSDL:portType - defines operations (conterpart of a method signature in a programming language) of a Web service, i.e., what XML messages can appear in the input and output data flows
  - WSDL:message - specifies which XML data types constitute various parts of a message (e.g., input and output parameters of an operation)
  - WSDL:type - describes complex data types

UDDI

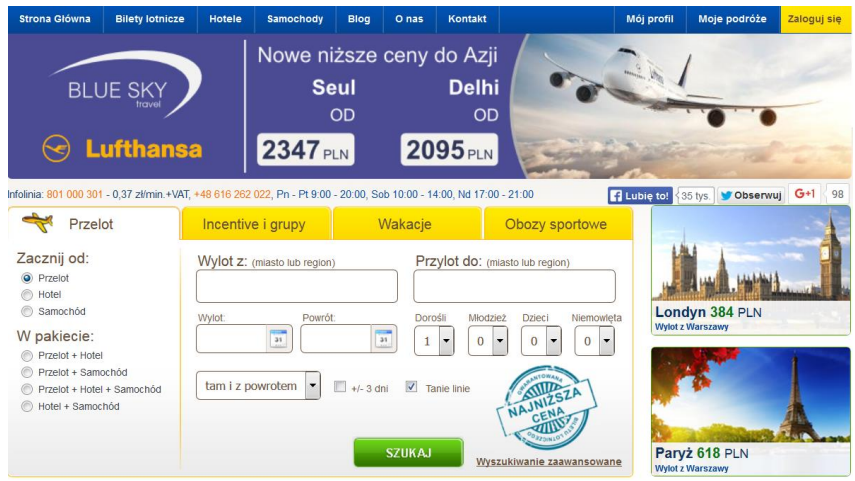
- Universal Discovery Description and Integration
- A UDDI directory entry is an XML file that describes a business and the services it offers
- Three parts to the entry in the UDDI directory
  - info about a company offering the service: name, address, contacts, ...
  - industrial categories based on standard taxonomies such as the North American Industry Classification System and the Standard Industrial Classification
  - description of the interface to the service

WEB Services' interaction

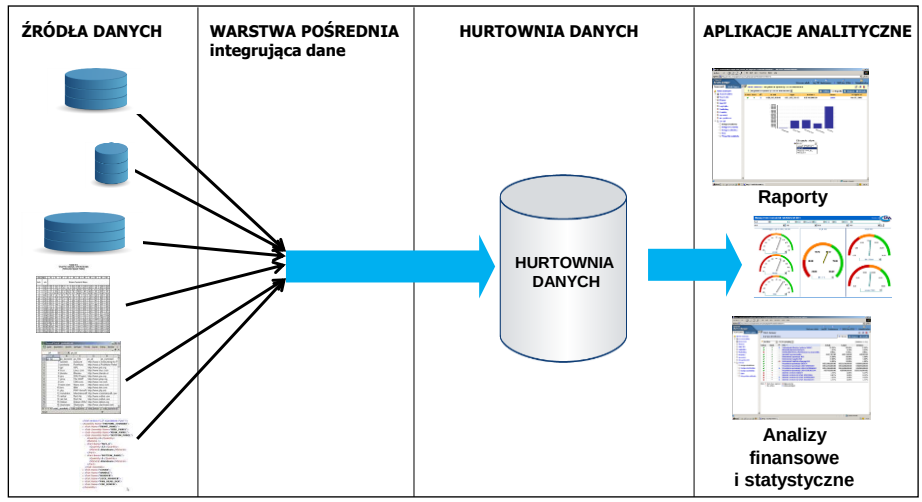


Web Services Conceptual Architecture (IBM)

Przykład wykorzystania WS

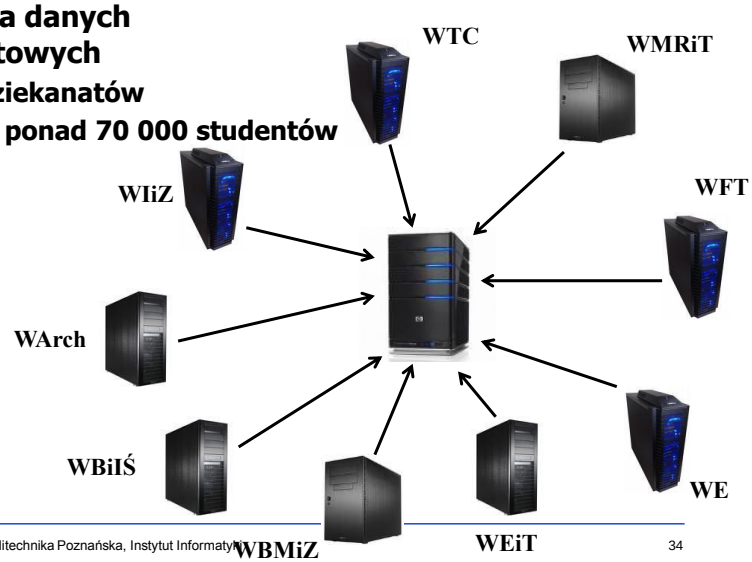


Hurtownia Danych



Jakość danych - studium przypadku

- Integracja danych dziekanatowych
  - 9 BD dziekanatów
  - łącznie ponad 70 000 studentów



Wartości unikalne

- **Nr indeksu**
  - **wartość unikalna w ramach jednej BD dziekanatu**
  - **wartość nieunikalna globalnie w ramach BD różnych dziekanatów ⇨ problem integracji danych**
  - **zidentyfikowano 49 par studentów o takim samym numerze indeksu, studenci w parze są innymi osobami**

Jakość danych nr indeksów

- **Postać nr indeksu: liczba + opcjonalnie litera**
  - **a - absolwent**
  - **d - drugi kierunek**
  - **s - skreślony lub zrezygnował**
  - **i - inne przypadki**

zawartość numeru albumu	liczba
tylko cyfry	65501
litera d na końcu	1247
litera a na końcu	433
litera s na końcu	992
litera i na końcu	982
litera p na końcu	603
litera x na końcu	358
litera g na końcu	146
litera r na początku	0
litera s na początku	225
wszystkie	71970
z literą na końcu	5333
z literą na początku	310
z literą	5625
białe znaki	293
inne znaki niż liczby	6471

Jakość danych PESEL

- **Różne osoby posiadające ten sam nr PESEL (również z tego samego wydziału)**
  - **13 powtarzających się numerów PESEL**
- **Poprawność wartości PESEL**

PESEL	liczba
wszystkie	71970
poprawny	56344
zła długość	8906
niepoprawne znaki przy dobrej długości	24
błąd sumy kontrolnej	6696
zła płeć przy założeniu M	763
zła płeć przy założeniu K	17

Jakość imion

- **Wykorzystano słownik imion (ponad 1700 imion;** http://piotr.eldora.pl/bazy-danych-kody-pocztowe-imiona-panstwa**)**

imię	liczba
wszystkie	86947
poprawne	86256
niepoprawne względem słownika	691
znaki inne niż litery	309
białe znaki	224
cyfry	57
białe znaki na początku lub końcu	16
dwa imiona	97
informacja o drugim kierunku	55
informacja o ITS	21
wielkimi literami	3
zła płeć przy imieniu M	30
zła płeć przy imieniu K	458

Jakość imion

imię	opis
Agata Agnieszka	wprowadzone dwa imiona
Natalia, Anna	wprowadzone dwa imiona z przecinkiem
Slawomir	zastąpienie polskiej litery diakrytycznej zwykłą
Joannna	literówka — nadmiarowa litera
Krzystof	literówka — brakująca litera
Przemysałw	literówka — przestawienie liter
Sikorska	nazwisko zamiast imienia
_Maciej	spacja występująca na początku lub końcu imienia
Marcin'	znak inny niż litera
wz.st.	informacja o wznowieniu studiów zamiast imienia
ITS	informacja o indywidualnym toku studiów zamiast imienia
-	brak imienia
Jakub 2 Kier.	informacja o drugim kierunku
Marcin S.Kazimierza	informacja o ojcu
Gniewosław	imię nie występuje w zbiorze poprawnych imion polskich, ale może być prawidłowe
Kevin	imię zagraniczne, może być prawidłowe

Jakość nazwisk

- **Wykorzystano słownik nazwisk (r.męski, 20000 najpopularniejszych nazwisk;** http://www.futrega.org/etc/nazwiska.html**)**

nazwisko	liczba
wszystkie	56607
niepoprawne względem słownika	11657
znaki inne niż litery, myślnik lub białe znaki	2
białe znaki	18
cyfry	1
białe znaki na początku lub końcu	2
wielkimi literami	1

Jakość danych słownikowych

- **Kategorie studiów**
  - **poprawne: {stacjonarne, niestacjonarne}**

kategoria studiów	liczba
Stacjonarne	9
Niestacjonarne	10
Trzeciego stopnia	7
Studia doktoranckie	1
Studia Wieczorowe	1
Niestacjonarne(wiecz)	1
Wieczorowe	1
Niestacjonarne(wieczorowe)	1

Jakość danych słownikowych

- **Rodzaje studiów**
  - **poprawne: {I stopnia, II stopnia, III stopnia}**
- **Kierunki studiów**

kierunek studiów	liczba
wszystkie	299
informacja o kategorii	137
informacja o rodzaju	154
informacja o miejscu	57
znaki inne niż litery i białe znaki	165
inaczej niż tylko pierwsza litera wielka	244

rodzaj studiów	liczba
Niestacjonarne I stopnia	11
Stacjonarne I stopnia	8
Niestacjonarne II stopnia	7
Stacjonarne II stopnia	7
Stacjonarne magisterskie	7
Niestacjonarne III stopnia	2
Niestacjonarne magisterskie	2
stacjonarne I stopnia	2
Stacjonarne III stopnia	2
Wieczorowe inżynierskie	2
Dzienne magisterskie uzupełniające	1
Niestacjonarne I stopnia (4 letnie)	1
Niestacjonarne II stopnia (uzupełniające)	1
Niestacjonarne II-stopnia	1
Niestacjonarne I-stopnia	1
Niestacjonarne magisterskie jednolite	1
Niestacjonarne uzupełniające	1
Stacjonarne I stopnia.	1
Stacjonarne II stopnia.	1
stacjonarne II stopnia	1
Stacjonarne II stopnia - 1,5-roczne	1
stacjonarne magisterskie	1
Stacjonarne magisterskie - 2-letnie	1
Stacjonarne magisterskie jednolite	1
Stacjonarne uzupełniające	1
Studia niestacjonarne	1
Studia stacjonarne	1
Wieczorowe	1
wieczorowe inżynierskie	1
Wieczorowe magisterskie uzupełniające	1
Zaoczne uzupełniające studia magisterskie	1
Zaoczne zawodowe	1

Jakość danych słownikowych

➲ Słownik miast

miejscowość	liczba
wszystkie	14293
znaki inne niż litery, myślnik lub białe znaki	567
cyfry	39
białe znaki na początku lub końcu	4
inaczej niż tylko pierwsza litera wyrazu wielka	11563

BYDGOSZCZ	9	605
Bydgoszcz	2	17
BYDGOSKIE	1	1
BYDGOSZCZ ADRES	1	0
GORZÓW WLKP.	7	173
Gorzów Wielkopolski	7	92
GORZÓW WIELKOPOLSKI	4	358
GORZÓW WLKP	1	33
GORZÓW WLKP-	1	5

miejscowość	liczba baz	liczba adresów
POZNAŃ	9	12633
Poznań	2	369
?????POZNAŃ???	1	1
POZNAN	1	73
KONIN	9	1455
Konin	2	95
konin	1	23
KONIN 2	1	0
KALISZ	9	1748
Kalisz	2	76
KaLISZ	1	39
kalisz	1	21
KALisz	1	6
KALISz	1	2
KaIISZ	1	2
kALISZ	1	2
kAIIsz	1	2
kaLISZ	1	2
kaliSZ	1	2
KALIsz	1	1
KAlisz	1	1
KaLiSZ	1	1
KaLiSz	1	1
KaLisz	1	1
kAIisZ	1	1
GNIEZNO	9	1517
Gniezno	3	45
PIŁA	9	1231
Piła	2	30
LESZNO	9	798
Leszno	2	25
WRZEŚNIA	9	732
Września	2	35
SWARZĘDZ	9	690
Swarzędz	2	33