EUROPEAN
STATISTICAL
SYSTEM

Web Intelligence Hub

Modernising the data collection process within the European Statistical System (ESS)

The project

The Web Intelligence Hub (WIH) is a toolbox for high-quality statistics from web content, aiming to integrate smart technologies and innovative data sources within the European Statistical System (ESS).

Implemented by Eurostat and the ESS as part of the ESS Innovation Agenda, the WIH focuses on creating new data pipelines and using cutting-edge technologies such as artificial intelligence (AI), machine learning (ML) and natural language processing.

The WIH applies its tools and methodologies to web content across multiple domains. Currently, the WIH most developed use cases are: :

- online job advertisements (OJA);
- online-based enterprise characteristics (OBEC);
- multinational enterprises (MNE)

Future plans include expanding to additional domains. The main goals of the WIH are to:

- develop new data pipelines for web scraping;
- create infrastructure to harvest and extract website content;
- produce microdata for statistical purposes

The motivation

The project is essential for modernising the data collection process within the ESS. By leveraging web scraping and natural language processing technologies, the WIH aims to provide up-to-date and detailed data, and enhance the accuracy and relevance of European statistics to support policymakers with timely and comprehensive insights.

Moreover, the WIH seeks to make use of web content to support the production of high-quality statistics. In today's digital age, web content is more than just an additional data source; it is a crucial, detailed and up-to-date resource. The WIH taps into this vast web content, ensuring that the data is of the highest quality and relevance. By complementing traditional data sources like surveys and administrative data, the WIH enables the production of more timely and detailed European statistics, with less burden on data providers.

The project's outputs, which include methodologies and data for several use cases, offer better insights into topics such as labour market trends and business activities, establishing an infrastructure for continuous and efficient content, and data extraction for statistical production.

In turn, these high-quality statistics can be used to improve decision-making capabilities for policymakers and generally support the public good, given that they are more reflective of the wealth of information available on the web.

The methodology

The WIH uses innovative techniques such as web scraping and crawling, as well as direct access via agreements with website owners, to gather content from various websites.

Advanced technologies like natural language processing and AI/ML are then employed to extract data from this content, ensuring it is of a high quality and ready for analysis.

To maintain high standards, the WIH implements methodologies to ensure the quality of new data types. It also transitions from experimental workflows to production pipelines, making the process more efficient and reliable.

The project deploys infrastructure to support web content retrieval and data extraction. It also addresses legal and organisational challenges associated with web scraping, ensuring compliance and security so that its users can have full trust in the data.

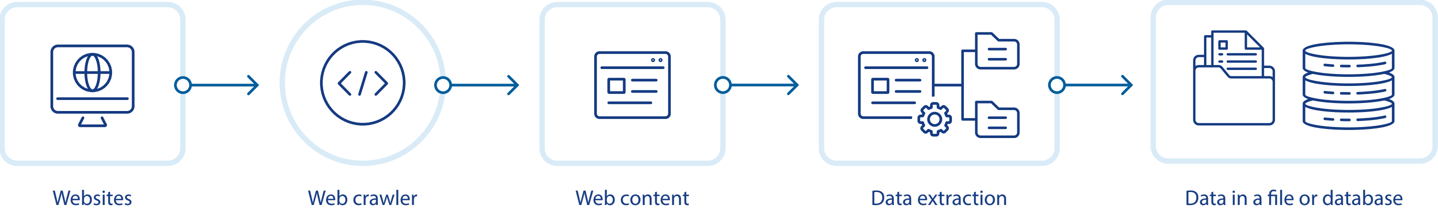
Additionally, the project collaborates with National Statistical Institutes – who make up the ESS alongside Eurostat – to guarantee the secure exchange of data.

The WIH is not just a concept, however – it is a tool to be used practically. One of the WIH's most advanced use cases focuses on OJA, providing a wealth of data, tools and methodologies.

The WIH gathers job advertisements from various online portals and websites. This content is then processed and transformed into data about the job market in Europe.

The WIH is currently extracting MNE data as these enterprises play a key role in the EU economy. As such, the WIH aims to further improve the availability of information on MNE for timely statistics. Currently, the WIH is working with around 1,500 such enterprises – a figure that should increase over time – and already has a set of developed methodologies.

© Shutterstock 2024



The team

The project team for the WIH is under the guidance of Eurostat.

Key stakeholders include Member States involved in the Web Intelligence Network (WIN), whose experts help to develop and implement the WIH, external partners such as the European Centre for the Development of Vocational Training (Cedefop) and policy Directorates-Generals (DG) like DG Employment, Social Affairs and Inclusion (EMPL).

The broader group of stakeholders include:

- ESS staff, the statistical community;
- website owners;
- policymakers and researchers who utilise the data

The timeline

- **2020:** Project launch and initial setup
- **2024:** Completion of the piloting phase and preparation for deployment
- **2025:** Full implementation and impact realisation

More information

- Web Intelligence Hub
- Web Intelligence Network