ESSnet Trusted Smart Statistics – Web Intelligence Network
Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 1
Coordination, support and dissemination

D1.5 WP1 Final technical report

2025-03-31

Prepared by:
Dominika Nowak, Statistics Poland (project coordinator until 31-10-2024)
Klaudia Peszat, Statistics Poland (project coordinator from 1-11-2024, k.peszat@stat.gov.pl)

Jacek Maślankowski, Statistics Poland
Olav ten Bosch, Statistics Netherlands
Alexander Kowarik, Statistics Austria
Sarah Phelps, ONS
Alessandra Righi, ISTAT
Massimo De Cubellis, ISTAT

Web Intelligence Network

Funded by the European Union

This document was funded by the European Union.

The content of this deliverable represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

Contents

Project partner organizations

No.	Name	Short name	Country
1	Glowny Urzad Statystyczny	GUS	Poland (PL)
2	Bundesanstalt Statistik Oesterreich	STATA	Austria (AT)
3	National Statistical Institute	NSI	Bulgaria (BG)
4	Tilastokeskus	TILASTOKESKUS	Finland (FI)
5	Institut National De La Statistique Et Des Etudes Economiques	INSEE	France (FR)
6	Direction De L'animation De La Recherche Et Des Etudes Statistiques	DARES	France (FR)
7	Amt Fur Statistik Berlin-Brandenburg	SSI-BBB	Germany (DE)
8	Statistisches Bundesamt	DESTATIS	Germany (DE)
9	Hessisches Statistisches Landesamt	HSL	Germany(DE)
10	Istituto Nazionale di Statistica	ISTAT	Italy (IT)
11	Lietuvos Statistikos Departamentas	SL	Lithuania (LT)
12	Centraal Bureau Voor De Statistiek	CBS	Netherlands (NL)
13	Instituto Nacional De Estatistica Portugal Statistics	POR	Portugal (PT)
14	Statistiska Centralbyran	SCB	Sweden (SE)
15	Statisticni Urad Republike Slovenije	SURS	Slovenia (SI)
16	UK Office For National Statistics	ONS	United Kingdom (UK)
17	Swiss Federal Statistical Office	FSO	Switzerland (CH)

Funded by the European Union

1. Introduction

The "WP1 Final technical report" of the ESSnet Trusted Smart Statistics – Web Intelligence Network project (ESSnet WIN) covers the information on the activities undertaken, deliverables and milestones completed during the entire project period.

The report outlines the activities of all work packages (WPs):

- Work Package 1 – Building the Web Intelligence Network (WIN) across the European Statistical System (ESS) and beyond, via Web Intelligence Hub (WIH)-related competency building, targeted knowledge sharing, user support and users' active engagement in WIH development;
- Work Package 2 – The advancement of the WIH, and moving the online job advertisements (OJA) and online-based enterprise characteristics (OBEC) use cases into statistical production stage;
- Work Package 3 – The exploration of the potential to extend the WIH by new data sources and use cases;
- Work Package 4 – The development of solid methodological and quality foundations for generating statistics within the WIH.

The information about the results achieved within the project can be found on the CROS portal and all public deliverables are also made available for long term access on GitHub.

The final project conclusions and possible future work in this theme are presented in the last chapter.

2. Summary of results

2.1. WP1 Coordination and Communication

WP1 was responsible for the implementation of the tasks related to:

- Project coordination;
- Promotion, communication and knowledge sharing;
- User support and project products dissemination;
- Web Intelligence uSER (WISER) group.

GUS, ONS and ISTAT were involved in the implementation of the tasks under WP1. In order to implement them, they cooperated with other members of the project consortium, with a particularly crucial role for the WP2, WP3 and WP4 leads.

Task 1.1 Project coordination

In order to ensure smooth coordination of the project, including planning and monitoring of work progress, bi-monthly WP leads meetings were held. They involved participation of project coordinator, WP2, WP3 and WP4 leads, WISER task coordinator, promotion and training task coordinators and representatives of Eurostat. WP leads meetings serve to monitor the work progress at the level of the WPs and the project as a whole, as well as to ensure alignment of project tasks with the developments of the WIH. The latter is particularly important due to the dynamic and rapidly progressing work on the WIH on the part of Eurostat, as well as the necessity to coordinate promotion, dissemination and training activities across the ESS.

In addition, WP leads organise on a regular basis technical meetings with WP teams, in order to ensure planning, coordination and implementation of the tasks assigned to WP members.

In order to support communication and documents exchange, the internal ESSnet WIN wiki was used to exchange WP working materials and minutes from the meetings between the project members, as well as to discuss project-related issues. As an external publicly available communication platform, a dedicated page on the CROS portal contains information about the project, specific issues related to the use of web data, and selected project deliverables.

The quality of the key project deliverables was ensured by the Review Board's (RB) evaluation. The RB consisted of independent professional experts, with statistical background and expertise in non-traditional data and projects.

Task 1.2 Promotion, communication and knowledge sharing

a) Promotion and communication

During the project's lifetime, the ESSnet Web Intelligence Network successfully delivered key initiatives, meeting the project's goals of building awareness of the WIN, WIH, and web data capability across Europe and beyond. The project reached a significant number of NSIs and other official bodies interested in web data sources, demonstrating its effectiveness.

The project's communication plan provided a framework of activities and specified which messages needed to be delivered and by which communication channels.

To promote the WIN, a straightforward and confident narrative was needed. The narrative needed to include using targeted messaging and communications channels to reach the audience at the right time and place. The overall approach was proactive, promoting the WIN project and its activities to NSIs and other producers of official statistics across the ESS. The ultimate goal was to ensure awareness of the opportunities the WIN project could offer to deliver official statistics with web data.

The plan, designed with adaptability in mind, needed to consider the as-is situation, i.e., the fact that the launch date for the WIH/P was unconfirmed. The plan was adapted to ensure that all communication and promotion goals were met by the end of the project. The adaptation, a testament to our project's resilience, included delivering regular updates on the project's work streams via webinars and blogs.

At each stage of the project, there was a need to identify the key messages that needed to be delivered and to take all opportunities to repeat these key messages over a prolonged period. This will ensure that our audience is well-informed and engaged with the project.

Effective communication was a cornerstone of the project's success. It required active input from project members, who shared relevant information to raise the project's profile and ensure critical information was disseminated to the audience. The project's priorities always guided communications, with the strategy ready to adapt to changing priorities at short notice. Within the project has been developed two central communication and promotion channels - electronic (social media activity) and physical/virtual (participation in events).

The Grant Agreement aimed for 17 webinars and workshops over the project's lifetime. The project has not only met but exceeded this target, delivering 19 webinars and workshops. The events have been promoted through the project's website, social media channels, an array of NSIs internal communication channels, and other professional bodies. Recordings of the webinars have been hosted on the project's YouTube channel, further showcasing the project’s achievements.

i) Social media engagement

Bearing in mind that the project targets a niche community, the social media channels have performed admirably. The promotion and communication team posts every Thursday and, when necessary, on Tuesdays. These regular updates kept the audience informed about the project’s progress and engaged with the goals.

At present, the project has 219 followers on X (formerly known as Twitter) and 763 on LinkedIn, and over the past 12 months, the number of followers has grown by 278 on LinkedIn. Regarding X, the followers are in decline, which mirrors the decrease in X users overall. Over the last 12 months, the WIN posted 60 times, including 93 videos and articles on both X and LinkedIn. The summary statistics are presented below:

- Number of impressions (the number of times the post has been shown to users’ feeds)
    - Project total for X: 732,772
    - Project total for LinkedIn: 126,341

- Page views:
  - Project total for LinkedIn*: 2,246 and unique view 962
- Videos and articles:
  - Project total for LinkedIn:

    126,341 impressions
    18,888 views
    1.91% click through rate (LinkedIn's average is 0.22%)
    3.49% engagement rate (LinkedIn's average is 2%)

Table 1. Statistics of project's social media engagement

	2022/23	2023/24	2024/25*	Total
Number of Tweets/LinkedIn Posts	63	80	60	207
X				
Impressions	138,475	460,822	133,475	732,772
Engagements	1,896	3,059	684	5.639
Engagement rate	1.37%	0.66%	0.51%	0.31%
Profile visits	148	291	76	515
Mentions	31	11	1	43
Followers	104	183	192	211
Link clicks	309	375	98	782
Retweets without comments	290	232	62	584
Likes	397	412	88	897
Video views	14,530	54,794	7,474	76,798
LinkedIn				
Page visitors	831	657	758	2,246
Unique visitors	318	279	365	962
Total followers	259	449	763	763
Impressions	19,232	22,971	84,138	126,341
Unique impressions	10,548	13,865	55,270	79,683
Clicks	643	621	1,115	2,418
Reactions	533	527	491	1,551
Reposts	161	179	22	362
Video/articles				
Impressions	8,319	13,789	60,435	82,543
Views	4,728	12,300	1,860	18,888
Click through rate	3.34%	2.70%	1.37%	1.91%
Engagement	6.95%	5.78%	2.07%	3.49%

*X analytics are only available up until the end of May 2024 as X removed free access to this information.

The project also has its own YouTube channel where the WIN shares recordings of the training webinars. The channel was first started in December 2022 and to date 10 videos were added. The summary statistics are presented below:

- Project total for YouTube:
  - The videos have had 2,191 views

- The channel has 79 subscribers
- 400 watched hours
- 21,438 impressions
- Impressions click through rate of 2.3%

ii) Blogs

The blogs have proved to be a valuable tool to generate interest in the project's work, promote projects' achievements and engage with a broader audience.

Members of the WIN published 27 blogs covering all work packages and corresponding use cases and promoting them via social media channels.

Table 2. CROS portal blogs:

ID	Blog	Publication date
Issue 1	Overview of the project	February 2022
Issue 2	Faster Economic Indicators - Traffic Camera Data	March 2022
Issue 3	Exploring Potential New Data Sources	April 2022
Issue 4	Developing and Improving Business Registers	June 2022
Issue 5	Path to a Quality Framework for OJA Data Source	July 2022
Issue 6	Exploring Potential New Data Source - Real Estates	September 2022
Issue 7	Measuring and predicting construction activities using online data	October 2022
Issue 8	Moving Big Data into statistical production	November 2022
Issue 9	Experimental indices in tourism statistics	January 2023
Issue 10	Quality aspects of web scraped data – Focus on landscaping and selection of sources	February 2023
Issue 11	Online job advertisements time series	February 2023
Issue 12	Statistical business registers – an important cornerstone in official statistics	April 2023
Issue 13	Data accuracy for hierarchical classification	June 2023
Issue 14	Exploring potential new data sources – real estate - update	July 2023
Issue 15	Web Intelligence Hub (WIH) introductory video	September 2023
Issue 16	Exploring new data sources update	October 2023
Issue 17	Measuring and predicting construction activities using data from online advertisements on internet real estate platforms – up date	October 2023
Issue 18	Getting enterprise characteristics based on website data	January 2024
Issue 19	Integrating Big Data into Tourism Statistics	February 2024
Issue 20	Online prices of household appliances and audio-visual, photographic and information processing equipment	March 2024
Issue 21	Gathering data on advertisements of house and apartments with a contract	April 2024
Issue 22	Business registers enhancements using web data	May 2024
Issue 23	Lessons learned from Eurostat's Deduplication Challenger	June 2024
Issue 24	Navigating the digital frontier: Unlocking the potential of web scraped data for official statistics	June 2024
Issue 25	Measuring Construction Activities Using Data from Online Advertisements on Internet Real Estate Platforms: Insights in Data Quality Discussions	September 2024
Issue 26	Online prices of household appliances and audio-visual, photographic equipment - update	October 2024
Issue 27	Harnessing Big Data for Tourism Statistics: Challenges and Future Steps	January 2025

b) Capability building initiatives

The project was scheduled to deliver each year at least four webinars / training sessions, participation in a selection of networking events and presentations at various statistical conferences (at least two events per year) and organize one virtual hackathon. The training programme delivered during the project exceeded the goal specified in the Grant Agreement.

i) Webinars

In total the project delivered 19 webinars and workshops. The topics which these training events have covered are presented in Table 3.

Table 3. Webinars delivered within the project

ID	Topic	Date	Type	Delegates registered	Delegates attended	% of attendance	Reach	
							Organisations	Countries
1	How to tutorials ESSnet WIH only Video	2022	Internal	N/A	57	N/A	Project members	
2	Quality and Methodology Training	2022	External	173	93	57%	42	24
3	Use Web Scraped Data to Enhance the Quality of the Statistical Business Register	2023	External	243	183	75%	88	43
4	Methods of Processing and Analysing Web Scraped Tourism Data	2023	External	286	162	57%	88	36
5	Web Intelligence in Practice. How to use content from the web for enterprise statistics?	2023	External physical	N/A	8	N/A	N/A	
6	Web Data in Official Statistics: Process, Challenges, Solutions - Online real estate	2023	External	166	99	60%	72	31
7	Project Training OJA Workshop	2023	Internal	100	38	38%	Project members	
8	Measuring Construction Activities using Data from the Web	2023	External	132	84	64%	51	23
9	Project Training OJA Workshop	2023	Internal	94	23	24%	Project members	
10	New avenues with Web Intelligence - Gaining additional value from cash register data by coming different sources	2023	External	128	65	51%	34	22
11	WISER Training WIH Training	2023	Internal		19		WISER members	
12	Measuring the quality of large-scale automated classification systems applied to online job advertisement data	2024	External	67	58	67%	54	16

13	WISER Training OJA Testing	2024	Internal	N/A	9	N/A	WISER members	
14	WISER Training Testing WIP	2024	Internal	N/A	9	N/A	WISER members	
15	Lessons learned from Eurostat's Deduplication Challenge	2024	External	68	37	54%	29	17
16	Artificial Intelligence and Machine Learning – Training at IAOS conference	2024	External physical	N/A	20	N/A	N/A	
17	WISER Training OBEC Testing	2024	Internal	N/A	8	N/A	WISER Members	
18	Introduction to the Data Acquisition Service (DAS) of the Web Intelligence Hub (WIH)	2024	External	121	71	59%	41	21
19	WISER Training Feedback Session	2024	Internal	N/A	5	N/A	WISER members	

For the external webinars the following interest was achieved:

- 58% attendance of those who registered.
- 48% classified their experience as beginners, 39% as intermediate, 13% as experts.
- Average attendance per external webinar was 54 delegates.
- The delegates came from 70 organisations from across 38 countries.

ii) Conferences

The project team members promoted their work in several conferences and international events. These conferences provided a key platform for raising awareness of the project and possible applications of web data sources in official statistics. Conferences attended by the WIN members included:

- IAOS (Poland), April 2022
- Q2022 (Lithuania), June 2022
- CARMA (Spain), June 2022
- NTTS (Belgium), March 2023
- WIH-CON (Belgium), June 2023
- ISI World Statistics Congress (Canada), July 2023
- European Statistics Day (Spain), October 2023
- European Innovation Network Plenary Meeting (Luxembourg), February 2024
- Conference on Foundations and Advances of Machine Learning in Official Statistics (Germany), April 2024
- IAOS (Mexico), May 2024
- UNECE Expert Meeting on Statistical Data Collection and Sources (Switzerland), May 2024
- Q2024 (Portugal), June 2024
- ICES (UK), June 2024
- Web Intelligence Network. From Web to Data (Poland), February 2025
- NTTS (Belgium), March 2025.

At the start of February, the WIN end-of-project conference, From Web to Data, was held in Gdansk, Poland. This conference not only showcased the work the WIN had completed over the last four years. It was also designed for the wider statistical and data science community, representatives from NSIs and other bodies that produce official statistics, academia, the private sector and anyone who works with web data. The event gathered nearly 80 participants.

The conference took place over two days and had presentations covering:

- Web scraping and infrastructure – 5 presentations
- OJA use case – 6 presentations
- OBEC use case – 5 presentations
- New use cases – 5 presentations
- Quality of web data – 5 presentations
- Methodology on using web data -– 4 presentations

All presentations and recordings are available on the CROS portal.

iii) Hackathon

The hackathon held by the WIN project was an online challenge of 6 weeks in autumn of 2024. The hackathon was promoted via the ESSnet WIN social media channels, NSIs involved in the project, and other external stakeholders, such as national statistical associations across Europe and international organisations. This led to the registration of 10 teams originating from different organisations registered for the challenge.

The challenge to be performed was described on the CROS portal pages of the project. In brief, the challenge was to develop open-source software, to be published under an open source license on a public GitHub repo, to score a given dataset of 4000 URLs of enterprises in 4 different countries (NL, AT, PL, DE). For each country 1000 URLs were available. The binary variables that had to be derived from the websites were e-commerce and social media use. The latter variable had subcategories Facebook, LinkedIn, X, Instagram, TikTok and YouTube.

The dataset was derived from public available entries on maps spread over different regions in the respective countries and varying in activity of the enterprises. The data was deduplicated and a sample was taken to arrive on the dataset for the challenge. A subset of 100 URLs per country was manually labelled by the project partners to arrive at a (secret) validation set to decide on the winner(s).

Two winners have been chosen. Their results can be found here:

- Roshna Omer (UNHCR)
  Enhanced Social Media and E-commerce Detector aka:
  https://github.com/RoshnaOmer/win-hackathon/

- Riccardo Corradini, Rita Lima (ISTAT)
  Freesoftwdreamer team:
  https://github.com/freesoftwdreamer/Web-Intelligence

One of the winners, Riccardo Corradini, presented their work in the WIN session at the NTTS 2025.

Task 1.3 User support and project product dissemination

User support, in the form of the helpdesk, was provided through a dedicated space on the CROS portal. The prepared space for the WIH helpdesk consisted of two components: a form for submitting comments on the functioning of the WIH and the Question and Answers (Q&A) section. With the migration of the CROS to a new environment, the helpdesk has been temporarily unavailable. In addition, given the fact that the WIH is not yet available to external users, the feedback from the WIH testing is provided by project members directly to Eurostat, mainly in a coordinated way via WP leads.

Task 1.4 Set-up of the WISER group

To facilitate the effective use of the WIH and ensure its adoption within the European Statistical System (ESS), the project established the Web Intelligence uSER (WISER) group. The WISER group is composed of experts from National Statistical Institutes (NSIs), academia, and international organizations, with representation from a diverse array of fields, including web data, methodology, and data science. This group was formed to serve as a bridge between the project's development teams and the external user community, providing insights, feedback, and independent assessments to refine the WIH's functionalities.

WISER members were carefully selected based on their expertise and were engaged as key contributors to the project. Their involvement includes evaluating project outputs, suggesting improvements, and identifying potential new functionalities for the WIH, particularly in the domains of online job advertisements (OJA) and online-based enterprise characteristics (OBEC). The group's role extends to promoting the WIH within their respective networks, supporting the dissemination of project findings, and fostering integration of the WIH's outputs within national statistical systems. The collaborative work of the WISER members is an essential contribution to achieving the project's goals, ensuring that the WIH meets the diverse needs of data users across Europe.

As part of the comprehensive plan for 2024, a series of training activities have been organized to support WISER members in effectively testing and providing feedback on the Web Intelligence Platform (WIP), the Online Job Advertisements (OJA) Datalab, and the Online-Based Enterprise Characteristics (OBEC) tools. These sessions aimed to provide members with hands-on experience across key functionalities of the platforms, with a particular focus on data acquisition, usability, and specific experimental outputs.

Following the three presentations held in May and June, WISER members were encouraged to independently test the platforms throughout the summer, giving them ample time to explore the tools and apply the training knowledge to real use cases. A series of follow-up meetings were organized in autumn in order to collect feedback on the tested tools, suggestions and recommendations. Summary of the WISER's recommendations is available in the D2.5 report.

2.2. WP2 OJA and OBEC software

The aim of WP2, as stated in the Grant Agreement, was to move the use cases of online job advertisements (OJA) and online-based enterprise characteristics (OBEC) into the production environment.

The implementation of the WP objective was carried out in the course of the following activities:

- Development of the scripts for OBEC data collecting, processing and analysing;
- Development of scripts contributing to OJA data collecting, processing and analysing;
- Preparation of the suggested indicators to move to production environment;
- Preparation of the list of possible improvements based on the feedback provided by the WISER group.

WP2 involved 11 member organisations from 10 countries, including: GUS, STATA, NSI, DARES, DESTATIS, HSL, ISTAT, SL, CBS, SURS and FSO which was not a project beneficiary but was involved in its implementation.

The first year of WP2 activities focused on addressing challenges related to defining the target population of companies in OBEC use case, legal aspects concerning the transfer of URLs included in business registers (for some NSIs), and issues regarding data comparability. For OJA use case, the criteria for evaluating existing datasets – including their alignment with official statistics published on NSIs' websites – were established.

OJA (online job advertisements) use case activities included the following tasks:

- review of OJA Classified Structural Metadata,
- overview of data in the OJA repository and the proposed list of OJA URLs for scraping,
- general approach to OJA data quality, covering:
  - various data assessment rules,
  - duplicates in OJAs across EU countries,
  - open OJAs,
  - new methodologies for evaluating OJA data quality (e.g., a brief overview of WP4's approach, including OJA data validation rules and a review of WIH-OJA validation rules).

OBEC (online-based enterprise characteristics) use case included the following tasks:

- defining the OBEC population,
- reviewing the software for URL scraping,
- conducting a requirement analysis for the OBEC platform and DataLab,
- assessing the current state of the OBEC platform and DataLab,
- developing a methodology for URL retrieval.

In particular, defining the requirements for the OBEC platform and DataLab gave the opportunity to create the useful and reliable environment to scrape and process enterprise data from websites. The complete software environment has been provided by Eurostat.

The second year of the project provided the following activities related to OJA and OBEC respectively:

- mapping the company and economic activity based on OJA data, including the methodology and results,
- annotation exercise to check the machine learning accuracy in detecting occupation – WP2 participated in design, conduct and provide the results of this exercise,
- suggesting changes in ontologies and Natural Language Processing algorithms,
- creating the set of possible OJA indicators to supplement existing data based on questionnaires,
- creating the URL Database to feed the Web Intelligence Platform,
- providing the methodology to evaluate e-commerce activity based on the websites,
- providing the in-person OBEC training course at the NTTS 2023 conference.

The third and four year of WP2 work consisted of creating and completing a set of tables and exchanging experiences between countries, in particular:

- providing training sessions to internal and external users of OJA and OBEC in WIH,
- creating the set of reliable indicators based on OJA data in the DataLab by the results of annotation exercises as well as accuracy tables received from Eurostat (provided by Lightcast),
- analysis of the OJA data in DataLab and calculation of the indicators based on occupation (ISCO 1st level) and regions (NUTS2), with the use of relative indicators,
- participation in the OBEC annotation exercise conducted by WP4,
- creating set of indicators on OBEC use case, i.e. social media presence, e-commerce and multilanguage versions of websites.

Compared to the previous year, it was possible to use the WIH infrastructure for testing the OBEC use case. URL lists for selected countries were uploaded to the WIH platform. Data was scraped and processed in the DataLab environment linked to the WIH platform for web scraping. The tasks performed involved:

- developing the indicator on enterprise social media presence, enterprises' e-commerce activity as well as multilanguage website indicators,
- the use of OJA DataLab to deliver high quality data on the demand for specific group of occupations.

In addition, an important task completed by WP2 involved delivering training to the project partners on the use of OJA data, including advanced methods to create indicators with the OJA data. Also, jointly with Eurostat, WP2 delivered training on the WIP to the WISER group.

The WP2 final results are summarised in the three deliverables:

D2.4 Suggested tables with experimental statistics and metadata including quality assessment

This deliverable is a result of working on OBEC and OJA use cases during the last year. With regard to the OBEC use case, it contains templates of tables and metadata, with data calculated by the use of the WIH platform and DataLab.

D2.5 List of requirements defined by the WISER group with the result of its implementation

This is the result of the training sessions provided by WP2 members and meetings with the WISER group. Nearly 20 different requirements were formulated, including the WIH, OBEC and OJA use cases.

D2.6 Technical guidelines on the implementation of the scripts produced for OJA and OBEC statistics

This deliverable shows all necessary information to move OBEC and OJA use cases into the production using the WIH environment. The deliverable has been positively reviewed by external reviewers.

Finally, the goal of WP2 has been achieved, as several experimental statistics based on OBEC and OJA data, with some limitations related to the data quality aspects, were produced. The experimental statistics include selected OJA attributes as well as OBEC-related indicators, such as e-commerce, multilanguage and social media presence.

2.3. WP3 New use cases

According to the grant agreement WP3 aimed to:

- explore the potential of new types of web data sources still not integrated into the WIH, with a view on their future integration in the WIH/P;
- produce experimental statistics for new types of web data sources demonstrating the potential to produce statistics.

During the last project's year WP3 has further improved the exploration of web data sources within use cases (UCs) and worked on wrapping-up the results of the four project years into the final deliverables per UC. As defined in the Grant Agreement the results of the UCs are not organised into one document for the whole work package as in the first 3 years, but instead they are to be published as specific documents and results per use case. Also, these final deliverables are the first WP3 deliverables that are to be published publicly. Therefore it made sense to make them as complete as possible and hence some of the contents of the non-public year reports were re-inspected to give the readers inside and outside of the WIN project the full picture. More in detail, the work on deliverables completed in the fourth year is listed below.

Deliverable D3.4 *Report on the results of the new data sources exploration and the conditions for using the data* from UC1 on Characteristics of the real estate market give an overview of the activities performed in this UC. Although the data sources studied in this UC were spread over different countries, different websites and were collected in different ways, there were similarities in the challenges faced at all these stages and the results achieved. The deliverable gives an overview for UC partners on data acquisition software used, methodologies applied, how issues were tackled such as missing data, or redundancy in data, and strategic questions around originality and preference of data sources. Some partners have reached a high stability level of web data for real estate and one of the conclusions is that the data can be a supplement to real estate market monitoring systems, or can be used to create early indices or to build hedonic indices or models classifying real estate in new cross-sections. The deliverable D3.5 *Experimental* *statistics for characteristics of the real estate market,* to be published on the project wiki, is backing up the research in D3.4.

Deliverable D3.6 *Report on methods and feasibility to track construction activities based on real estate web portals* describes the activities from UC2 on measuring construction activities from web data. The deliverable describes the role of official statistic on the theme of construction, the overall aim of this UC, and points out the connection between UC1 and UC2. Furthermore, it contains a definition of “newly constructed” properties, lists the real estate web portals used in the study, as well as a minimal set of indicators to extract and the IT Choices of the UC partners. Also, it presents some overarching questions regarding quality aspects of the data. The deliverable contains an overview of the country specific data sources, the data preparation steps performed by each partner as well as the results achieved, including quality aspects, such as missings, duplicates and coverage. One of the conclusions of the report is that while there is a difference between target population and survey population, data from real estate web portals are a valuable source which allows analysis of and insights into a specific and important part of economy and society.

Deliverable D3.7 *Report on methodology and results for online prices* describes the activities from UC on measuring online prices of household appliances and audio-visual, photographic and information processing equipment. Statistics Sweden and Statistics Bulgaria worked together to explore web data in their respective countries for this UC goal. The work consisted of exploration of the data sources (two in Sweden and four in Bulgaria), development of web scraping software, data acquisition and recording and processing of the data to get estimates of the online market. At some point Statistics Sweden concluded that from a statistical viewpoint it could be more valuable to focus on opportunities for finding and estimation distributions regarding the correlation between popularity of each item and how much they actually sell by combining web data with administrative data. Hence, the deliverable also contains the results of a Swedish study on this particular subject. Deliverable D3.8 *Experimental statistics for online prices …etc.,* to be published on the project wiki, is backing up the work in D3.7.

Deliverables D3.9 *Report on methods for analysing hotel price data and computing various indices of interest* and D3.10 *Report on methods to be used for imputation of price data in price statistics* describe different aspects of the work in UC4 on experimental indices in tourism statistics. D3.9 describes the results of scraping one of the leading hotel platforms, which appears to be valuable for analysing both the accommodation base in tourism (supply side of tourism) as well as for studying tourists' travel patterns and expenditures (demand side of tourism) for balance of payments purposes. The deliverable describes the web scraping, data linkage, metrics, possible use of the data in surveys an interesting image deduplication technique and raises some strategic questions. D3.10 describes the results of scraping two other platforms for validating and imputing missing records in sample surveys of tourist travel and spending (demand side) conducted for balance of payments and tourism statistics. The conclusion states that while the methodological work is still ongoing, these early results indicate the value web scraping offers in enhancing the accuracy of travel expenditure survey.

Deliverable D3.11 *Report on methodology and results to use online data for business register enhancement* from UC5 on business register enhancement contains detailed descriptions on the two important phases: URL finding and enhancing the business register (with a focus on NACE prediction). After a thorough general introduction in both subjects, the country activities on each topic are described. This exemplifies the variety in this work. For example, URL finding can be performed by using search engines, or linking third party data or using other web data such as map data. In addition to NACE prediction, other results are improving contact information or searching for new establishments. The deliverable also contains the results of a literature review and a discussion on future work including ideas for other indicators to be derived from these data sources.

Deliverable D3.12 *Report on assessment of challenges and opportunities* from UC6 on faster economic indicators using new data sources contains a detailed description on early work on traffic activity indicator from public camera images. It describes the objectives, data sources, processing pipeline, and deep learning model and reflects on some complexity concerns measuring busyness. The report describes the scaling up of the work to other countries, in particular the proof of concept in Sweden, which led to challenges on data availability, infrastructure, and accuracy. One of the conclusions is that the methodology is technically portable across countries, but also that practical challenges such as differences in weather conditions and privacy concerns limit full realization of this concept. The report also highlights some related work and future opportunities.

Concluding, although data streams which have been set up in early project years have been continued as much as possible during the fourth project year, much of the work in this final year has been spent on writing down the experiences in a set of deliverables which are hoped to give a good overview of the new data sources examined and their potential to produce official statistics or be used as an auxiliary source.

2.4. WP4 Methodology and Quality

The aim of WP4, as per the grant agreement, was to:

- consolidate knowledge gained in the ESSnet WIN in the area of methodology and quality when using web data in the statistical production process;
- extend and enhance deliverables of the ESSnet Big Data II (especially WPF & WPK) focusing on web data;
- transition from quality guidelines to quantitative quality indicators as well as the first cross-national quality assessment for online job advertisements (OJA) and online-based enterprise characteristics (OBEC).

The work package was divided into four distinct tasks: Quality (led by Statistics Austria), Quality Assessment (led by Statistics Finland), Methodology (led by Statistics Netherlands), and Architecture (led by ISTAT).

The work completed by WP4 in the project involved preparation of five content deliverables:

- D4.1 Minimal guidelines and recommendations for implementations
- D4.5 Quality guidelines for acquiring and using web scraped data
- D4.6 Methodology report on using web scraped data
- D4.7 BREAL – Big data reference architecture and layers for web scraped data
- D4.8 Quality assessment for the statistical use of web scraped data.

The deliverable D4.1 *Minimal guidelines and recommendations for implementations* outlines practical guidelines and strategies for integrating web scraped data into official statistics. The document focuses on methodology, quality management, and software architecture to enhance statistical processes using web intelligence and is also structured along these 3 topics. The report served as a foundational guide for newcomers and experts in leveraging web intelligence for statistical purposes at the beginning of the project.

The deliverable D4.5 *Quality guidelines for acquiring and using web scraped data* aims to provide generic guidelines for NSIs and other organizations on the use of web data for the production of official statistics. The report outlines systematic approaches to landscaping (data sources selection), and quality aspects important for all phases of the statistical production process (input and throughput phase I and II). It also includes recommendations for a central web scraping infrastructure. This deliverable acts as a comprehensive resource for statisticians and data scientists involved in integrating web data into official statistics, offering advice for quality monitoring and assurance.

The goal of the deliverable D4.6 *Methodology report on using web scraped data* is to identify generically applicable methods that can be used to produce web-based statistics of high quality. The report includes three major topics: sampling for web data, the web based statistical process with a special focus on the sources of bias, and the methods used for dealing with web scraped data (such as URL finding and linking, web scraping methods, dealing with over- and under-coverage, deduplication of units, detection of concept drift, correcting model-induced bias). The report underscores the need for continuous improvement in methodologies to address the dynamic and complex nature of web data in statistical contexts.

D4.7 *BREAL - Big Data REference Architecture and Layers for web scraped data: hands-on experiences and architectural challenges* details the development and implementation of the Big Data REference Architecture and Layers (BREAL) framework, designed to support the collection, processing, and integration of web data for official statistics. The work emphasizes enhancing statistical production systems through centralized and shareable infrastructure, known as the Web Intelligence Hub (WIH). The document emphasizes the need for a shared vision and holistic approach at the EU level, it offers a framework for collaboration and innovation within the European Statistical System.

D4.8 *Quality assessment for the statistical use of web scraped data* evaluates the quality of web scraped data with a focus on online job advertisements (OJA) and online-based enterprise characteristics (OBEC). It assesses data stability, accuracy, and applicability for official statistics. For OJA quality indicators to assess data sources' relevance, stability and ranking over time were introduced. This led to a detection of fluctuations in source stability and ranking of the current available OJA data from Eurostat raising concerns about data reliability for time series. The two OJA annotation exercises for classification accuracy showed declining accuracy at more detailed ISCO classification levels. Both quality assessments, of source stability and classification accuracy, showed that the current state of the OJA data is far from being usable for the production of official statistics.

OBEC data assessment analysed the accuracy of automated URL finding and the models for detecting social media and e-commerce on enterprise websites. It found reasonably high accuracy in URL linkage and for the two indicators.

Overall, these two examples stress the importance of rigorous quality assessment while adopting web data for statistical production.

All reports were reviewed twice by the project’s review board and comments from reviewers were considered during finalisation of the documents.

3. Final project conclusions

Web data has been, remains, and will continue to be an important source of information for official statistics. It offers the opportunity to create data that are timely and rich in information, complementing the traditional data collection methods of national statistical institutes. However, despite continuous efforts, the integration of web data into the production process of official statistics is still in its early stages. While exceptions are to be noted where integration did succeed (particularly in price statistics), or is coming close to being achieved, the majority remain experimental and the quality of ad hoc estimates made directly and only on web data is in general poor.

Moving from experimental applications to fully integrated official statistics presents not just a technological challenge, it also and more particularly raises the methodological issue of dealing with the poor data quality within the wider perspective of making trusted official statistics at the level of a whole statistical population.

One major obstacle is ensuring that official statistics based on web data meet the rigorous quality standards required for official use. Therefore, we envision that web data cannot stand alone – it must be combined with other data sources to produce meaningful and reliable statistics. A purely web-based approach is in many cases insufficient, making integration with other already established data sources such as survey or administrative data a key factor in its successful application.

Another critical issue is that the information/statistical output requested from NSIs often lacks important information or sufficient granularity, thus prompting NSIs to also adopt web data for statistical production. This might convince management that a strong push in the use of web data would be needed. For instance, a key question is whether job vacancy data, or more generally labour market indicators, could be delivered sooner using online job advertisements (OJA) data. Currently OJA experimental statistics fall far below the quality standards set for official statistics, often even lacking a proper estimation methodology needed to make statements, based on inference, on the population.

The evolving regulatory landscape within the European Union also highlights the need to keep improving methodologies, such as non-probability or small area methods that are able to work in a multi-source setting with web data as one source. With new EU regulations on the horizon, adapting strategies to meet emerging requirements is essential. Instead of merely requesting official statistics, future approaches should prioritize problem-solving. One example in case is the issue mentioned above already, whether online job vacancy (OJV) data can be made available earlier.

Future projects should focus on specific problems that need to be solved, e.g. the previously mentioned fast indicators for the labour market rather than a specific data source.

Several use cases demonstrate high potential for integration into statistical production at a later stage, while still requiring additional development:

- **Real estate market analysis –** Rental market statistics in particular remain a weak point in official data, making web data a valuable supplementary source.

- **Multi-source data approaches –** The combination of mobile network operator (MNO) data with administrative registers or the integration of web data with tourism statistics could significantly improve accuracy and comprehensiveness.
- **Large Language Models (LLMs) for URL discovery –** The application of LLMs in identifying and categorizing relevant web data sources is a promising avenue for further exploration.

Finally, the great achievement of this project was to establish a **Web Intelligence Network (WIN) community** that is very much interested in sharing developments and exchanging ideas. Maintaining and expanding this community is an important objective for the future. Interest in the topic extends far beyond the current consortium, with other NSIs eager to engage and a strengthening cooperation with academia, particularly in the development of sound methodologies, will be essential for advancing web data research and ensuring its sustainable integration into official statistics.

Funded by the European Union