WEBFOSS-AT

Grant Agreement Number: 101225462 — STATA-WEB-SSIG - SMP-ESS-2024-WIH-IBA

Work Package 1: Open Source Tools for URL Finding and Information Extraction, Statistical Scraping Interest Group (SSIG)

Deliverable 1.3: Report on SSIG1

2025-10-23

Prepared by:

Alexander Kowarik, Statistics Austria

Johannes Gussenbauer, Statistics Austria

Statistical Scraping
Interest Group

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.

Content

1. Introduction

The Statistical Scraping Interest Group (SSIG) was established ad continuation of the network established during the Web Intelligence Network (WIN) project. It connects several national initiatives that are related to using web scraping for statistical production.

The group and the meeting are co-organized by Statistics Netherlands and Statistics Austria. The first took place 16-17 September 2025 in Vienna, Austria at Statistics Austria.

There are in total four SSIG meetings planned, the second meeting will be held in the Hague 15-16 April 2026.

The SSIG functions as an informal community of practice: share experiences, learn from examples, explore new approaches. The group has the long-term goal to develop a formal statistical scraping theory.

The SSIG meetings provide the opportunity for presentations but also give plenty of time for discussion and general exchange.

All presentations of the meeting are publicly available on Github.

Funded by the European Union

2. Participants

FIRSTNAME	LASTNAME	Institution	Country
Alexander	Kowarik	Statistics Austria	AT
Christian	Salwiczek	Statistik Nord	DE
Raul	Fernandez	INE	ES
Olav	ten Bosch	Statistics Netherlands	NL
Dominik	Blatt	Statistics Netherlands	NL
Luuk	Haarman	Statistics Netherlands	NL
Luca	Gramaglia	European Commission - Eurostat	EC
Joan	Fernández-Navarro	Universitat Politecnica de Valencia	ES
Femke	Bosman	Statistics Netherlands	NL
Dennis	Pipenbring	Statistics Denmark	DK
Sarah Valentina	Möller	Statistics Denmark	DK
Jacco	Daalmans	Statistics Netherlands	NL
Tobias	Gramlich	Statistik Hessen	DE
Nina	Niederhametner	Statistics Austria	AT
Johanna	Bolli-Kemper	University of Zurich	CH
Brandon	Sepulvado	NORC at the University of Chicago	US
Gussenbauer	Johannes	Statistics Austria	AT

Brandon Sepulvado presented online due to unforessen circumstances.

3. Agenda

16 September at 9:30
9:30 - 9:40 Welcome & Organisational (Alexander Kowarik)
Item 1. 9:40 - 10:10 Statistical Scraping – history, concept, where we are, SSIG (Olav ten Bosch)
10:10 - 10:30 Introduction round: who is here
Item 2. 10:30 - 11:00

- Webscraping Use-Cases for Statistics Austria (Johannes Gussenbauer, Bernhard Meindl, Alexander Kowarik)
- R-Package for Selective Web-Scraping (Bernhard Meindl, Johannes Gussenbauer, Alexander Kowarik)

Item 3. 11:30 - 12:00

- Classification of Companies' Economic Activity: A Web Content and Machine Learning-Based Approach (Joan Fernández Navarro, Ana Debón, and Josep Domenech)

Item 4. 13:30 - 14:30

- WEB-FOSS-NL: Statistical Scraping Explored in the Netherlands (Olav ten Bosch, Luuk Haarmans, Dominik Blatt, Femke Bosman, Jacco Daalmans)

Item 5. 15:00 - 15:45

- Statistical Scraping – Related Initiatives in Eurostat (Luca Gramaglia)

15:45 - 16:00 Closing first day

Funded by the European Union

17 September at 9:30
9:30 - 9:40 Start of the second day
Item 6. 9:40 - 10:10

- Statistical Scraping at Statistics Hesse (Tobias Gramlich)

Item 7. 10:10 - 10:40

- Adjusting Non-Probability Samples in Job Vacancy Statistics (Johanna Bolli-Kemper)

Item 8. 11:00 - 11:30

- Leveraging AI for Statistical Web Scraping (Brandon Sepulvado)

Item 9. 11:30 - 12:00 Round Table / Discussion:
Especially participants without presentation can freely explain their project / work / experience
12:00 - 12:20 Feedback on the SSIG1, Ideas for next SSIG meetings
12:20 - 12:40 Closing – Outlook next meetings

4. Meeting summary

1. Statistical Scraping: history, concept, where we are, SSIG (Olav ten Bosch)

The presentation provided an overview of the history, concept, and current state of statistical scraping, as well as the role of the Statistical Scraping Interest Group (SSIG).

CBS (Statistics Netherlands) has over 15 years of experience with web data: fuel prices, real estate, air tickets, webshop data for CPI, enterprise websites, social media, job portals, tourism portals, Wikipedia, DNS, municipal and school portals.

Concept of Statistical Scraping

- Defined as a targeted, methodological approach to link web data to statistical units, registers, and classifications.
- Aims to reduce representation errors and enable calculation of quality indicators.
- Emphasizes careful handling of sensitive inputs and avoiding “quick and dirty” scraping.

Statistical scraping offers potential, but robust methodology for design, validation, and estimation is still needed. There is balance between bulk scraping (for population discovery) and targeted scraping (for efficient data collection) needed.

There is an ongoing debate on how far web data can substitute or complement traditional surveys.

2. Webscraping Use-Cases for Statistics Austria & R-Package for Selective Web-Scraping (Johannes Gussenbauer, Bernhard Meindl, Alexander Kowarik)

The presentation showcased two pilot use-cases for applying webscraping in official statistics at Statistics Austria. The first use-case focused on scraping wine prices from Austrian vineyard websites, with the aim of using these prices as a proxy for the Agricultural Producer Price Index (API). The approach starts with vineyard units from the Statistical Business Register, identifies their websites and online shops, and extracts wine price lists to estimate average prices across different types. Challenges include finding sufficient websites, ensuring that prices are consistently available and in usable formats, and deciding whether a full census or a representative sample of vintners would be more appropriate .

The second use-case addressed NACE classification in the context of the upcoming revision. With a massive re-classification effort ahead, the idea is to support NACE coding by scraping and analyzing product information from enterprise webshops. The workflow envisages identifying relevant enterprises, scraping their online shop content, and classifying products—potentially with the help of COICOP categories—to derive rules or machine-learning models that map businesses into the new NACE 2025 structure. Open questions include the feasibility of large-scale webshop scraping (and risk of blocking), how much data per site is needed, and whether methods like word embeddings (e.g. cosine similarity) suffice for reliable automated classification.

The second part of the presentation introduced their work on an R package for selective webscraping, designed to provide reproducible, configurable workflows with minimal dependencies, parallelization, retry logic, customizable outputs, and integration with Dockerized Selenium, aiming to make statistical scraping accessible and easily adaptable for different projects.

3. Classification of Companies' Economic Activity: A Web Content and Machine Learning-Based Approach (Joan Fernández Navarro, Ana Debón, Josep Domenech)

The contribution explored the potential of using company websites to automatically classify firms by economic activity. Manual classification is costly, slow, and prone to errors, while the widespread availability of corporate websites creates an opportunity for automation. The project used the SABI database (2.9 million companies, 151,000 with available websites) and applied a pipeline from web content extraction to embeddings generation (using OpenAI's multilingual text-embedding-3-small model), followed by machine learning training. Preprocessing steps included synthetic oversampling (SMOTE, ADASYN, Borderline SMOTE) and embedding aggregation (truncate, max pooling, mean pooling, sum pooling), with classification carried out at the division level (CNAE level 2).

Results showed that mean pooling combined with SMOTE achieved the best performance, with a validation F1 score of around 0.56. Comparisons between flat and hierarchical classifiers indicated that hierarchical models achieved better generalization and robustness, though at significantly higher computational cost. Error analysis highlighted challenges such as ambiguous or unrepresentative web content and misclassifications in the original database, but the model was able to correct some errors, suggesting value in error detection. The conclusions stressed that automated classification using web content is both feasible and scalable, with future work focused on diversifying data sources, exploring advanced oversampling tailored to text, using LLMs for synthetic data generation, and experimenting with deep learning architectures to further improve accuracy.

4. WEB-FOSS-NL: Statistical Scraping Explored in the Netherlands (Olav ten Bosch, Luuk Haarmans, Dominik Blatt, Femke Bosman, Jacco Daalmans)

The presentation described current Dutch efforts to structure and test building blocks for statistical scraping. The approach is centered on linking the Statistical Business Register to enterprise websites, then applying a modular chain of URL finding, focused scraping, and text mining/LLMs to extract specific information, with initial focus on job vacancies (JVs). The work is organized in phases: developing building blocks independently, integrating them into a process chain, and testing on subsets of the register to validate feasibility.

The URL finding component is being refactored as open-source software, using search APIs and potentially EU projects like OpenWebSearch, to reliably locate enterprise websites. The focused scraping module aims to extract only the relevant subpages (e.g. careers pages), using keyword-based approaches tailored to country and language contexts (with examples for NL, AT, PL, UK). The presentation highlighted open questions such as the depth and breadth of scraping, the role of sitemaps, and the best format for outputs (raw HTML, structured text, or clean text) to be passed to classifiers.

The text mining/LLM block is building on earlier CBS and AIML4OS hackathon work, using prompts to detect job vacancies or classify webpage content. Two options are being explored: zero-shot LLM prompting for direct detection, and feature extraction approaches where LLMs generate vectors that can be passed to traditional classifiers. A prototype is being developed on INSEE's Onyxia platform with GitHub collaboration, aiming at reusable open-source tools for broader applications such as hotels or sport clubs. Key challenges remain around multilinguality, scalability to millions of pages, and the need for labeled validation data.

WEB-FOSS-NL exemplifies a modular, open-source, collaborative approach to statistical scraping, where URL finding, scraping, and text mining components can be iteratively improved and reused across domains. The project positions itself as experimental, encouraging cross-country contributions to refine the methods and test practical applicability, with future milestones including SSIG2 in The Hague (spring 2026) for follow-up results and community engagement

5. Statistical Scraping – Related Initiatives in Eurostat (Luca Gramaglia)

A first focus was on Online Job Advertisements (OJAs): Eurostat has been collecting OJAs to produce experimental labour market statistics, such as the OJA rate and ICT labour demand. While traditional NLP methods have been used for classification, a recent study tested LLM-based embeddings, vectorising both OJAs and ESCO occupation descriptions. Using expert-labelled samples and farthest-point sampling for diversity, the study showed that embedding models outperform previous methods. Next steps include validating robustness and integrating the approach into OJA production.

A second initiative covered extracting information on top-tier multinational enterprises (MNEs) from Wikipedia, including employees, revenue, and assets. Wikipedia offers scraper-friendly, standardised info boxes, though freshness and accuracy vary. Around 50% of enterprises were matched with the relevant Wikipedia page automatically; the rest required manual matching. Quarterly updates are now shared with the Statistical Business Registers Working Group, and future work may extend to scraping financial reports from company websites.

Other topics included the European Statistics Awards, which crowdsource solutions in nowcasting and web intelligence (e.g., deduplication and classification of OJAs, financial data discovery/extraction), and a review of the legal framework around scraping, which will take into account changes in EU copyright and data access laws. Eurostat will also update its landscaping of job portals in 2025, refining the methodology from 2021 to take into account the comments and recommendation from the WIN ESSnet.

6. Statistical Scraping at Statistics Hesse (Tobias Gramlich)

The presentation outlined how Statistics Hesse has implemented productive web-scraping systems in official statistics. Four main applications were highlighted: (1) tourism statistics by scraping hotel booking portals to identify new units for inclusion in the sampling frame; (2) scraping the commercial trade register (due to the absence of an API) to update the business register and support economic activity classification; (3) URL finding for units from the statistical business register (up to 10,000 Google searches per day), linking URLs to register IDs and storing website texts; and (4) keyword search on these websites to help classify economic activity and refine target populations.

All applications are accessible via a central entry point that can be used by all 15 statistical offices in Germany. More than 120 users are registered, with separate environments for development, testing, pre-production, and production. The platform is built on R, Python, Shiny/ShinyProxy, Kafka, and Keycloak, with strict synchronization between network zones to meet data protection and IT security requirements. Future priorities focus less on new developments and more on stabilization and incremental improvements, such as enhanced ML/LLM-based URL linking, updating stored URLs, extracting imprint information, and potentially providing APIs for integration with other systems.

7. Adjusting Non-Probability Samples in Job Vacancy Statistics (Johanna Bolli-Kemper)

The presentation addressed challenges and solutions when using online job vacancy (JV) data in Switzerland. The Swiss Job Market Monitor (SJMM) has collected job ads since 2000, combining a probability sample of firms (scraped websites and surveys on posting channels) with non-probability samples (NPS) from job boards (since 2006) and press archives (until 2018). While job boards cover a large share of online vacancies, they are cheaper to scrape but introduce strong selection bias, as not all firms or vacancies are represented. This creates the central methodological problem: how to use non-probability data for representative official statistics.

Three model-based approaches to adjust NPS were discussed. The quasi-randomization (QR) approach estimates pseudo-inclusion probabilities from a reference sample to weight NPS observations; the superpopulation (SP) approach models variables of interest directly and imputes them into the reference sample; and the doubly robust (DR) approach combines both, ensuring consistency if at least one model is correct. In practice, SP and DR are well-suited for official statistics with a limited set of variables, while QR is better for producing scientific use files where weights are needed for broader analysis. Key QR techniques include post-stratification and inverse propensity weighting. Looking ahead, the SJMM aims to implement these adjustments, improve data integration between probability and adjusted non-probability samples, and strengthen infrastructure for deduplication and representativeness.

8. Leveraging AI for Statistical Web Scraping (Brandon Sepulvado)

The presentation explored how large language models (LLMs) and AI agents can support scraping tasks in official statistics. Using the case of healthcare practices acquired by private equity, the study investigated whether AI could help identify and extract establishment-level information from websites more efficiently than traditional scraping methods. The analytical workflow combined constructing an organizational dataset, using the Wayback Machine to capture pre-acquisition website versions, scraping establishment addresses, linking them to outcomes, and applying an event study design.

Two scraping approaches were compared. The traditional method relied on downloading entire sites, analysing their structure, and coding extraction routines, while the AI-based approach simply gave an LLM a URL and requested practice addresses. Results showed about 80% overlap between the two methods, with the traditional one producing slightly higher counts of practices. However, AI significantly reduced effort—by up to 80–90%—and handled issues like deduplication and address formatting more effectively. Challenges remain, particularly regarding security risks of AI agents and potential systematic biases in coverage, but the study concluded that AI approaches can provide higher efficiency and complementary results for statistical web scraping.

9. Discussion

The concept of "focused scraping" was highlighted in the Dutch presentation and echoed in the R package currently under development at Statistics Austria, both applying similar methodologies. The group agreed that this represents an important area for joint methodological development. As a first step, the current implementation will be shared, and ideas will be collected on possible enhancements beyond the keyword-based approach—for example, leveraging website sitemaps.

It was further agreed to coordinate communication with the Open Web Search Initiative (OWS), with Eurostat providing updates on its long-term sustainability and potential integration into URL-finding tools. Another presentation introduced the use of the Wayback Machine to access historical versions of websites as a means of enriching information about entities, a practice that could be further explored in a statistical context.

Denmark shared their experience using DuckDuckGo for URL finding and will be able to share R code that implements this soon.

The usage of inhouse vs. publicly hosted LLMs was discussed briefly. The platform provided by INSEE for all ESS members in the project AIML4OS includes also LLMs that are callable via APIs from external sites.

Future meetings could be joined/back-to-back meetings with all Eurostat web data grants recipients. SSIG3 and SSIG4 locations are not fixed yet and the organizers welcome volunteers.

Finally, all presentations will be made available on the SSIG GitHub site, where links to related projects will also be added.