

od **STUDENTA** *do* **SPECJALISTY** **BUSINESS** *konferencja* **INTELLIGENCE**

**„Dynamiczne bazy danych – platforma szkoleniowa Warszawskiej
Wyższej Szkoły Informatyki”**

Andrzej Ptasznik



Geneza projektu „Dynamiczna populacja”

► Poszukiwanie inspiracji

- AeroSym – symulacja ruchu w przestrzeni powietrznej
- Via Salus (e-recepta) – model działania służby zdrowia e-NFZ
- Pro – symulacja działania komunikacji w Warszawie



Geneza projektu „Dynamiczna populacja”

- **Specjalizacja w WWSI**
- **Brak szkoleniowych baz danych**
- **Ciekawe wyzwanie**
- **Zapewnienie ciągłości prac w kole naukowym**



Co, po co i dlaczego ...?

- Utworzenie bazy danych będącej obraz populacji ludności w Polsce
- Opracowanie symulatora zmian populacji i wybranych zdarzeń zachodzących w Polsce
- Opracowanie scenariuszy wykorzystania systemu



Prace przygotowawcze

- ▶ Podstawą jest system Teryt udostępniony przez GUS, który musiał zostać odpowiednio przygotowany
- ▶ Do każdej lokalizacji zostały dowiązane współrzędne geograficzne(ok. 300 000 lokalizacji)
- ▶ Każda osoba populacji „Herkules” jest dowiązana do lokalizacji
- ▶ Dane będą wykorzystywane do zobrazowań przestrzennych



Prace przygotowawcze

Przykładowy
fragment danych
systemu TERYT

IdTeryt	Name	Hid	typID	geo	SHid	Gmina	Powiat	Wojewodztwo	KodTeryt	CzyMiasto	SIMC
1	Dzierżonów	0x5B5AB0	1	0xE6100000010CB25D465623A8304017B83CD68C5D4940	/1/2/2/1/	Dzierżonów	dzierżoniowski	DOLNOŚLĄSKIE	0202021	1	0983988
2	Dziewa	0x6D35CA	1	0xE6100000010C80D250A390783240518138B46D5D4A40	/2/9/1/6/	Dąbrowa Biskupia	inowrocławski	KUJAWSKO-POMORSKIE	0407022	0	0083859
3	Dziewiątło	0xB73B61	1	0xE6100000010C99B44E01D534354085F4CAAB29594940	/13/7/2/4/	Iwaniska	opatowski	ŚWIĘTOKRZYSKIE	2606022	0	0793609
4	Dziewule	0x9F0F6DD1	1	0xE6100000010C1D120885526736408962F20698064A40	/7/30/13/8/	Zbuczyn	siedlecki	MAZOWIECKIE	1426132	0	0696361
5	Dzięcielec	0xAF8331D5	1	0xE6100000010C16E8C9EB77E93140FD288F13CB3F4B40	/11/20/5/10/	Łęczycę	wejherowski	POMORSKIE	2215082	0	0167214
6	Dzięgiele Oleckie	0xBB7DB5C0	1	0xE6100000010C38CB76FCB78364048B76FFDFBF64A40	/14/15/2/11/	Olecko	olecki	WARMIŃSKO-MAZURSKIE	2813043	0	0763650
7	Dzikowice	0x86DCB440	1	0xE6100000010CF8167BD4043A2F40DF3BC505FBCB4940	/4/13/6/8/	Szprotawa	żagański	LUBUSKIE	0810073	0	0914600
8	Dzikowiec	0x5D3A7440	1	0xE6100000010CA88FC01F7E923040F15C84CE10494940	/1/9/9/8/	Nowa Ruda	kłodzki	DOLNOŚLĄSKIE	0208112	0	0854140
9	Dzikowo	0x85F0D9	1	0xE6100000010CF8B0CD7CDBD52D409EF3AED579FB4940	/4/3/4/12/	Gubin	krośnieński	LUBUSKIE	0802052	0	0909609
10	Dzikówko	0xC11B2B9C	1	0xE6100000010C5B4EF85BF14F2E40FC3FE2B2657E4A40	/16/12/1/7/	Barlinek	myśliborski	ZACHODNIOPOMORSKIE	3210013	0	0177603
11	Dziunin	0x9E59EA80	1	0xE6100000010CAEE98BCE43B534403A15BA5C0EDC4940	/7/6/7/10/	Mogielnica	grójcki	MAZOWIECKIE	1406073	0	0629270
12	Dziwiszów	0x5CED78	1	0xE6100000010C6B5FE57569982F406CC5484096784940	/1/7/2/3/	Jeźów Sudecki	jeleniogórski	DOLNOŚLĄSKIE	0206062	0	0189760
13	Dziwogóra	0xC11C176DB0	1	0xE6100000010CB65A71BBE10A30407948E7B287E44A40	/16/19/2/13/	Połczyn-Zdrój	świdwiński	ZACHODNIOPOMORSKIE	3216033	0	0310002
14	Dziwy	0x9F2CDF18	1	0xE6100000010CC24AA5BE2CF3340D8744EFDCD8D4A40	/7/41/3/5/	Lubowidz	żuromiński	MAZOWIECKIE	1437032	0	0118523
15	Dzwonek	0x9EFBEC80	1	0xE6100000010C5C4C7D6A06B635409BB0581EFF7A4A40	/7/15/3/12/	Czerwin	ostrolęcki	MAZOWIECKIE	1415032	0	0508201
16	Dzwonkowie	0x8F05E78C	1	0xE6100000010C17E41D96BC4E344023FF829323F14940	/5/19/7/5/	Nowy Kawęczyn	skiemiewicki	ŁÓDZKIE	1015072	0	0734222
17	Dubicze Cerkiewne	0xAB2C3080	1	0xE6100000010C52BAF42F49713740B541DC8B1E534A40	/10/6/4/4/	Dubicze Cerkiewne	hajnowski	PODLASKIE	2005052	0	0027915
18	Dublany	0xAAD9EC80	1	0xE6100000010C3836960BF0E23740BD5BB45B26864A40	/10/2/7/12/	Michałow	białostocki	PODLASKIE	2002073	0	0034252
19	Duchów	0x9ECBE880	1	0xE6100000010C3DCFA47D37BA3540625FFC7E7B274A40	/7/12/3/8/	Dobre	miński	MAZOWIECKIE	1412062	0	0670290



Prace przygotowawcze

Klasyfikacja chorób wg WHO - kod ICD10

Przykładowy
fragment
danych
systemu ICD10

Wykorzystywane do określenia przyczyn zgonów i diagnoz lekarskich

id	name	parent
D34	Nieżłośliwy nowotwór tarczycy	D10-D36
D35	Nieżłośliwy nowotwór innych i nieokreślonych gruczołów wydzielania wewnętrznego	D10-D36
D50	Niedokrwistości z niedoboru żelaza	D50-D53
D50-D53	Niedokrwistości z niedoborów pokarmowych	D50-D89
D50-D89	Choroby krwi i narządów krwiotwórczych oraz niektóre choroby przebiegające z udziałem mechanizmów autoimmunologicznych	
D51	Niedokrwistości z niedoboru witaminy B12	D50-D53
D52	Niedokrwistość z niedoboru kwasu foliowego	D50-D53
D55	Niedokrwistość spowodowana zaburzeniami enzymatycznymi	D55-D59
D55-D59	Niedokrwistości hemolityczne	D50-D89
D56	Talasemia	D55-D59
D57	Zaburzenia połączone z sierpowatokrątkowością	D55-D59
D58	Inne dziedziczne niedokrwistości hemolityczne	D55-D59
D60	Nabyta aplazja czysto czerwonekrwinkowa (erythroblastopenia)	D60-D64
D60-D64	Niedokrwistości aplastyczne i inne	D50-D89
D61	Inne niedokrwistości aplastyczne	D60-D64
D62	Ostra niedokrwistość pokrwotoczna	D60-D64
D63	Niedokrwistość w przebiegu chorób przewlekłych sklasyfikowanych gdzie indziej	D60-D64
D65	Rozsiane wykrzepianie wewnątrznaczyniowe	D65-D69
D65-D69	Zaburzenia krzepnięcia, płamice oraz inne skazy krwotoczne	D50-D89



Prace przygotowawcze

Statystyki,
statystyki,
statystyki

A	B	C	D	E	F
Kod	Jednostka terytorialna		Lata	rostka m	Wartość
1100000000	ŁÓDZKIE	ogółem	2011	-	23952
1100000000	ŁÓDZKIE	12	2011	-	0
1100000000	ŁÓDZKIE	13	2011	-	2
1100000000	ŁÓDZKIE	14	2011	-	1
1100000000	ŁÓDZKIE	15	2011	-	15
1100000000	ŁÓDZKIE	16	2011	-	58
1100000000	ŁÓDZKIE	17	2011	-	148
1100000000	ŁÓDZKIE	18	2011	-	281
1100000000	ŁÓDZKIE	19	2011	-	413
1100000000	ŁÓDZKIE	20	2011	-	533
1100000000	ŁÓDZKIE	21	2011	-	653
1100000000	ŁÓDZKIE	22	2011	-	772
1100000000	ŁÓDZKIE	23	2011	-	914
1100000000	ŁÓDZKIE	24	2011	-	1180
1100000000	ŁÓDZKIE	25	2011	-	1428
1100000000	ŁÓDZKIE	26	2011	-	1818

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
kod teryt	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
0	68708	59393	51943	51003	52871	51634	59649	68950	76061	79290	78760	72828	72418	73453
1100000000	4394	3803	3467	3384	3638	3406	3753	4316	4488	4897	4854	4500	4213	4392
1140000000	8263	7547	6868	6849	7101	7098	7447	8734	9506	9823	9934	9236	9307	9164
2120000000	4972	3914	3402	3349	3479	3364	3992	4623	5012	5227	5420	5106	4955	4792
2240000000	9802	8275	7236	7285	7465	7446	8493	9985	10720	11219	11075	10015	9830	9807
3060000000	3410	3025	2525	2581	2664	2742	3227	3496	3795	3722	3638	3430	3457	3319
3180000000	2629	2322	2151	2094	2302	2340	2599	3041	3258	3150	3163	2999	3081	2919
3200000000	1846	1605	1383	1381	1456	1446	1603	1713	1893	2112	1994	1889	2001	1960
3260000000	1998	1716	1409	1372	1482	1451	1577	1879	2150	2097	2043	2027	2008	1977
4080000000	2251	1903	1615	1467	1559	1509	1790	2146	2485	2537	2485	2377	2313	2554
4300000000	5508	4479	3895	3758	3900	3773	4716	5485	6204	6804	6835	6267	6209	6407
4320000000	4049	3530	3037	2941	2995	2767	3294	3903	4405	4598	4753	4261	4005	4168
5020000000	6526	5764	4913	4718	4902	4723	5632	6335	7266	7375	7221	6725	6879	7311
5160000000	1837	1603	1528	1553	1573	1525	1800	2200	2163	2267	2254	2102	2017	2095
6040000000	4341	3744	3290	2999	2887	2834	3545	3908	4419	4754	4694	4172	4259	4461
6220000000	4189	3698	3283	3200	3229	3112	3725	4558	5113	5324	5233	4690	4765	4879
6280000000	2893	2465	2141	2072	2239	2098	2456	2628	3184	3384	3364	3032	3119	3248

Bardzo dużo pracy polegało na wyszukiwaniu odpowiednich statystyk i przygotowaniu ich do postaci możliwej do zaimportowania do bazy danych



Efekty

- Wygenerowany obraz populacji na 31.12.2011 (według spisu GUS)

IdOsoby	Nazwisko	Imie	DataUrodzenia	DPPesel	IdTeryt	CzyŻyje	CzyKobieta	IdStanuCywilnego	InneDane	IdWykształcenia	RokUr	Wiek
-2000000000	Ozimłowska	Karolina	2011-06-26	20110626999075	271360	1	1	1	<Dane />	1	2011	2
-1999999999	Dranka	Karolina	2011-01-12	20110112999056	71745	1	1	1	<Dane />	1	2011	2
-1999999998	Czurlok	Jagoda	2011-07-28	20110728999000	214996	1	1	1	<Dane />	1	2011	2
-1999999997	Domkow	Aleksandra	2011-06-10	20110610999050	203418	1	1	1	<Dane />	1	2011	2
-1999999996	Kowalczyk	Olena	2011-09-13	20110913999059	25687	1	1	1	<Dane />	1	2011	2
-1999999995	Piestrak	Oliwia	2011-10-25	20111025999077	264851	1	1	1	<Dane />	1	2011	2
-1999999994	Przemieniecki	Amelia	2011-04-18	20110418999074	167512	1	1	1	<Dane />	1	2011	2
-1999999993	Przystaj	Małgorzata	2011-10-03	20111003999057	162644	1	1	1	<Dane />	1	2011	2
-1999999992	Poziemkiewicz	Maria	2011-05-14	20110514999062	49479	1	1	1	<Dane />	1	2011	2
-1999999991	Glenek	Izabela	2011-11-23	20111123999071	90299	1	1	1	<Dane />	1	2011	2
-1999999990	Grawirska	Maja	2011-05-20	20110520999057	117247	1	1	1	<Dane />	1	2011	2
-1999999989	Wutke	Pola	2011-01-06	20110106999061	197152	1	1	1	<Dane />	1	2011	2
-1999999988	Sapiewska	Marika	2011-09-01	20110901999046	37826	1	1	1	<Dane />	1	2011	2
-1999999986	Kręblewska	Wiktoria	2011-05-31	20110531999067	197152	1	1	1	<Dane />	1	2011	2
-1999999985	Kowalska	Hanna	2011-11-10	201111110999055	167512	1	1	1	<Dane />	1	2011	2
-1999999984	Pogoda	Antonina	2011-02-23	20110223999066	220029	1	1	1	<Dane />	1	2011	2
-1999999983	Taparek	Maja	2011-12-18	20111218999079	170016	1	1	1	<Dane />	1	2011	2
-1999999982	Smaudzik	Hanna	2011-06-10	20110610998050	256493	1	1	1	<Dane />	1	2011	2
-1999999981	Prześlakiewicz	Maja	2011-07-29	20110729999003	225740	1	1	1	<Dane />	1	2011	2
-1999999980	Gmośńska	Pola	2011-07-07	20110707999064	54148	1	1	1	<Dane />	1	2011	2
-1999999979	Osiatkowska	Zuzanna	2011-09-05	20110905999058	224944	1	1	1	<Dane />	1	2011	2
-1999999978	Górska	Zuzanna	2011-06-27	20110627999078	256493	1	1	1	<Dane />	1	2011	2
-1999999977	Brzutek	Maja	2011-10-08	20111008999072	256493	1	1	1	<Dane />	1	2011	2
-1999999976	Artus	Izabela	2011-11-05	20111105999063	38898	1	1	1	<Dane />	1	2011	2
-1999999975	Machtycz	Małgorzata	2011-06-15	20110615999065	140173	1	1	1	<Dane />	1	2011	2
-1999999974	Loesche	Julia	2011-09-26	20110926999075	101464	1	1	1	<Dane />	1	2011	2
-1999999973	Organista	Marta	2011-07-27	20110727999078	35549	1	1	1	<Dane />	1	2011	2
-1999999972	Panaczek	Julia	2011-12-29	20111229999008	257309	1	1	1	<Dane />	1	2011	2



Efekty

- ▶ Opracowany został mechanizm losowania i przypisywania cech
- ▶ Opracowano postaci komunikatów, które ma otrzymywać system
- ▶ Opracowano procedury generowania informacji o zgonach, urodzeniach i zmianach stanu cywilnego
- ▶ Opracowano procedurę generowania komunikatów USC

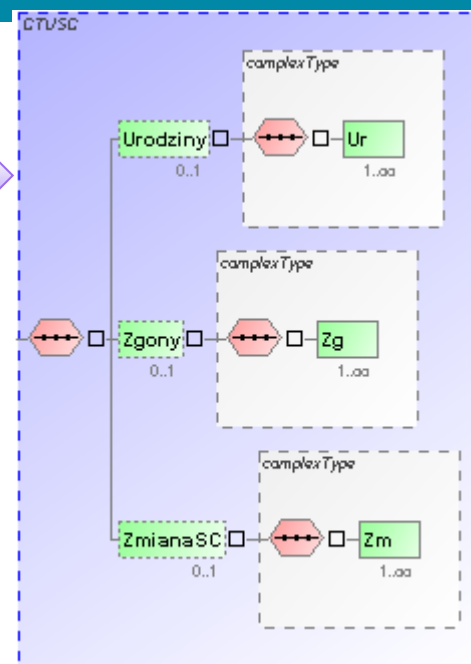
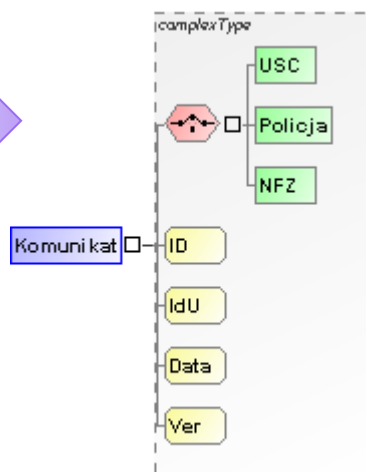
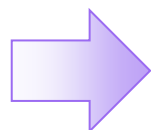




Efekty

Schemat
XSD

Przykładowy
komunikat
XML



```
<Komunikat Id="1" IdU="18781" Data="20130318" Ver="1">
  <USC>
    <Urodziny>
      <Ur Pesel="" Imie="Jacek" Nazwisko="Kowalski" PeselO="19940724899909" PeselM="19940202904812" Data="20130315" Plec="M"/>
      <Ur Pesel="" Imie="Olga" Nazwisko="Masłowska" PeselO="19820918864961" PeselM="19820210870844" Data="20130316" Plec="K"/>
    </Urodziny>
    <Zgony>
      <Zg Pesel="19390803914876" Data="20130311" ICD="Q01"/>
      <Zg Pesel="19350912927875" Data="20130314" ICD="Y85"/>
      <Zg Pesel="19340504959911" Data="20130314" ICD="T51"/>
    </Zgony>
    <ZmianySC>
      <Zm Pes1="19851125860873" Naz1Po="Iżewicz-Friess" Pes2="19790221861961" Data="20130314" Typ="M"/>
      <Zm Pes1="19830927864868" Naz1Po="Wermajster-Hilber" Pes2="19701113905901" Naz2Po="Hilber-Wermajster" Data="20130314" Typ="M"/>
    </ZmianySC>
  </USC>
</Komunikat>
```



id	prob
335	1735955024
336	584222602
337	1638184554
338	1499081245
339	290197005
340	576227314
341	93387669
342	1926567948
343	737828276
344	1388889391
345	872587780
346	47645424
347	134523865
348	1035617817
349	1406092054
350	1199339493
351	556736577
352	1515019777
353	1984903838

Tabela pomocnicza

**100 mln liczb losowych z
przedziału 0 .. 2 mld.**

**Mechanizm losowej
wymiany liczb**



Losowanie

Imię	Nazwisko	Prawdopodobieństwo	Liczba losowa
Jan	Maliniak	0,6	1 200 000 000
Piotr	Nowak	0,7	1 400 000 000
Szymon	Kowalski	0,3	600 000 000



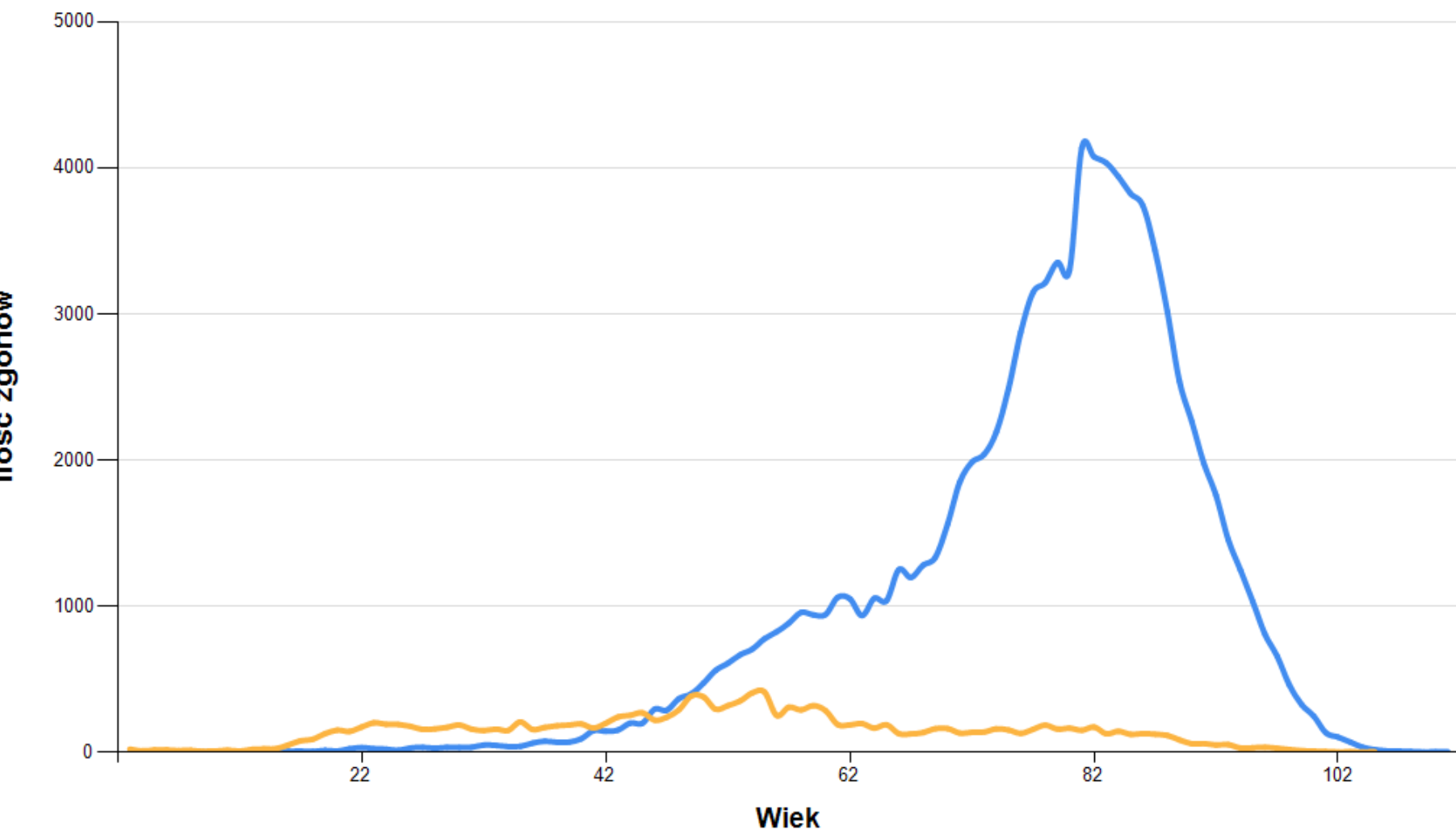
IdOsoby	Prob	liczba losowa
3458	13755550	234576
3459	22187654	1112343456
3460	503560345	567334658
3461	111765456	987576354
3462	4565687	23525



Osoba	Liczba losowa	Wartość od	Wartość do	Przyczyna zgonu
Jan Nowak	16578	1	300	Przyczyny zewnętrzne
Jan Nowak	16578	301	1000000	Nowotwory
Jan Nowak	16578	1000001	1240000	Choroby krążenia
Jan Nowak	16578	1240001	3000000	Choroby zakaźne

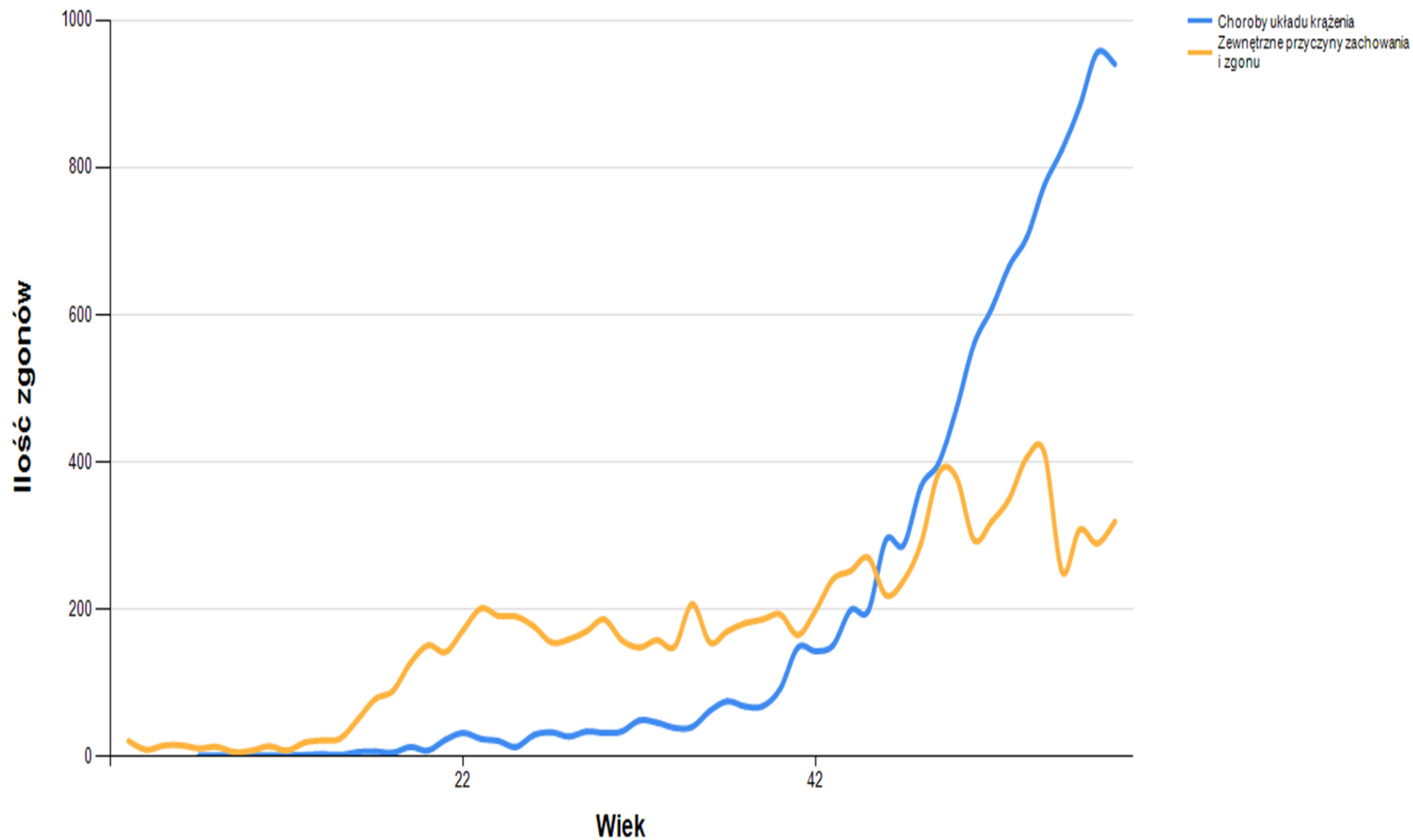


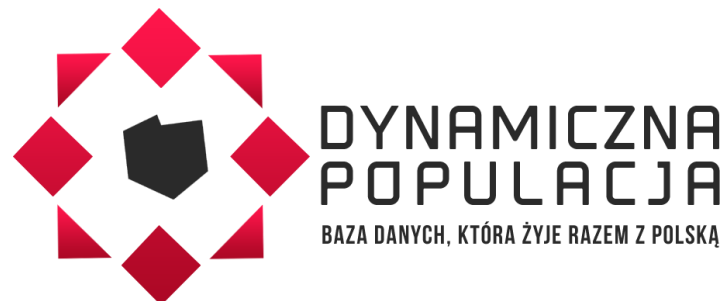
Wyniki





Wyniki





- Uruchomienie testowego generowania danych – zaczęliśmy od zgonów
- Przeprowadzenie analizy statystycznej wyników symulacji
- Uruchomienie pełnej wersji komunikatu USC
- Uruchomienie pełnej symulacji
- Budowa systemów analitycznych
- Tworzenie alertów i scenariuszy wykorzystania



Ale było mało

W ciągu ostatniego roku powstały trzy nowe dynamiczne bazy danych :

	AUD	dolar australijski
	CAD	dolar kanadyjski
	CHF	frank szwajcarski
	CZK	korona czeska
	DKK	korona duńska
	EUR	euro
	GBP	funt brytyjski
	HUF	forint węgierski
	NOK	korona norweska
	SEK	korona szwedzka
	USD	dolar amerykański



Kursy waluty od 2002 roku
Wygenerowano ok..3 mln faktur



Plik konfiguracji generatora w XML



Codziennie generowane są zdarzenia i ich obsługa



Ale było mało

Aktualnie trwają prace nad kolejnymi dynamicznymi bazami danych :



Depesze SYNOP z ok. 50 stacji pomiarowych aktualne + historia



Dane ze stacji pomiarowych

Repertuar Kin
sprawdź go przez komórkę

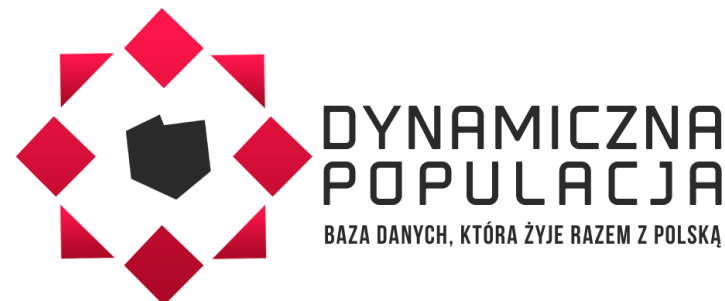
 m.kina.wp.pl

sprawdź teraz

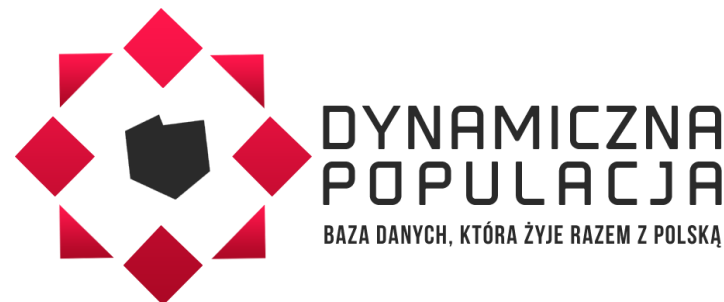
Dane o filmach + repertuar + symulacja



Problemy



- ▶ Ograniczone zasoby serwera
- ▶ „Ucieczka” zaawansowanych uczestników projektu
- ▶ Bariery czasowe dla uczestników
- ▶ ... więcej „grzechów” nie pamiętamy 😊



Co z tego wyniknie – nie wiemy 😊

Dla studentów jest to doskonała platforma zapoznania się z problemami „dużych ilości danych”

Tworzenie hurtowni danych

Wykorzystanie systemu do zadań Data Mining

Wykorzystanie do prac badawczych

„Kopalnia tematów do prac dyplomowych”

NIE ZNALEŹLIŚMY INFORMACJI O PODOBNYM SYSTEMIE