BIK

BIURO INFORMACJI KREDYTOWEJ

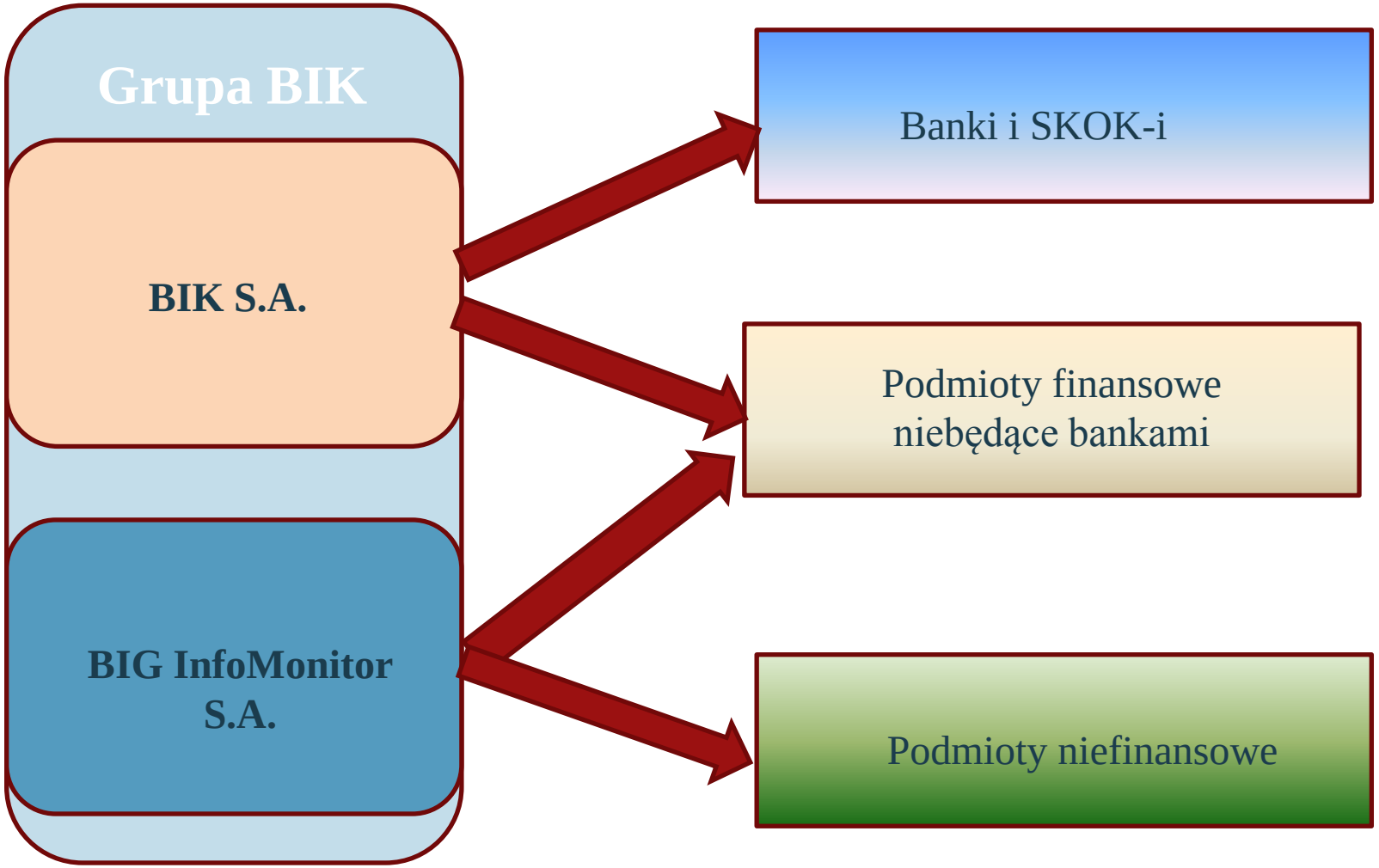
Czyszczenie i standaryzacja danych adresowych

Michał Słoniewicz, Biuro Informacji Kredytowej

Warszawa, 19 kwietnia 2012 r.

LAUREAT KONKURSU
TERAZ POLSKA

Współpraca z Grupą BIK



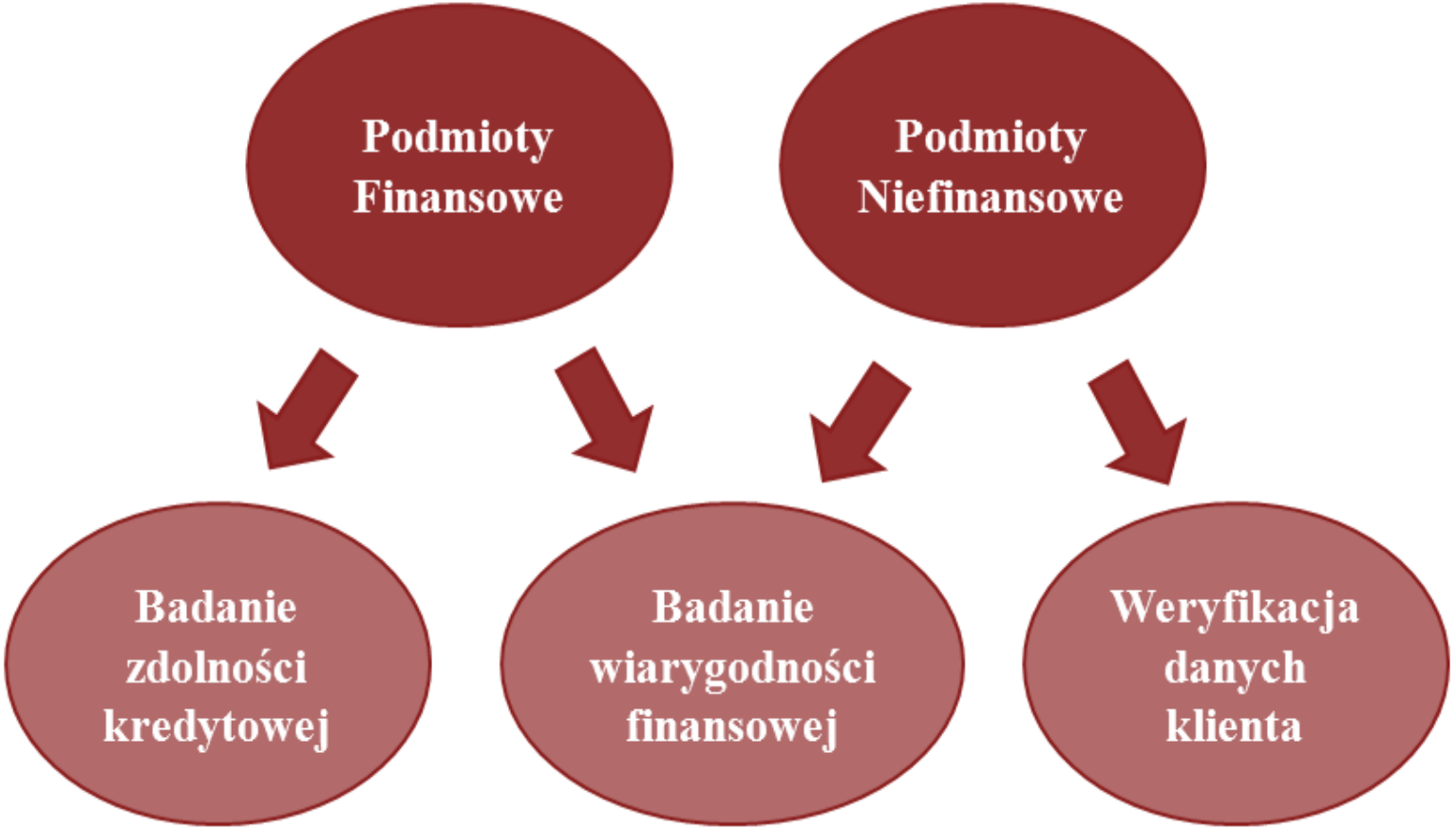
Oferta Grupy BIK

Oferta Grupy BIK skierowana jest do:

- ✓ Banki i SKOK-i
- ✓ Podmiotów finansowych, w szczególności:
  - firm leasingowych,
  - firm udzielających pożyczek,
  - firm udzielających kredytów konsumenckich.

- ✓ Podmiotów niefinansowych, które z uwagi na charakter prowadzonej działalności gospodarczej, mogą być w istoty sposób zainteresowane produktami Grupy BIK (np. firmy telekomunikacyjne).

Korzyści ze współpracy z Grupą BIK



Korzyści ze współpracy z Grupą BIK

- Dostęp do danych niezbędnych dla oceny ryzyka kredytowego konsumenta
- Lepsza ocena wiarygodności finansowej klienta
- Dodatkowe źródło potwierdzenia danych identyfikacyjnych klienta

Czyszczenie danych adresowych – tło

Braki, **błędy**, **niespójności** oraz **brak standaryzacji** w danych adresowych zapisanych w bazie danych SI BIK - KI, to czynniki, które nie pozwalają lub w istotnym stopniu ograniczają potencjalne możliwości użycia w/w danych w produktach i usługach wykorzystujących przeszukiwanie, porównywanie i identyfikację danych adresowych oraz w analizach przestrzennych.

Jednym z głównych czynników warunkujących poprawne funkcjonowanie w/w produktów i usług jest zaimplementowanie w BIK narzędzi i procesów czyszczenia danych adresowych.

Zasady przekazywania danych adresowych do BIK

Dwie możliwości:

- Dane **zagregowane** (miejscowość, ulica, nr domu i nr lokalu w jednym polu + osobne pole dla kodu pocztowego),

- Dane **w osobnych polach** (miejscowość, ulica, nr domu i nr lokalu, kod pocztowy).

Przykłady – dane przekazywane do BIK

```
GNIEZNO SZCZYTNIKI DUCHOWNE
POZNAŃ UL.JANA HENRYKA DĄBR
WŁOCŁAWEK STARODĘBSKA 0
WEJHEROWO  WIERZBOWA
TFUTD6UID  0
TYCHY  28
----------------------------------------
KRAPKOWICE UL.KILIŃSKIEGO 29M/21
DĄBROWA TARNOWSKA ŚW. BRATA CHMIELOWSKIEGO 32
WARSZAWA GROCHOWSKA 309/317 M 40 309/317
```

Definicje

- **Źródłowe bazy** - bazy danych BIK, z których pochodzą dane adresowe. W bazach źródłowych adresy pozostają w postaci nie wystandaryzowanej.

- **Standaryzacja** - czynność w procesie czyszczenia polegająca na parsowaniu i poprawieniu adresu w oparciu o słownik adresów i ujednoliceniu formy zapisu. Standaryzacja jest płaska i nie zależy od innych adresów klienta. W jej wyniku powstaje wystandaryzowany adres.

- **Uzupełnianie adresu** - czynność w procesie czyszczenia polegająca na uzupełnieniu braków w adresie na podstawie innych adresów klienta.

Definicje c.d.

- **Wyczyszczony adres** - adres, który przeszedł wszystkie kroki procesu czyszczenia (standaryzacja i uzupełnianie danych adresowych, wskazanie duplikatów), niezależnie od nadanego mu statusu czy jest poprawny czy nie. Był czyszczony i uzupełniany w ramach wszystkich adresów podmiotu danego typu. Wyczyszczony adres jest zawsze w kontekście klienta, wystandaryzowany – nie.

- **Referencyjna baza adresów** - baza wyczyszczonych adresów połączona referencją z danymi źródłowymi.

- **„Złoty" adres** – wybrany spośród wszystkich wyczyszczonych adresów klienta w danym kontekście biznesowym.

Na przykład…

DANE ŹRÓDŁOWE DLA ADRESÓW TEGO SAMEGO KLIENTA

ID ADRESU	KOD POCZTOWY	MIEJSCOWOŚĆ	ULICA	NUMER DOMU
1*	02-679 WARZSAWA UL. Modzelewskiego 77			
2*	02-679 WARSZAWA Modzelewskiego 77			
3	02-679	WARZSAWA	Modzeleswkiego	77
4	02-670	WARSZAWA	Z. Modzelewskiego	77
5	02-679	WARSZAWA	UL. Modzelewskiego	
*dane zagregowane				

DANE ŹRÓDŁOWE DLA ADRESÓW TEGO SAMEGO KLIENTA

ID ADRESU	KOD POCZTOWY	MIEJSCOWOŚĆ	ULICA	NUMER DOMU
1*	02-679 WARSZAWA UL. Modzelewskiego 77	WARSZAWA	UL. Z. Modzelewskiego	77
2*	02-679 WARSZAWA Modzelewskiego 77	WARSZAWA	Modzelewskiego	77
3	02-679	WARZSAWA	Modzeleswkiego	77
4	02-670	WARSZAWA	Z. Modzelewskiego	77
5	02-679	WARSZAWA	UL. Modzelewskiego	

PO PARSOWANIU
PO STANDARYZACJI

ID ADRESU	ID ADRESU	KOD POCZTOWY	KOD POCZTOWY	MIEJSCOWOŚĆ	ULICA	NUMER DOMU	TERYT
1	1	02-679	02-679	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX
2	2	02-679	02-679	WARSZAWA	ul. Modzelewskiego	77	XXXX
3	3	02-679	02-679	WARSZAWA	ul. Modzelewskiego	77	XXXX
4	4	02-679	02-670	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX
5	5	02-679	02-679	WARSZAWA	ul. Modzelewskiego		XXXX

Na przykład cd…

PO STANDARYZACJI

ID ADRESU	KOD POCZTOWY	MIEJSCOWOŚĆ	ULICA	NUMER DOMU	TERYT
1	02-679			77	XXXX
2	02-679			77	XXXX
3	02-679			77	XXXX
4	02-679			77	XXXX
5	02-679				XXXX

Adres Złoty Adresu Zamieszkania klienta id1

ID ADRESU	1
KOD POCZTOWY	02-679
MIEJSCOWOŚĆ	WARSZAWA
ULICA	ul. Z. Modzelewskiego
NUMER DOMU	77
TERYT	1465058 13192
ID KLIENTA	id1
ROLA	Adres Zamieszkania
WSKAŹNIK	ZŁOTY

ID ADRESU	KOD POCZTOWY	MIEJSCOWOŚĆ	ULICA	NUMER DOMU	TERYT
1	02-679			77	XXXX
2	02-679			77	XXXX
3	02-679			77	XXXX
4	02-679			77	XXXX
5	02-679			77	XXXX

ID ADRESU	KOD POCZTOWY	MIEJSCOWOŚĆ	ULICA	NUMER DOMU	TERYT	ID KLIENTA	ROLA	WSKAŹNIK
1	02-679	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX	id1	Adres Zamieszkania	ZŁOTY
2	02-679	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX	id1	Adres Zamieszkania	DUPL
3	02-679	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX	id1	Adres Zamieszkania	DUPL
4	02-679	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX	id1	Adres Zamieszkania	DUPL
5	02-679	WARSZAWA	ul. Z. Modzelewskiego	77	XXXX	id1	Adres Zamieszkania	DUPL

Statystyka przetwarzania

Opis		Wartość
OGÓLNE		
wszystkich adresów przed przetwarzaniem (źródłowe)		103 mln
wszystkich adresów po przetwarzaniu (źródłowe + powstałe w wyniku czyszczenia)		135 mln
po przetwarzaniu adresów doszło (powstałe w wyniku czyszczenia)		32 mln
STANDARYZACJA		
liczba adresów, które udało się wystandaryzować		98 mln
ADRESY CZYSTE – PO UZUPEŁNIENIU		
liczba adresów czystych (po uzupełnieniu)		28 mln
liczba adresów czystych (po uzupełnieniu) w zależności od roli adresowej	ID_ROLI_ADR	LICZBA
	Poprzedni adres zamieszkania (Z)	60 000
	Adres korespondencyjny (R)	170 000
	Adres zamieszkania (Z)	14 400 000
	Adres zamieszkania (R)	20 920 000
	Adres korespondencyjny (Z)	600 000
	Adres zatrudnienia (Z)	320 000
	Adres zatrudnienia (R)	4 290 000

Dziękuję za uwagę

LAUREAT KONKURSU
TERAZ POLSKA