New Use-cases of web data for official statistics

Web intelligence Network Conference
From Web to Data
4 Feb 2025, Gdansk

Trusted Smart Statistics – Web Intelligence Network
Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

WIN work package 3 on new use cases

- Exploration of ‘new’ web data sources for the production of official statistics, as primary or auxiliary datasource
- 6 use cases (UCs):

• **UC1** Characteristics of the real estate market	**PL**, BG, DE-HSL/BBB, FI, FR
• **UC2** Construction activities	**DE-HSL**, DE-BBB, SE
• **UC3** Online prices of household appliances and audio-visual, photographic and information processing equipment	**SE,** BG
• **UC4** Experimental indices in tourism statistics (hotel prices)	**PL,** BG
• **UC5** Business register quality enhancement	**NL,** AT, DE-HSL, SE, FI
• **UC6** Faster Economic Indicators using new data sources	**SE,** UK

Challenges

Weighting issues

Hedonic prices

Deduplication

Missing data

Find the right urls / texts

ML models

Search-engine leakage

API legals

Changed privacy concerns

Lack of data

Lack of standardisation among sites

Duplicate offers

Coverage

Stability

UC1: characteristics of the real estate market (1)

Aim:

- to monitor the ***real estate market*** that responds ***quickly*** to the economic cycles and is ***not fully covered by administrative data***

Potential use:

- Flash estimates
- In-depth research, hedonic indices, modeling real estate characteristics

Landscaping phase:

- Each candidate portal was assessed using a standardised ***checklist*** for assessment of data sources

- Consistent indicators across countries :

Common basic indicators:

- number of offers
- average price per square meter (sale)
- share of offers by price per square meter classes (sale)
- average surface area in square meter
- share of offers by surface area classes
- average number of rooms
- share of offers by number of rooms
- average price (rent)
- share of offers by price classes (rent)

UC1: characteristics of the real estate market (2)

Monitor trends

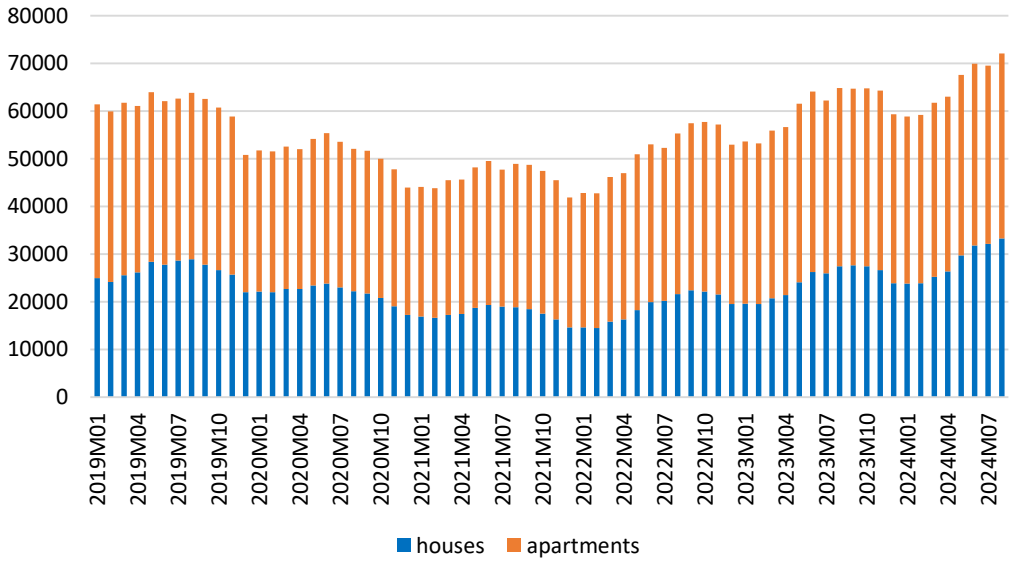


Figure 30. Number of offers for sale on Oikotie 2019M01 to 2024M08

Some sources had good coverage and were stable over time. Yearly cycles in real estate sales market are visible (less offers in winter holiday seasons).
The COVID pandemic in 2020 to 2021 decreased the numbers of sale offers considerably and the pre-pandemic levels were reached again in 2023

UC1: characteristics of the real estate market (3)

- Countries have shown ***high stability*** level of web data and good level of ***reflecting the current market situation*** in real estate

- The data that have been acquired can certainly be ***a supplement*** to the current real estate market monitoring system or be used to create ***leading indices*** where the data flow in classic researches is long.

- These data can also be used to build ***hedonic indices*** *or* ***models*** classifying real estate in new cross-sections

Deliverable 3.4: Report on the results of the new data sources exploration and the conditions for using the data

Characteristics of the real estate market

UC2 construction activities

Measuring construction activities using advertisements, Tobias Gramlich, session IV

Next presentation !

UC3: online prices of household appliances and audio-visual, photographic and information processing equipment

Results:

- ***Weekly*** price collection in ***Bulgaria*** (4 data sources initially, later 3) and ***Sweden*** (2 online sources) to calculate average prices and ***price indices*** over ***longer period*** (>2 yrs) for specific product categories (blenders, steam irons, coffee machines, TVs, washing machines).
- Swedish data integration pilot on combining ***web data*** with ***cash register*** data to estimate ***weights*** of products, products groups or COICOPs, where this information is not available.
- Proof of concept executed on laptops and TVs and cash register data from 2 companies.
- Open issues: weight estimation algorithm, bias in web data, scaling up, scaling up of method, timing of CPI data versus cash register dataflow,
- Further input welcome!

Deliverable 3.7:
UC3: Report on methodology and results for online prices

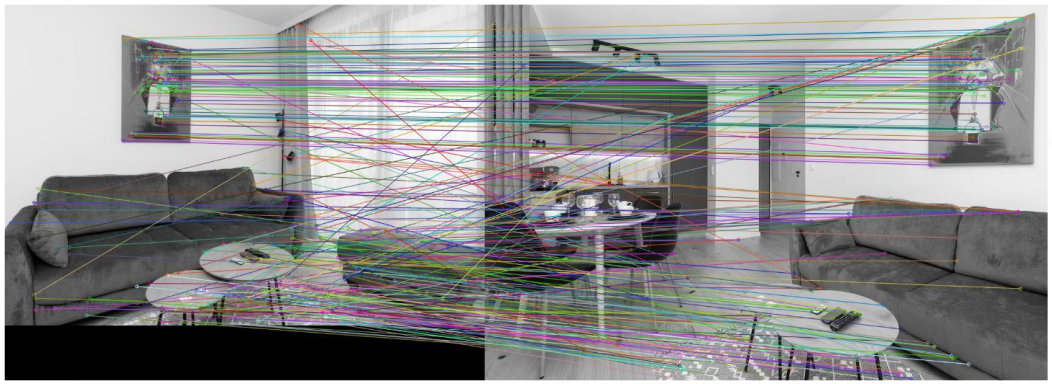
UC4: experimental statistics in tourism (1)

Aim:

- to develop experimental indicators based on data collected through web scraping from online platforms for the purpose of conducting statistical research in the field of tourism
- Use of web data for ***accommodation base*** in tourism (supply side of tourism) and ***tourists' travel patterns and expenditures*** (demand side of tourism)
- Web data can be used for ***validating and imputing*** missing records in sample surveys of tourist travel and spending (demand side of tourism)

Deduplication methodology:

- SIFT (Scale-Invariant Feature Transform) algorithm



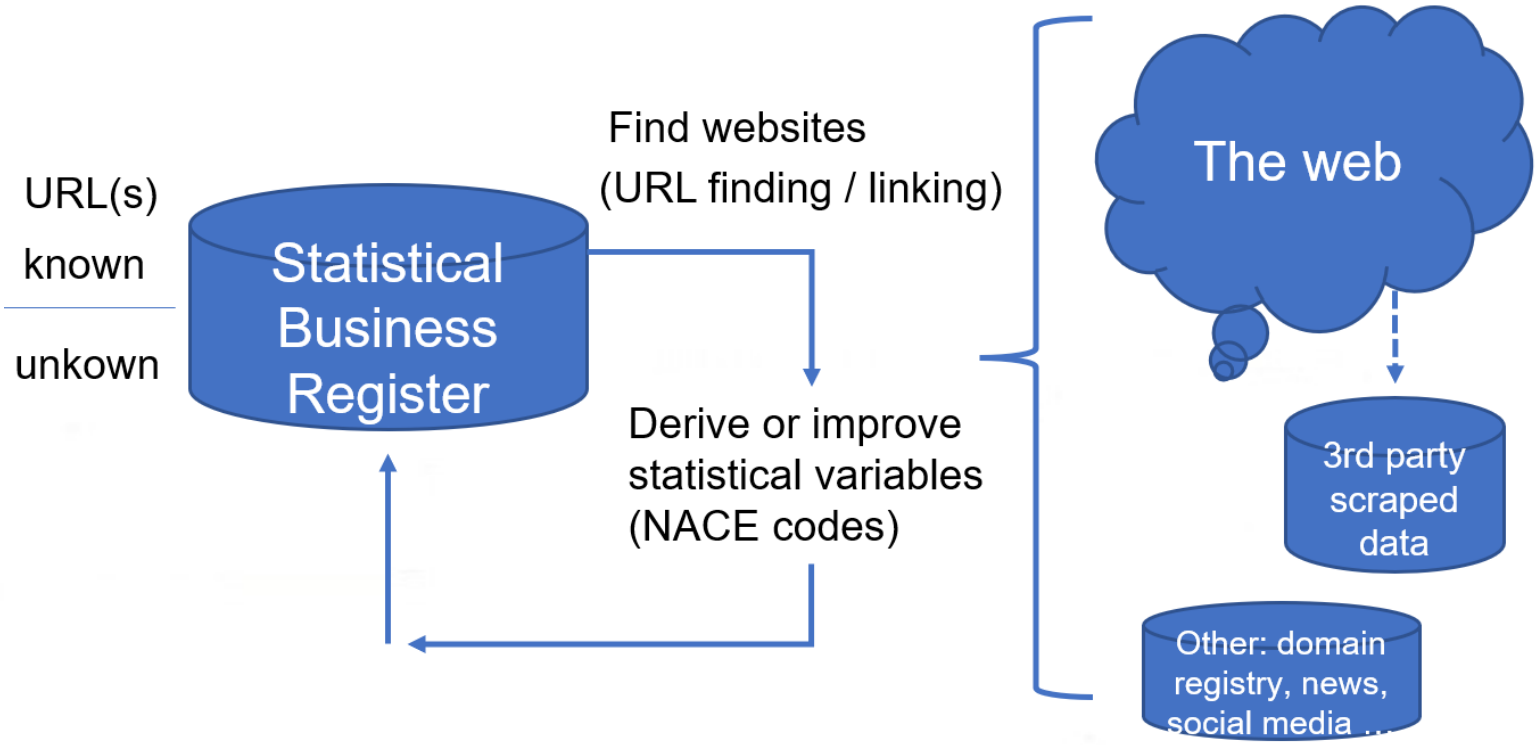
Deliverable 3.9: WP3 UC 4

Report on methods for analysing hotel price data and computing various indices of interest

UC4: experimental statistics in tourism (2)

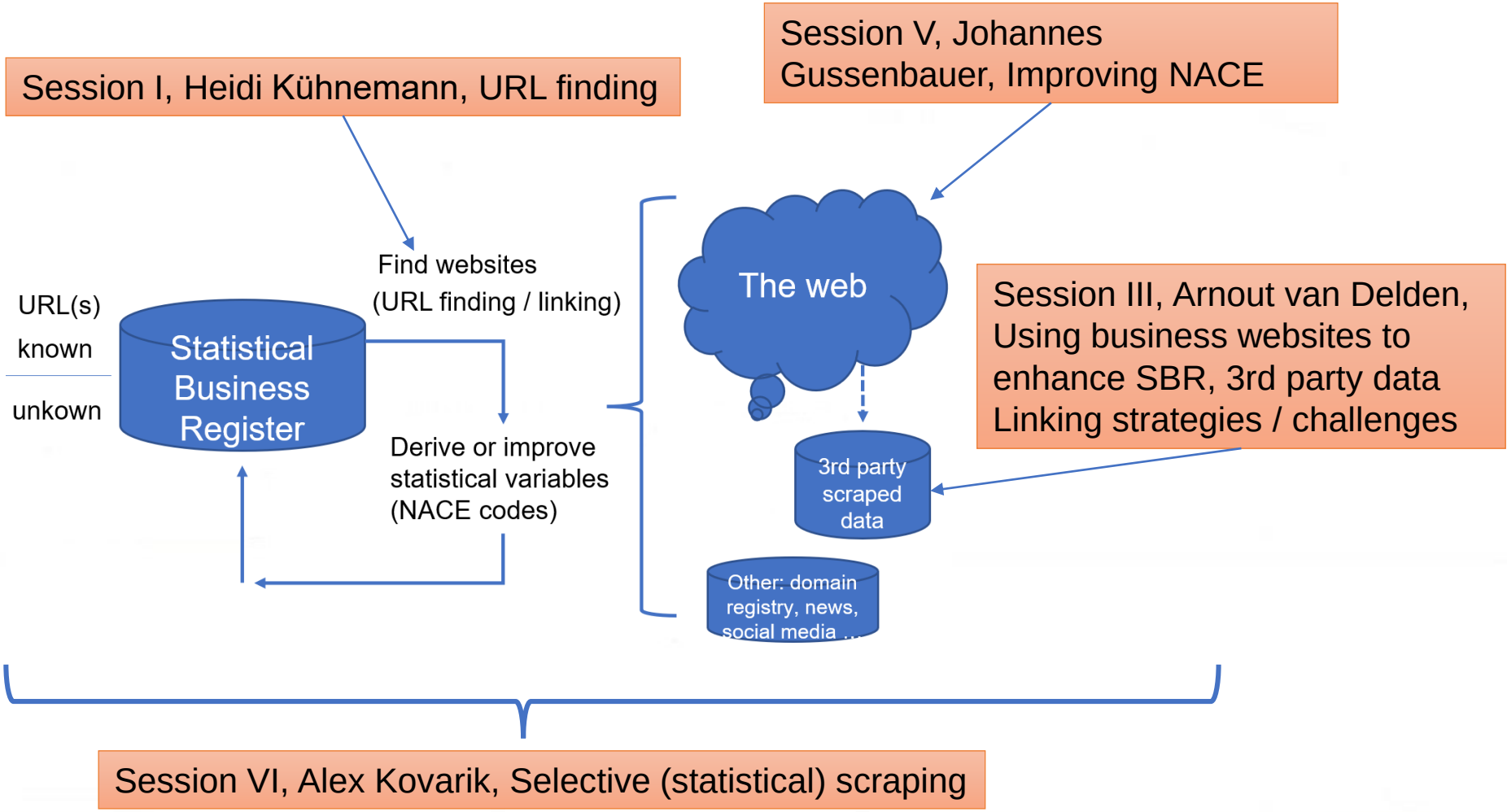
- Using web scraping repeatedly, it is possible to obtain data on accommodation ***rental prices*** for specific tourist ***destinations*** and during a specified ***period of time***.
- Indicators regarding the ***average rental prices of accommodations*** in ***Poland*** and ***Bulgaria*** have been developed. The rental prices exhibited distinct ***seasonality***.
- This data is also useful for ***validating*** and ***imputing missing data*** in ***sample surveys*** on travel and tourist expenditures conducted by NSIs within the EU.
- Comparing images across platforms that offer accommodation bookings can assist in the ***deduplication*** of accommodation property datasets.
- The research demonstrates potential in utilizing web scraping techniques to estimate ***travel-related expenditures***. While the methodology is still under refinement, the initial findings are promising

UC5: business register enhancement (1)



The SBR serves as sampling frame producing official economic statistics.
Improvements pave the way to control representational errors in all business statistics and contribute to overall **quality**.

UC5: business register enhancement (2)

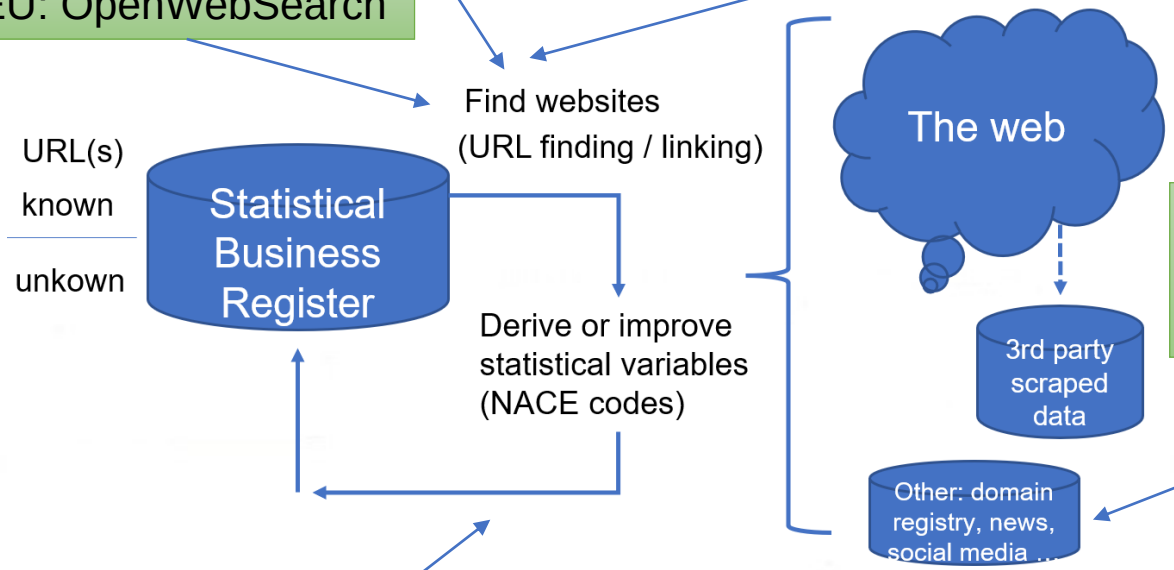


UC5: business register enhancement (3)

FI: URL finding via domain registry API

SE: pre-feasibility study
Google search API

EU: OpenWebSearch



FI: Search for new establishments using
Data from (Google) mapping APIs

Contact information discovery (Hesse)
- extracting email addresses
- names of business executives

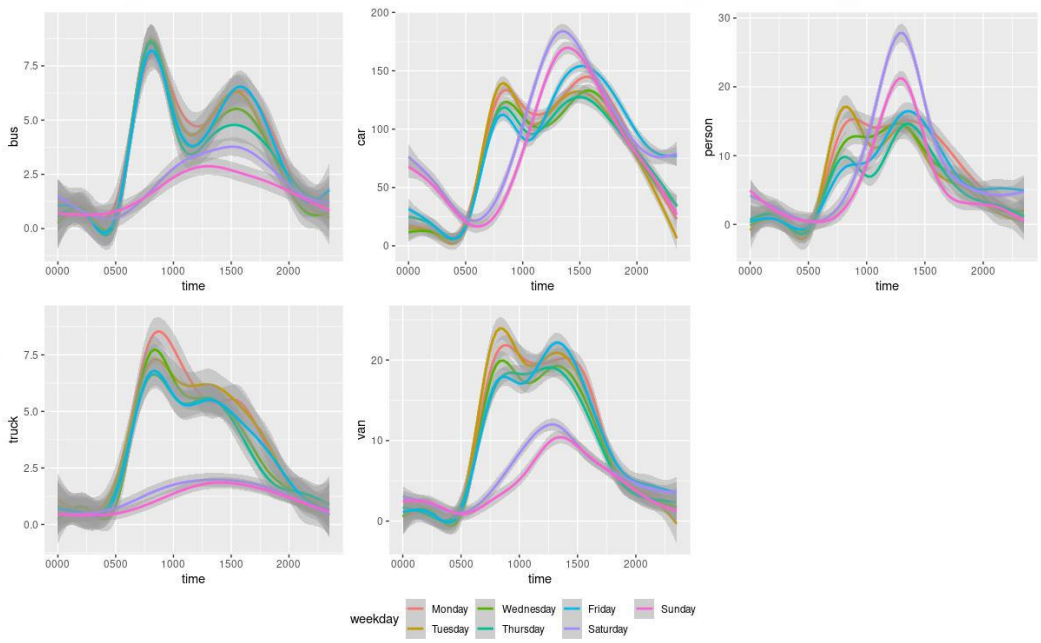
Deliverable 3.11:

UC5: Report on methodology and results to use online data for business register enhancement

UC6: faster economic indicators, traffic cams

Aim:

- to explore the applicability of using ***publicly available traffic camera images*** to calculate a ***busyness indicator***, successfully piloted in the UK, and to adapt it for use in other countries, with ***Sweden*** as a case study



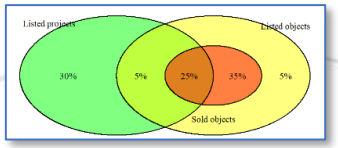
- Method is technically & methodologically feasible and portable across countries
- Different weather conditions

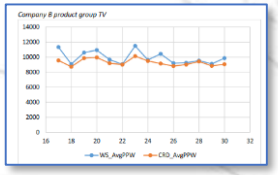
However:

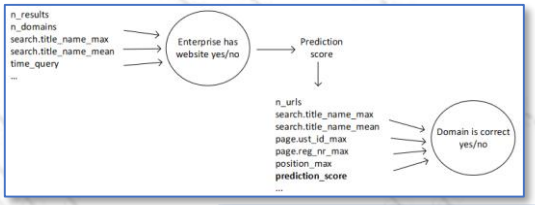
- Evolving social attitudes towards privacy => less camera's images available

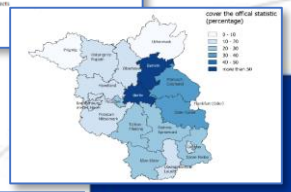
Future: combine with:

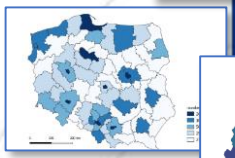
- Combine such data with MNO / sensor data / citizen science initiatives (i.e. telraam)

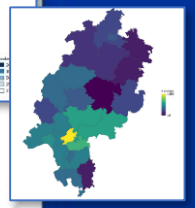


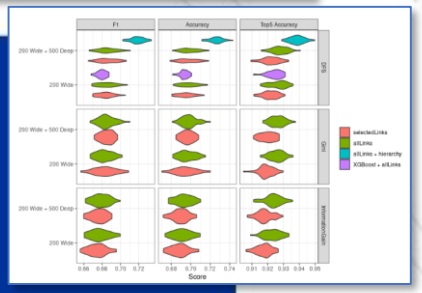


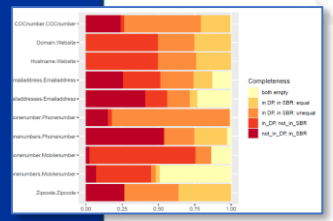


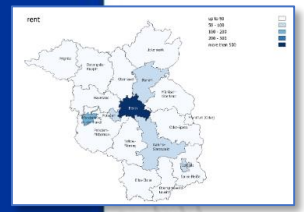












THANK YOU

Thanks to many project partners of WP3

Leads: Galya Stateva/Olav ten Bosch

Year	Month	Objects in Poland	Avg. Price in Poland	Objects in Bulgaria	Avg. Price in Bulgaria
2023	January	8276	259,11	3702	252,44
2023	February	9233	259,39	4888	272,00
2023	March	8993	302,74	4812	341,86
2023	April	10017	300,15	4397	305,61
2023	May	10687	314,45	4579	330,20
2023	June	10721	310,02	3506	310,88
2023	July	6886	411,97	1933	394,55
2023	August	8499	376,28	2645	376,06
2023	September	9490	365,03	2910	321,77
2023	October	11573	311,24	3454	320,90
2023	December	10396	300,44	3045	308,71

Web Intelligence Network

Funded by the European Union