

Dane internetowe w europejskiej statystyce publicznej – doświadczenia projektu ESSnet Web Intelligence Network

Klaudia Peszat, Dominik Dąbrowski, Dominika Nowak

Trusted Smart Statistics – Web Intelligence Network



**Web Intelligence
Network**



**Funded by
the European Union**

Nowe źródła danych w statystyce publicznej

- Nowe informacje w wielu obszarach społeczno-gospodarczych
- Większa szczegółowość danych
- Bardziej aktualne dane (w tym analiza w czasie rzeczywistym)
- Zmniejszenie obciążenia respondentów
- Redukcja kosztów badań
- Odpowiedź na rosnące potrzeby użytkowników

NOWE ŹRÓDŁA DANYCH  WDROŻENIE DO PRODUKCJI STATYSTYCZNEJ

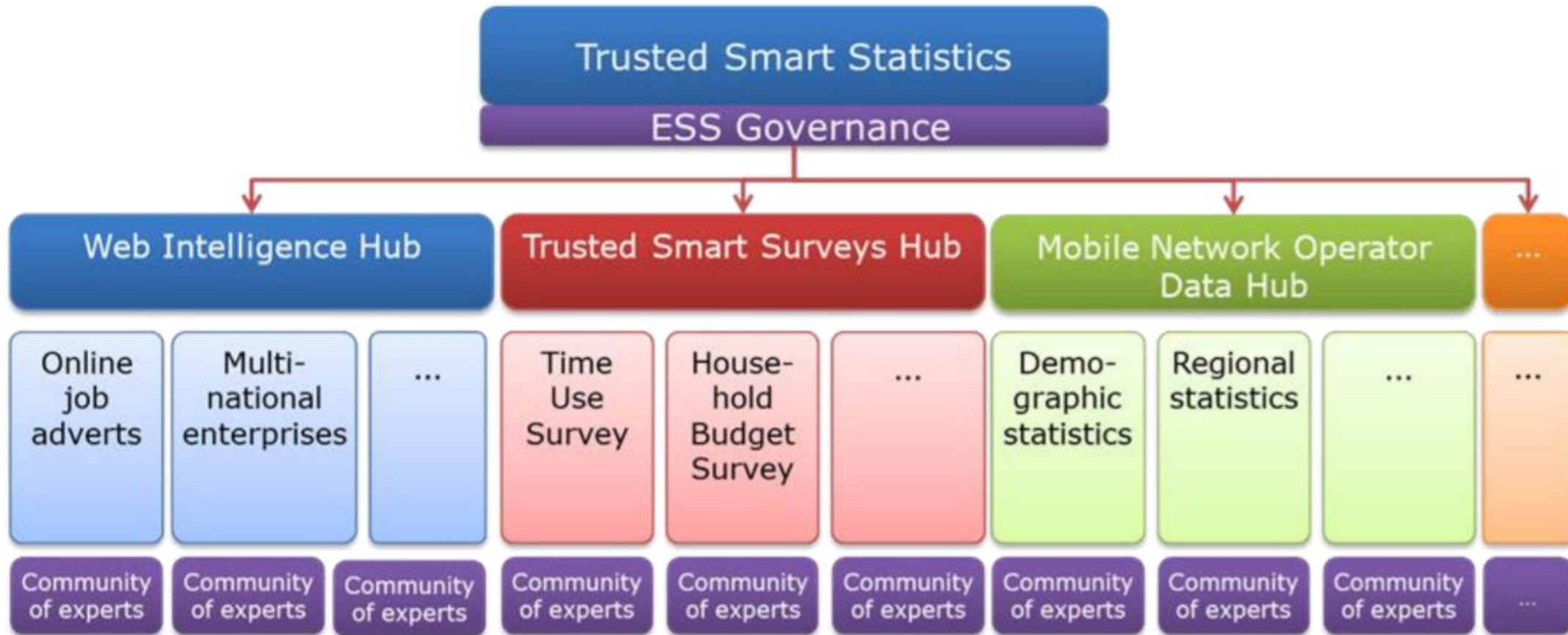


Web Intelligence
Network

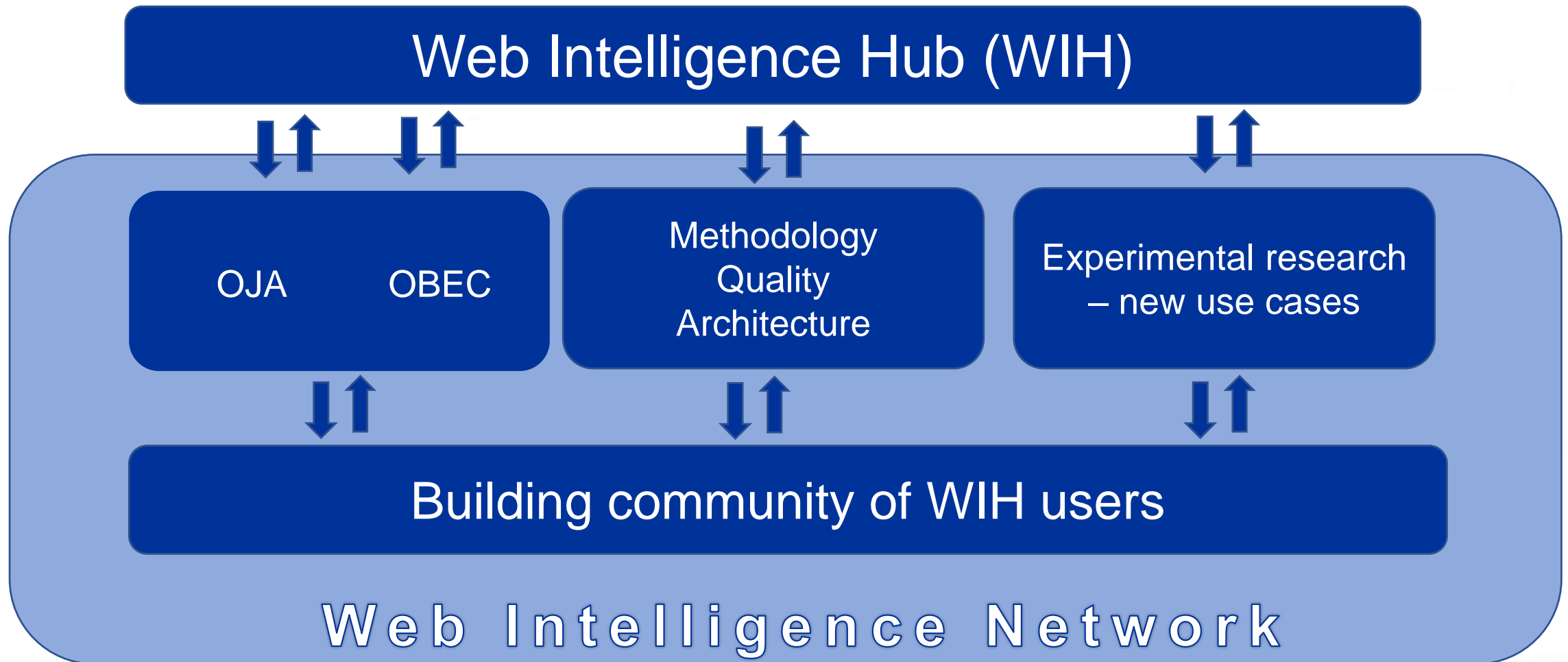


Funded by
the European Union

Dane internetowe w europejskiej statystyce publicznej



Projekt ESSnet Web Intelligence Network (WIN)



Budowanie społeczności użytkowników WIH

Dotarcie do wszystkich krajów ESS

Wykorzystanie środowiska WIH

Integracja WIH z krajowymi systemami produkcji



Web Intelligence
Network



Funded by
the European Union

Potencjalne dziedziny w Web Intelligence Hub

Najbardziej zaawansowane dziedziny

Internetowe oferty pracy (OJA)

Charakterystyki przedsiębiorstw
(OBEC)

Nowe dziedziny

Rynek nieruchomości

Nowe budownictwo

Internetowe ceny sprzętu AGD,
sprzętu audiowizualnego,
fotograficznego i przetwarzającego
informacje

Turystyka

Poprawa jakości rejestrów
statystycznych

Szybkie szacunki ekonomiczne



Potencjalne dziedziny w Web Intelligence Hub

Najbardziej zaawansowane dziedziny

Internetowe oferty pracy (OJA)

Charakterystyki przedsiębiorstw (OBEC)

Nowe dziedziny

Rynek nieruchomości

Nowe budownictwo

Internetowe ceny sprzętu AGD, sprzętu audiowizualnego, fotograficznego i przetwarzającego informacje

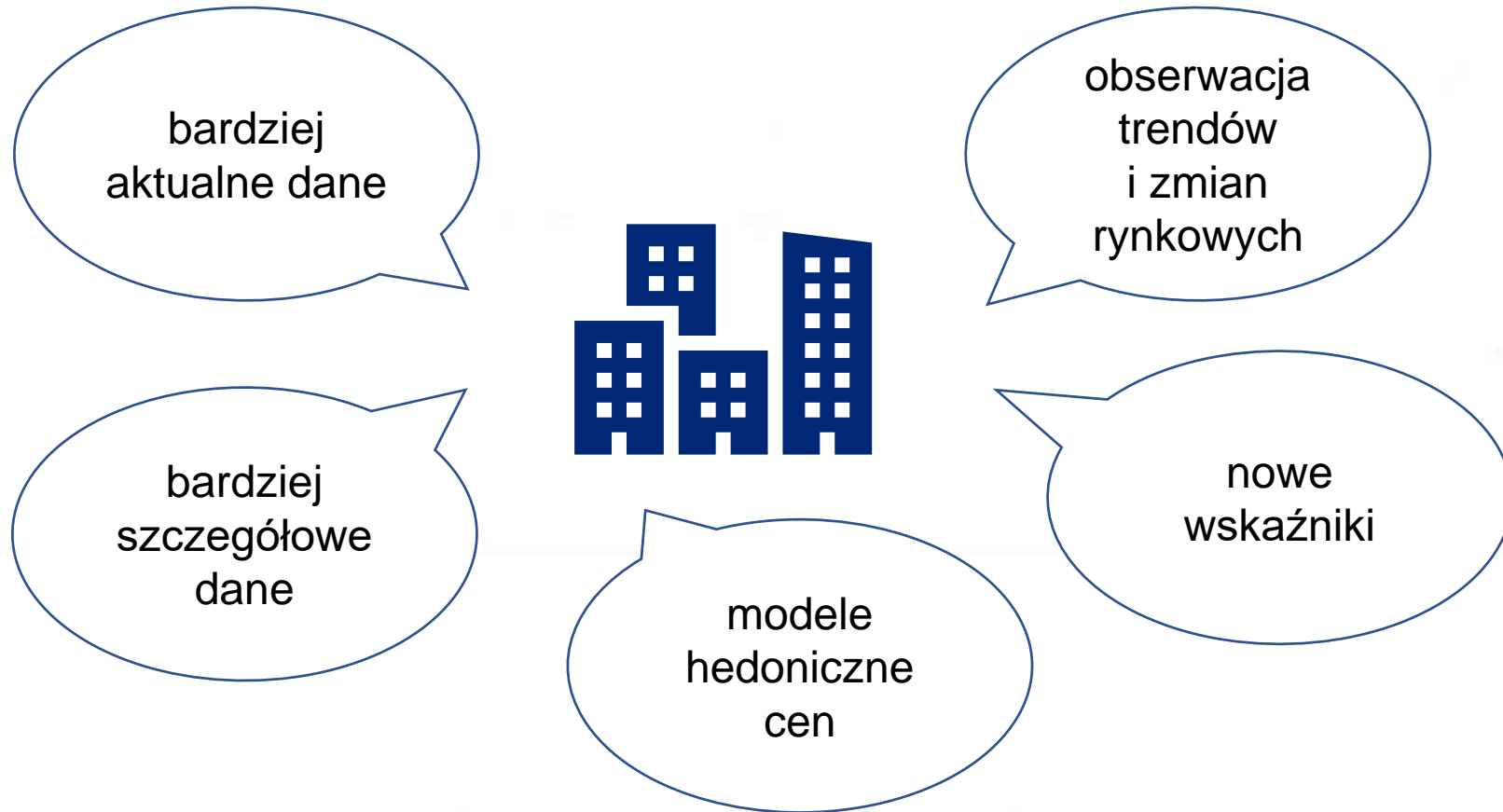
Turystyka

Poprawa jakości rejestrów statystycznych

Szybkie szacunki ekonomiczne



Zastosowanie danych internetowych w statystyce nieruchomości



Główne etapy procesu badawczego

- Eksploracja nowych źródeł danych
- Przygotowanie oprogramowania
- Pozyskiwanie danych
- Przetwarzanie danych
- Modelowanie i interpretacja danych
- Publikacja statystyk eksperymentalnych



Web Intelligence
Network



Funded by
the European Union

Główne etapy procesu badawczego

- Eksploracja nowych źródeł danych
- Przygotowanie oprogramowania
- Pozyskiwanie danych
- Przetwarzanie danych
- Modelowanie i interpretacja danych
- Publikacja statystyk eksperymentalnych



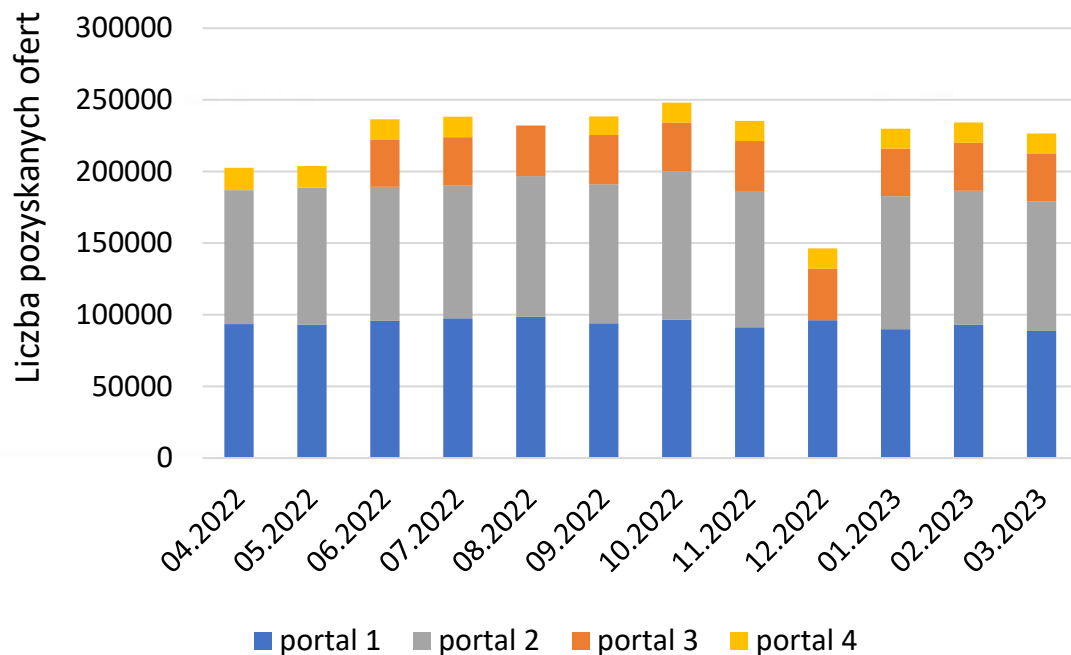
Wybór internetowych źródeł danych – kryteria

- ✓ znaczenie źródła internetowego (liczba ofert, udział w rynku)
- ✓ ograniczenia techniczne (np. blokowanie robotów, robots.txt)
- ✓ mechanizmy poruszania się po stronie oraz struktura strony
- ✓ kompletność informacji
- ✓ aktualność treści

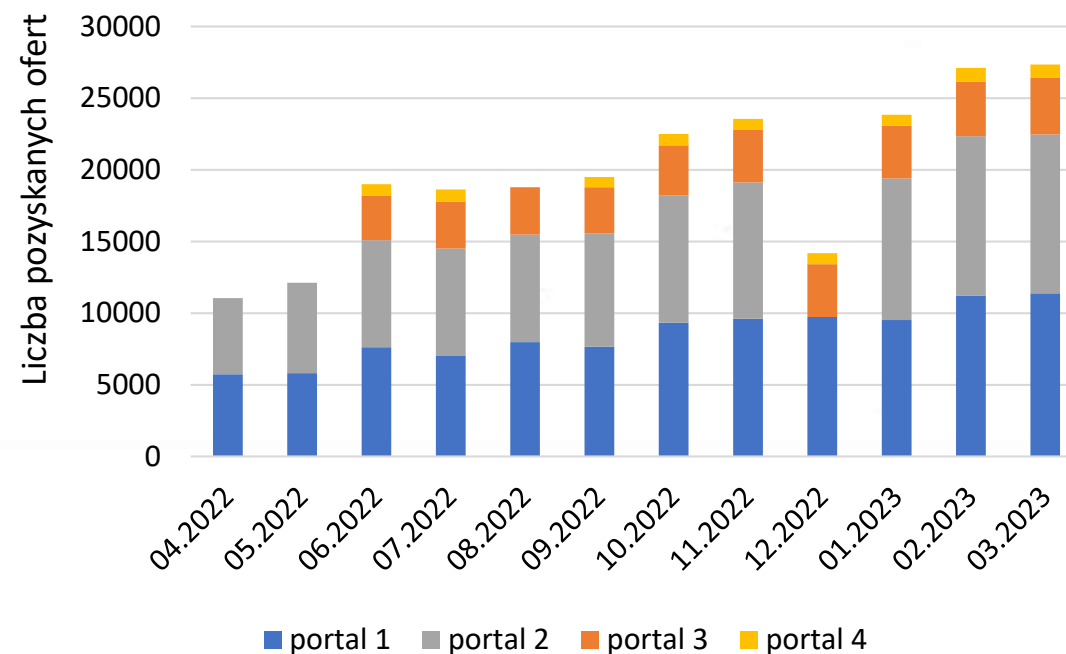


Stabilność internetowych źródeł danych w czasie

Oferty mieszkań na sprzedaż



Oferty mieszkań na wynajem



Wstępna ocena jakości danych

Wskaźnik	Opis	Wartość
Wielkość próby	Liczba ofert na portalu	102566
Braki odpowiedzi	Liczba ofert z negatywnym kodem odpowiedzi	2487
Pobrane oferty	Liczba ofert z pozytywnym kodem odpowiedzi	100079
Braki danych	Liczba pustych lub nieprawidłowo wypełnionych wartości*	20089
Procent braków danych	Procent pustych lub nieprawidłowo wypełnionych wartości*	4,01%

* w zmiennych obowiązkowych

Źródło: Portal 1, oferty sprzedaży i najmu, marzec 2023.



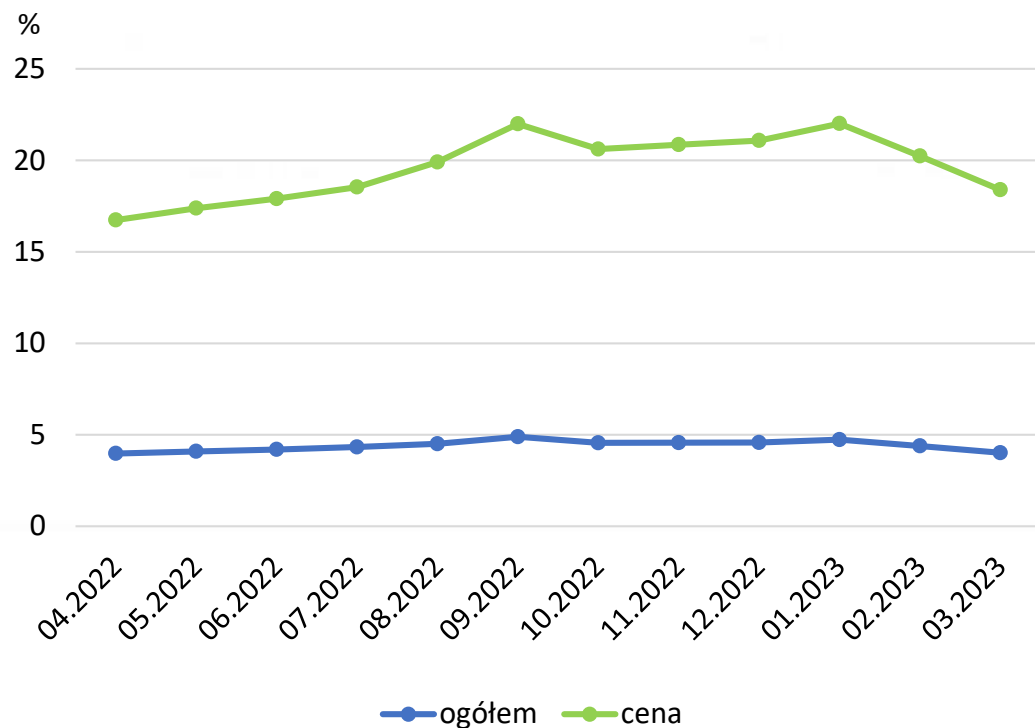
Web Intelligence
Network



**Funded by
the European Union**

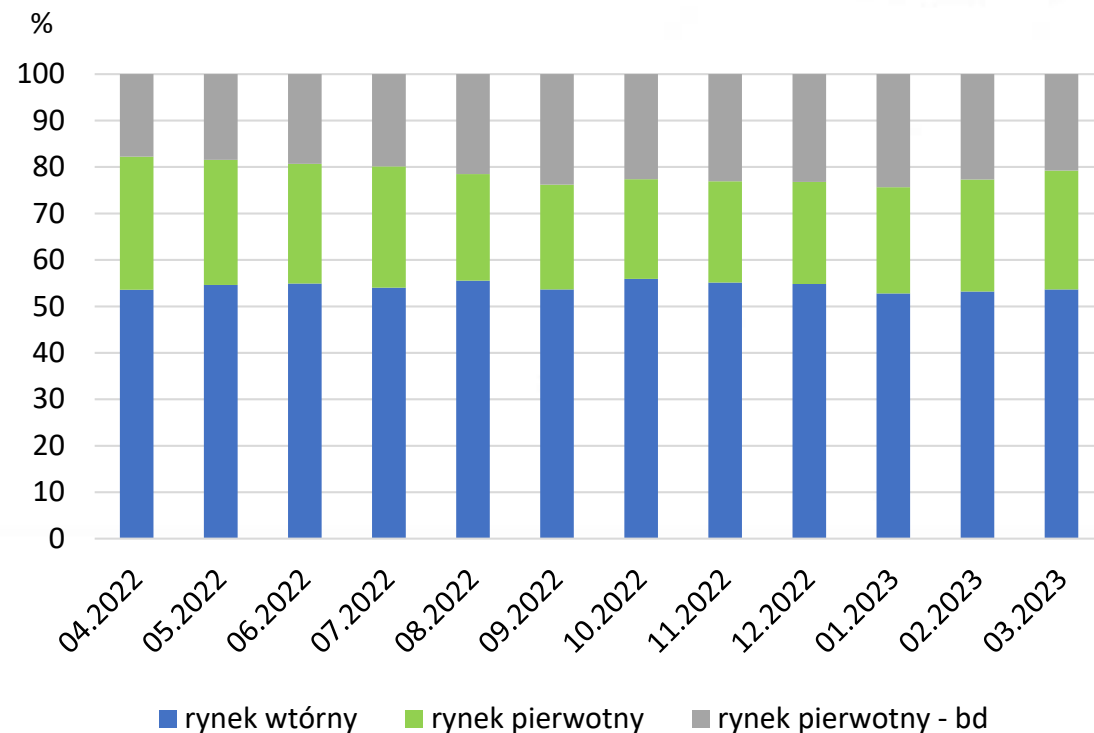
Braki danych

Udział braków danych



Źródło: Portal 1, oferty sprzedaży i najmu.

Udział braków danych w zmiennej cena wg typu rynku



Źródło: Portal 1, oferty sprzedaży.



Web Intelligence
Network



Funded by
the European Union

Główne etapy procesu badawczego

- Eksploracja nowych źródeł danych
- Przygotowanie oprogramowania
- Pozyskiwanie danych
- Przetwarzanie danych
- Modelowanie i interpretacja danych
- Publikacja statystyk eksperymentalnych



Zagadnienia metodologiczne

- Różnica między cenami ofertowymi i transakcyjnymi
- Niezdefiniowana populacja (brak referencyjnego źródła dla rynku najmu)
- Duplikacja ofert (w ramach jednego portalu, jak i pomiędzy różnymi portalami)
- Multioferty (mieszkania na rynku pierwotnym)
- Braki danych (np. cena mieszkań na rynku pierwotnym)



Wyzwania



Stabilny dostęp do źródeł danych (web-scraping, dostawcy danych)



Ocena jakości źródeł danych, procesu przetwarzania, wyników



Zapewnienie porównywalności danych na poziomie europejskim



Utworzenie infrastruktury IT w ramach urzędu



Budowanie odpowiednich umiejętności i kompetencji pracowników oraz współpraca z ekspertami merytorycznymi



Dziękuję za uwagę!

Więcej informacji o projekcie:



CROS: https://ec.europa.eu/eurostat/cros/WIN_en



Twitter: <https://twitter.com/EssnetWin>



LinkedIn: <https://www.linkedin.com/company/essnet-project-web-intelligence-network/>

Kontakt:

Klaudia Peszat: k.peszat@stat.gov.pl

Dominik Dąbrowski: d.dabrowski@stat.gov.pl

Dominika Nowak: do.nowak@stat.gov.pl (koordynator projektu)



Web Intelligence
Network



Funded by
the European Union