

# Tapping into web data for European statistics – challenges and experiences of the ESSnet Web Intelligence Network

UNECE Expert Meeting on Statistical Data Collection and Sources

May 22, 2024

Klaudia Peszat, Dominika Nowak (Statistics Poland)



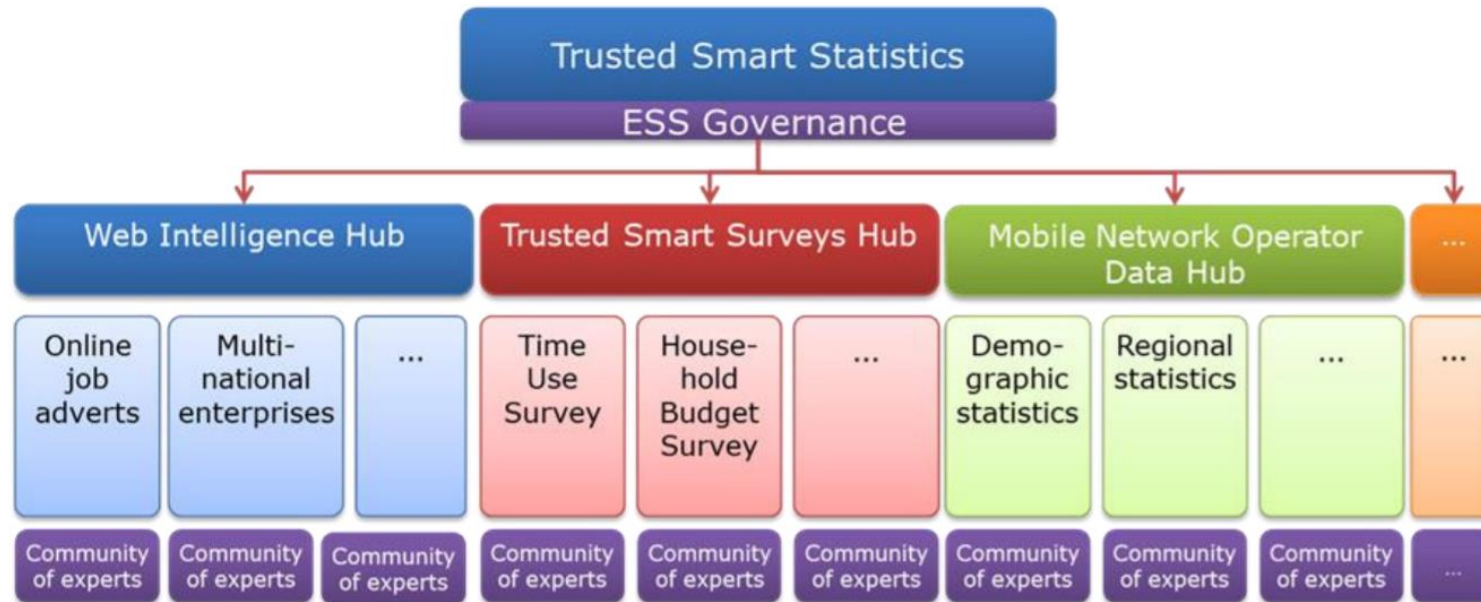
**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Web Intelligence Hub (WIH)

- Started as a concept – evolved towards tangible tools
- Web data acquisition, processing & analysis environment
- Centralized, shared system, pan-European platform



# Rationale behind the Web Intelligence Hub (WIH)

- Different capacity of NSIs across Europe to use web data
- Different competency levels, scarcity of data science skills
- Infrastructure with big data capabilities required
- More efficient use of resources
- Overcoming technical/legal issues with accessing web data

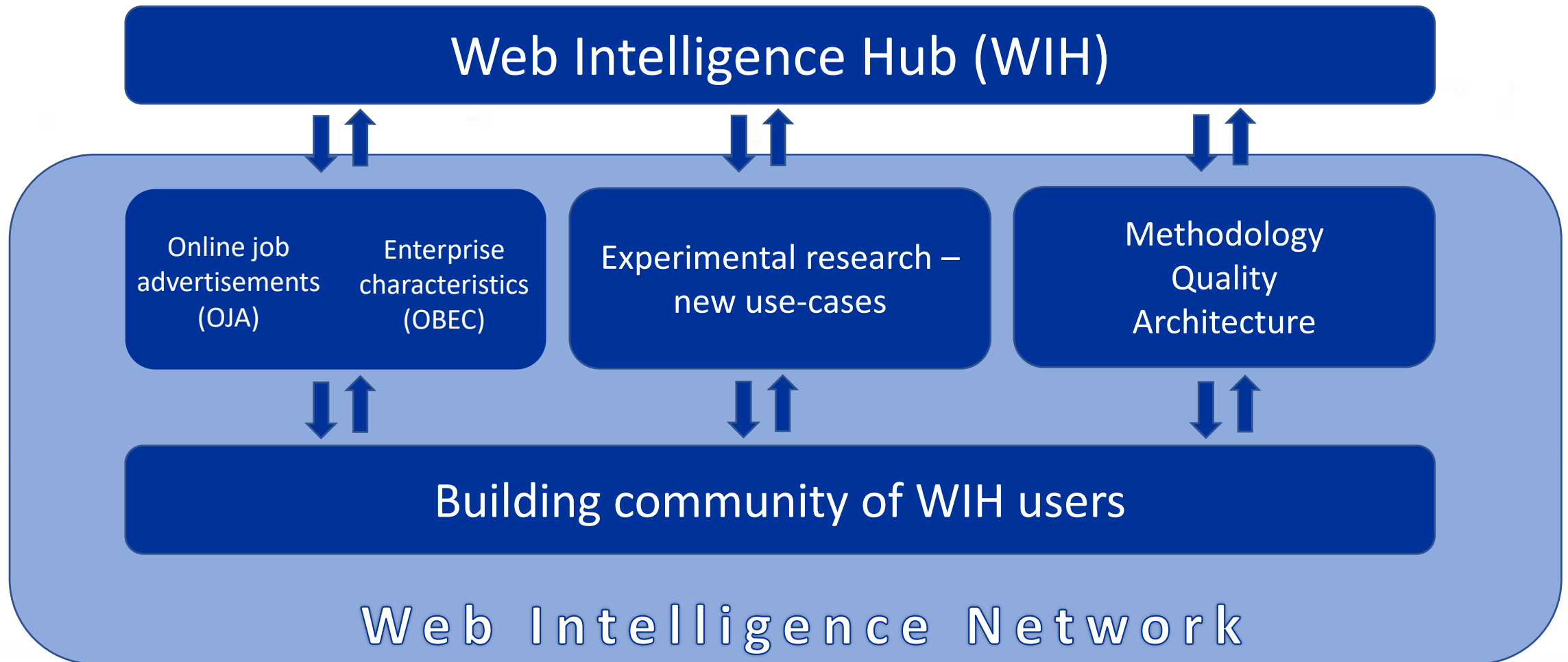


**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# The role of the Web Intelligence Network (WIN)



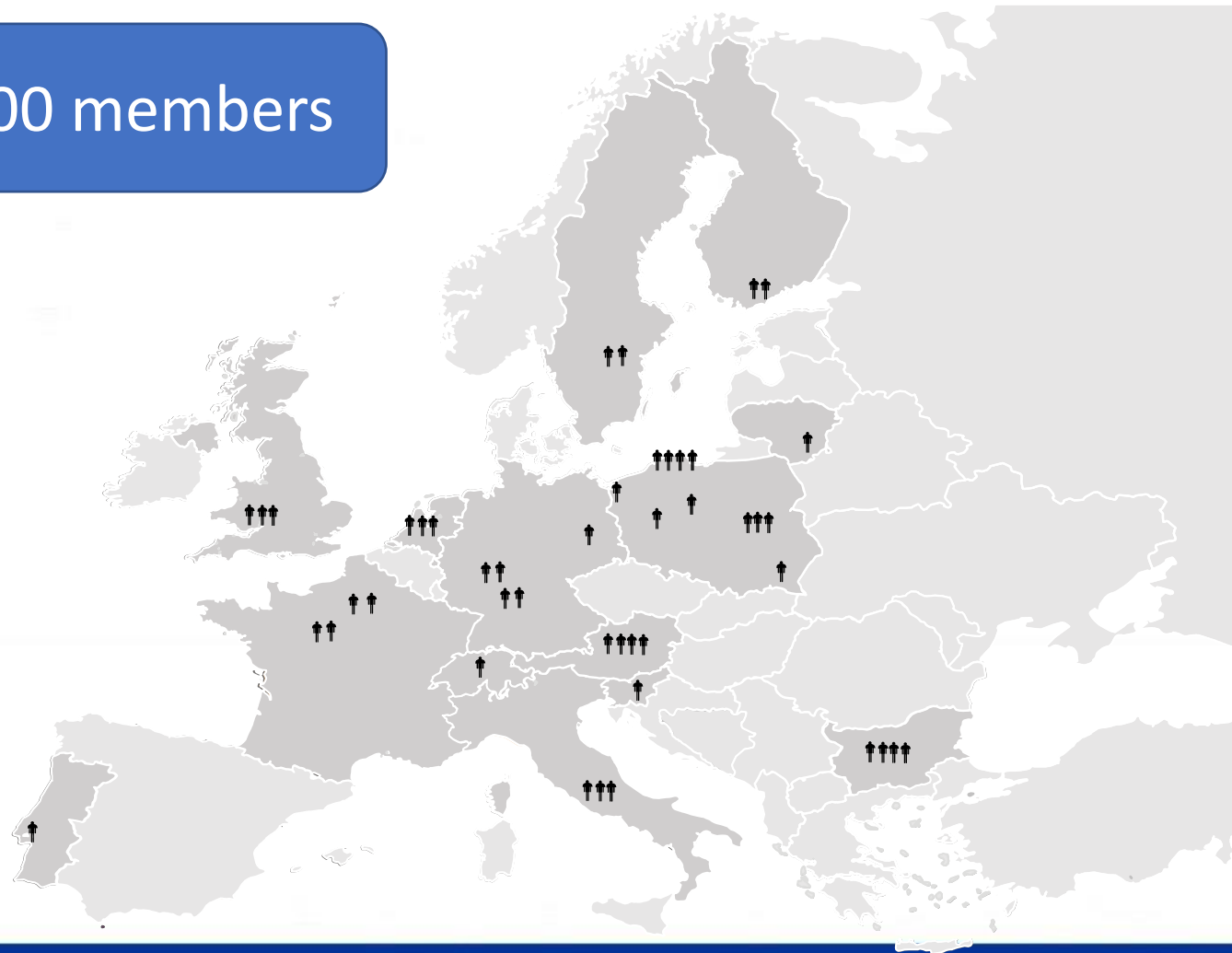
# Web Intelligence Network (WIN)

14 countries, 17 organizations, ~100 members

Contribute to the development  
of the Web Intelligence Hub

Reach out to **all ESS countries**

Use web data, use the WIH



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Topics WIN is looking to

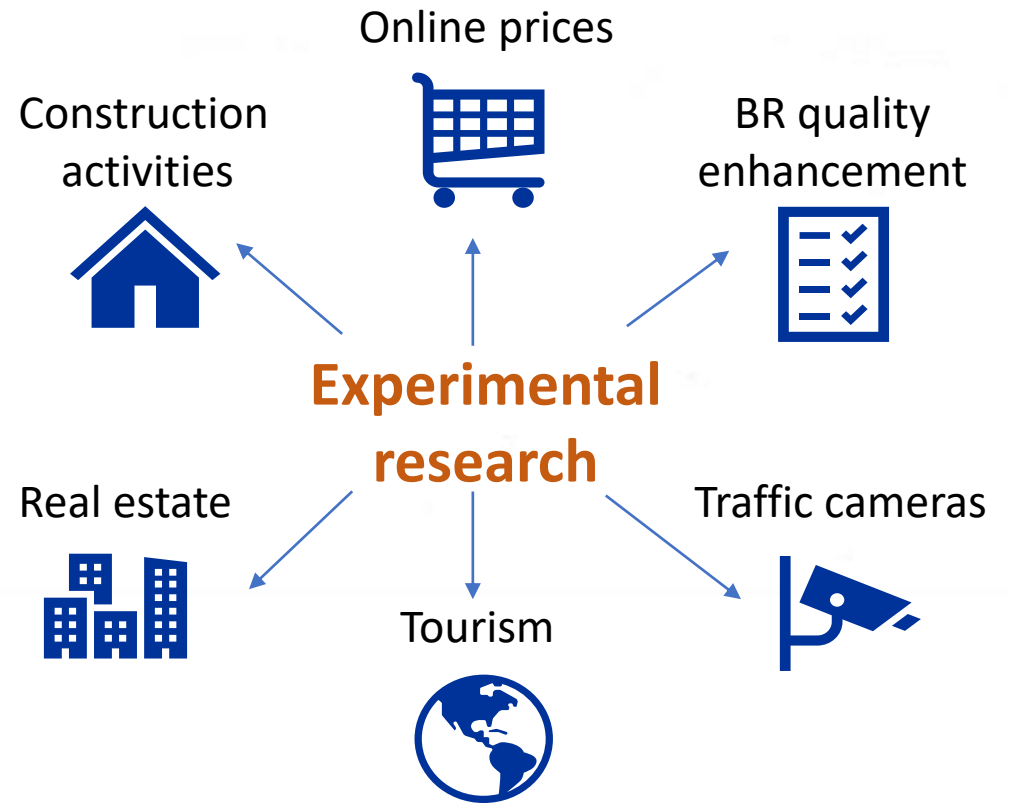
## Most mature use-cases



Online job advertisements (OJA)



Enterprise characteristics (OBEC)



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# OJA – data classification quality

occupation1d	AT	BG	FI	FR	IT	PL	SI
NA	0	4 / 3.6	2.4 / 2.7	2.3 / 1.9	2.3 / 2.9	0	12 / 13
correct	49 / 52.7	48.5 / 52.7	62.4 / 63.8	63.3 / 59.7	61.9 / 61	44.7 / 50.1	50.5 / 53.2
incorrect	51 / 47.3	47.5 / 43.8	35.2 / 33.5	34.3 / 38.4	35.8 / 36.2	55.3 / 49.9	37.5 / 33.8

working_time	AT	BG	FI	FR	IT	PL	SI
NA	0	13.5 / 12.7	0.3 / 0.4	4.3 / 4.7	3.2 / 3.4	0.2 / 0.1	17.9 / 17.1
correct	75.2 / 73.3	62 / 67	67.9 / 76	49 / 54.7	61.6 / 65.5	50.7 / 58.8	55.1 / 67.5
incorrect	24.8 / 26.7	24.6 / 20.3	31.8 / 23.5	46.7 / 40.6	35.2 / 31.1	49.2 / 41.1	26.9 / 15.4

\* unweighted / weighted



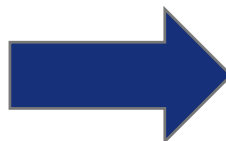
**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Tourism – detection of duplicates

## The key-point detection



370 good matches from 1786 all matches



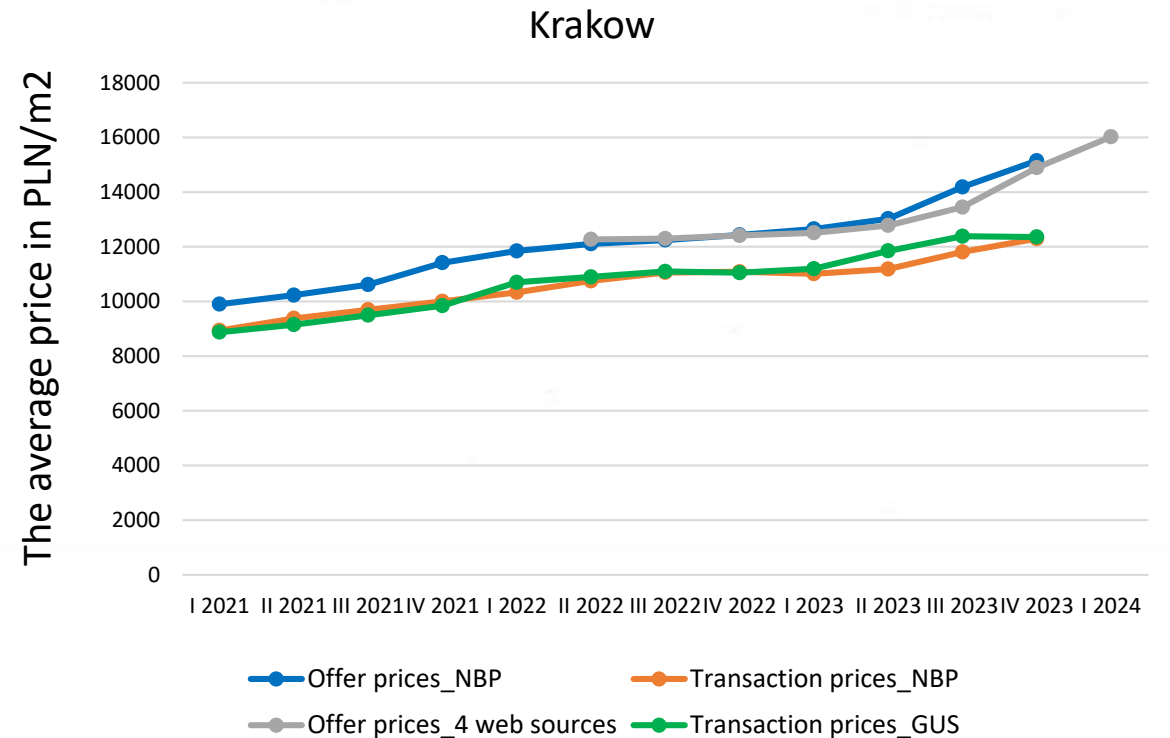
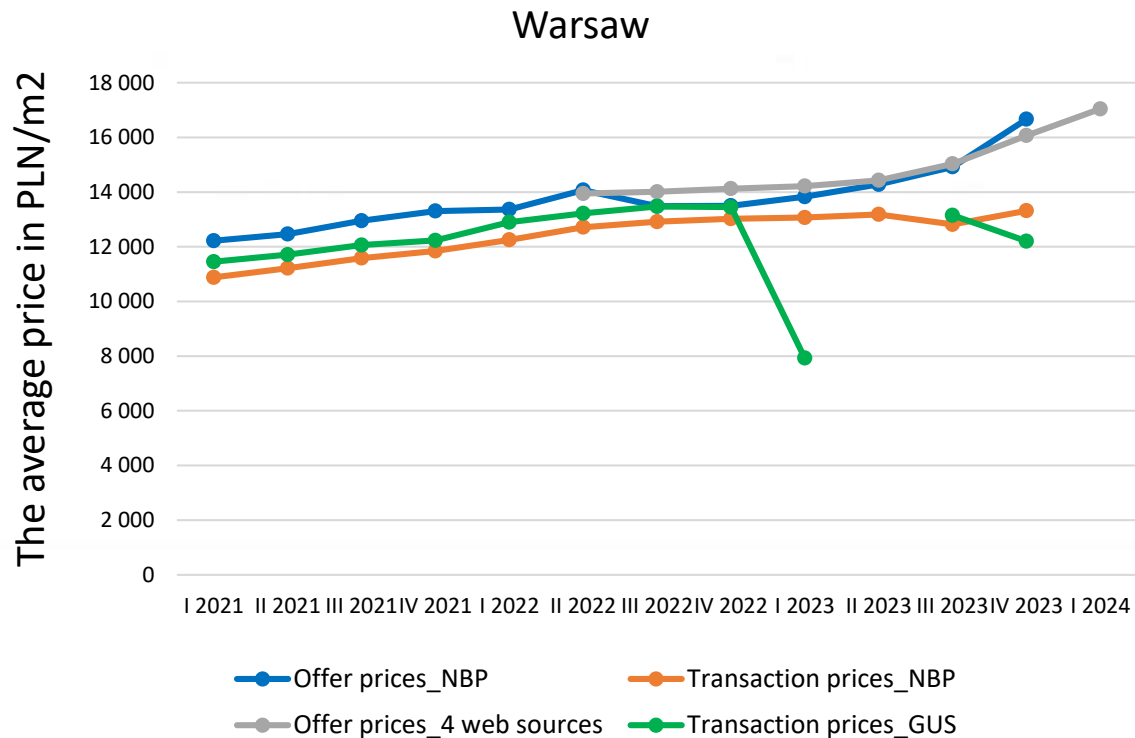
Web Intelligence  
Network



Funded by  
the European Union



# Real estate – comparison between web data and official statistics



# Issues WIN is dealing with

- **Quality assessment:**
  - Data source selection and stability
  - Quality of the data (e.g. over and under coverage, de-duplication, dealing with missing data etc.)
  - Quality of data classification (e.g. ISCO, NACE, NUTS)
- **Legal / organizational constraints:**
  - Legal grounds for using register data in a web scraping pipeline
  - Agreements with private companies, purchasing data from third parties
- **Technical:**
  - WIH – still in the development stage



# Building community of WIH and web data users

Meet us at  
conferences

Visit us on-line

Look for our training,  
webinars, tutorials

**NTTS conferences**



Statistical  
Business  
Registers

Tourism data

Online real  
estate market

Architecture,  
methodology  
and quality

Construction  
activity

Web  
Intelligence  
in Practice -  
OBEC

OJA Training  
for WIN and  
WISER

And more...

Join WISER  
(Web  
Intelligence  
uSERs  
Group)



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Thank you.

[k.peszat@stat.gov.pl](mailto:k.peszat@stat.gov.pl) – Klaudia Peszat

[do.nowak@stat.gov.pl](mailto:do.nowak@stat.gov.pl) – Dominika Nowak



**Web Intelligence**  
Network



**Funded by  
the European Union**