EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

From Web Content to Quality Data: Rules, Roles, and Reliability in the Web Intelligence Hub

Fernando Reis

Eurostat, WIH methodology team

Raquel Faria Paulino

WIH methodology team, Sogeti Luxembourg

Vladimir Kvetan

Cedefop

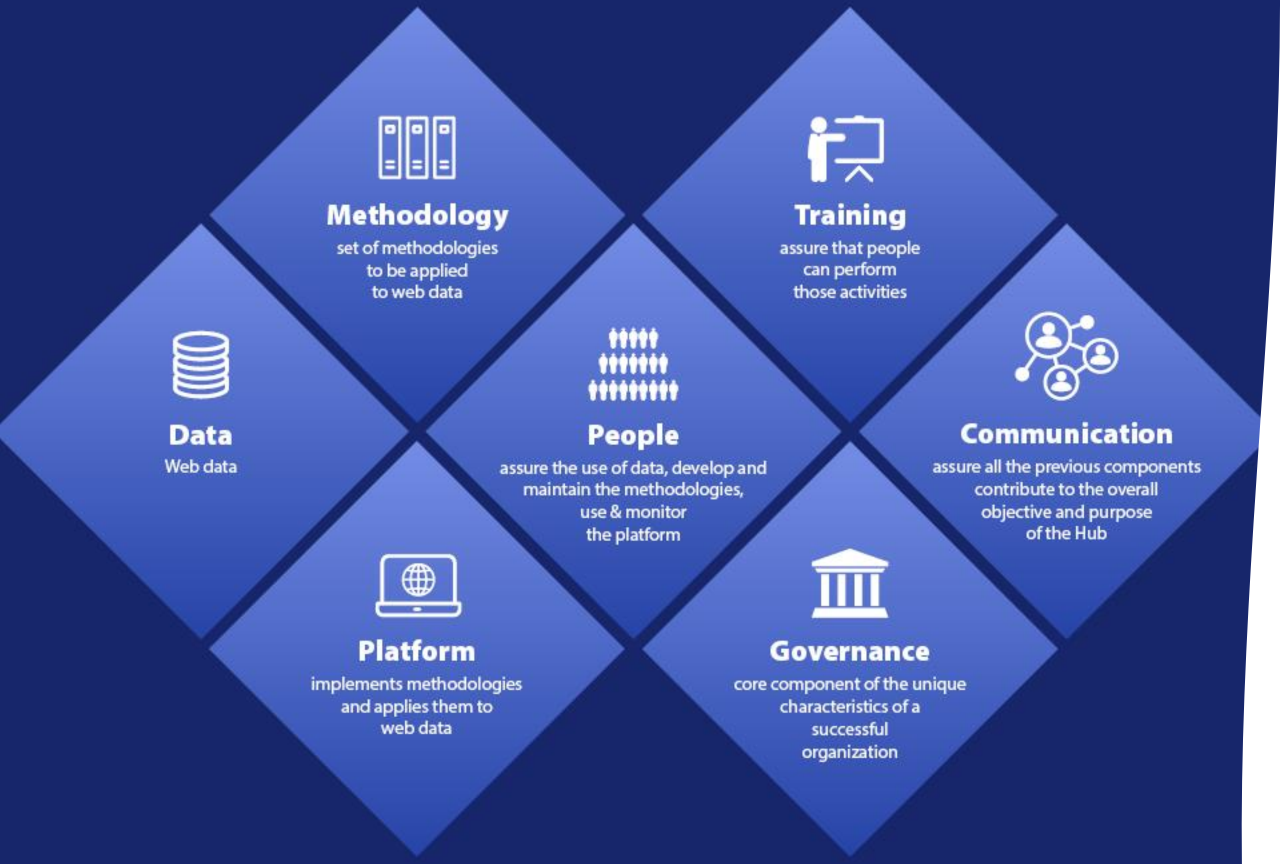
Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Web
Intelligence
Hub

Web
Intelligence
Hub



Methodology
set of methodologies to be applied to web data

Training
assure that people can perform those activities

Data
Web data

People
assure the use of data, develop and maintain the methodologies, use & monitor the platform

Communication
assure all the previous components contribute to the overall objective and purpose of the Hub

Platform
implements methodologies and applies them to web data

Governance
core component of the unique characteristics of a successful organization

Web Intelligence Hub (WIH)

□ The **WIH** is the **pillar** of TSS that provides the fundamental building blocks for harvesting information from the web to produce statistics

□ **Mission**: *“a high-quality source of data extracted from web content, methodologies and algorithms, ready to be used to produce European and national official statistics”*

□ **Collaborative effort:** Eurostat, NSIs, statistical authorities and partners

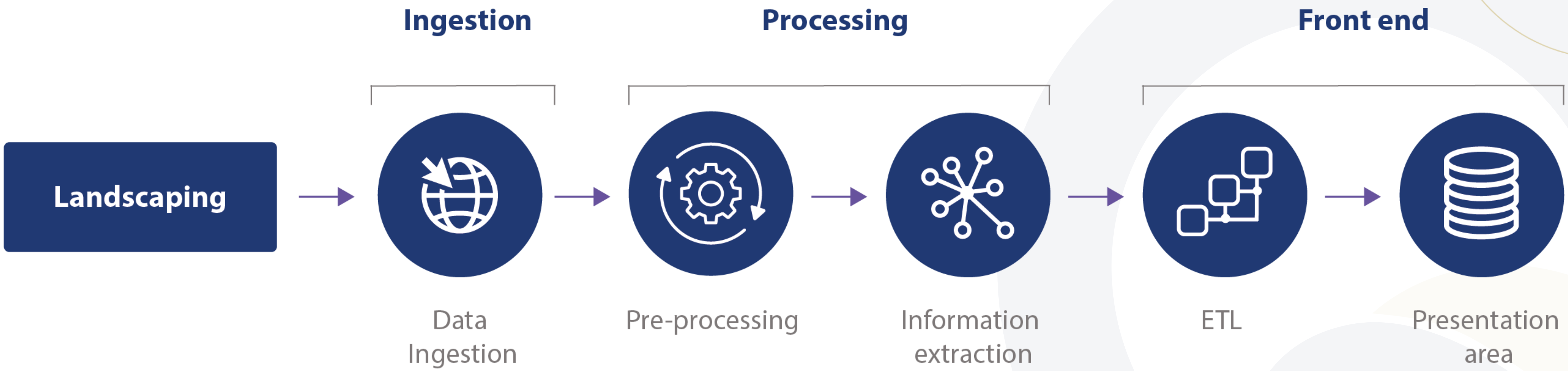
□ **Community of experts**: Web Intelligence Network, CEDEFOP

□ **WIH Platform**: technical components and services

□ **Current use cases**:

- **Online Job Advertisements**,
- Online Based Enterprise Characteristics (OBEC),
- Multinational Enterprises (MNE)

WIH data production and use



Ingestion

Processing

Front end

Landscaping

Data Ingestion

Pre-processing

Information extraction

ETL

Presentation area

First WIH data-based experimental statistics, based on OJA data

❑ Experimental statistics on Labour market demand for ICT specialists in online job advertisements

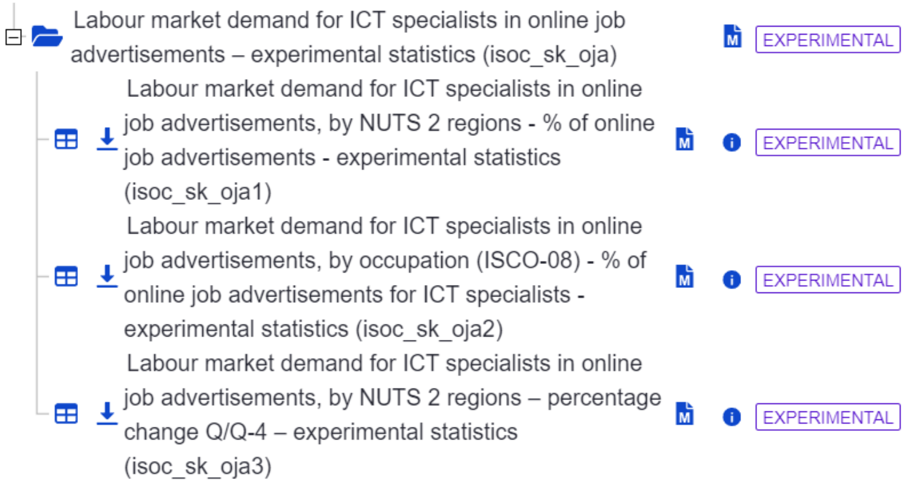
- Proportion of OJAs for digital occupations
- Released December 2023

❑ Top & Trending Skills Web Tool

- Most required skills in OJAs, by occupation and country
- Released December 2023

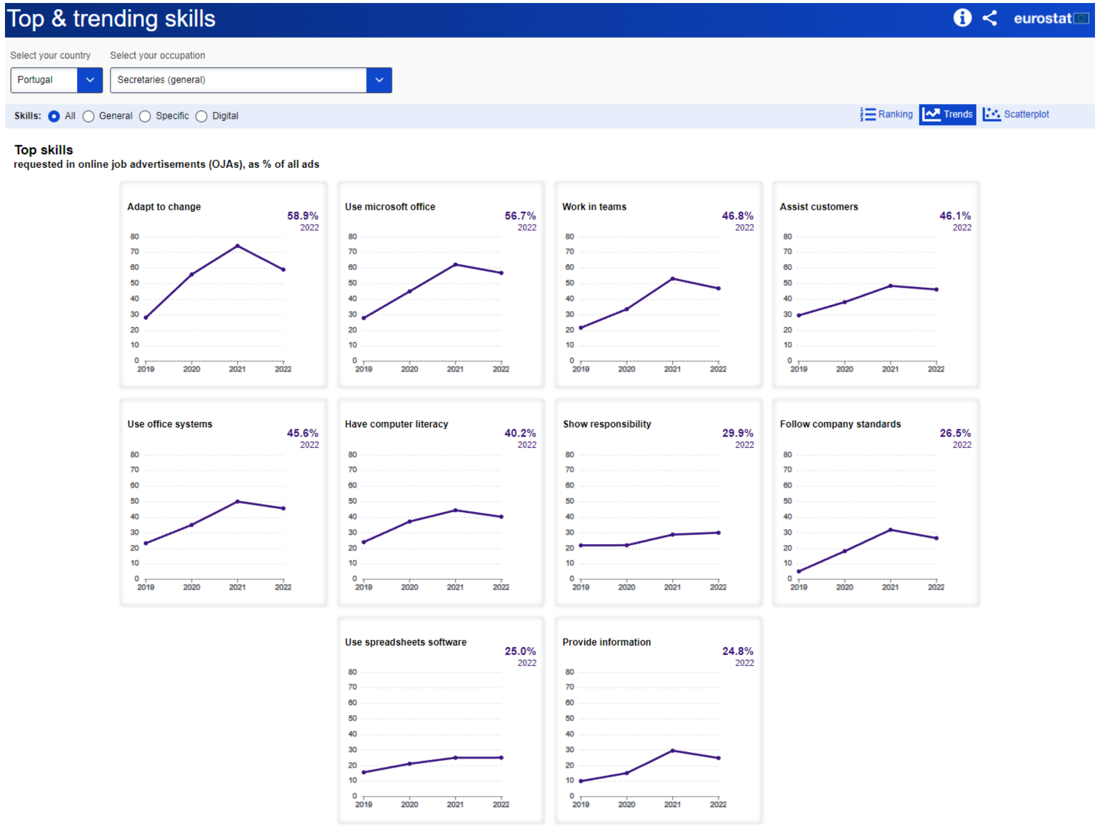
❑ Experimental statistics on labour shortage indicator

- Ratio job advertisements / employment by occupation and NUTS
- To be released 3rd quarter 2024



Labour market demand for ICT specialists in online job advertisements – experimental statistics (isoc_sk_oja) EXPERIMENTAL

- Labour market demand for ICT specialists in online job advertisements, by NUTS 2 regions - % of online job advertisements - experimental statistics (isoc_sk_oja1) EXPERIMENTAL
- Labour market demand for ICT specialists in online job advertisements, by occupation (ISCO-08) - % of online job advertisements for ICT specialists - experimental statistics (isoc_sk_oja2) EXPERIMENTAL
- Labour market demand for ICT specialists in online job advertisements, by NUTS 2 regions – percentage change Q/Q-4 – experimental statistics (isoc_sk_oja3) EXPERIMENTAL



Top & trending skills

Select your country: Portugal

Select your occupation: Secretaries (general)

Skills: ● All ○ General ○ Specific ○ Digital

Ranking | Trends | Scatterplot

Top skills
requested in online job advertisements (OJAs), as % of all ads

eurostat
Adapt to change 58.9% 2022
Use microsoft office 56.7% 2022
Work in teams 46.8% 2022
Assist customers 46.1% 2022
Use office systems 45.6% 2022
Have computer literacy 40.2% 2022
Show responsibility 29.9% 2022
Follow company standards 26.5% 2022
Use spreadsheets software 25.0% 2022
Provide information 24.8% 2022
80
70
60
50
40
30
20
10
0
2019
2020
2021
2022

WIH Rules and Procedures

❑ Modular set of rules and procedures, for easier adaptation in the future

❑ First set of modules (Annexes)

- ❑ Web content Retrieval
- ❑ Web Sources Agreements
- ❑ Access to Content and Data

❑ Potential future modules (Annexes)

- ❑ Retention of Content and Data
- ❑ Access to Source Code

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

WIH Web Content Retrieval (WCR)

- **Purpose of rules**: ensure transparency in web content retrieval practices by the Web Intelligence Hub (WIH) and foster collaboration with website owners
- **Scope**: Applies to all WIH content retrieval activities, targeting specific websites within certain use cases
- **Principles**: transparency, minimizing burden on website owners, meet statistical needs, ensure methodological soundness, adhere to internet conventions
- **Website Selection**: based on quality and relevance considerations, with agreements potentially made with the most relevant sites' owners
- **Identification & Crawling**: identify itself to websites and crawls, balancing website owners' rights with the public interest in statistics
- **Sensitive Content**: Handles various types of sensitive data, including personal and statistically confidential data, in compliance with relevant legal provisions.
- **Dissemination and communication**: Communicate retrieval purpose, rules, and impacts to website owners (when significant server impact is expected). Engage with the website owners to gather insights for and jointly discuss standards and use cases.

WIH Web Sources Agreements (WSA)

- **Purpose**: Clarify practices around making agreements with website owners for content retrieval
- **Scope**: Provide guidance for formal agreements with website owners, considering the interests of both the WIH and the website owners
- **Principles**: Equal treatment of all website owners, transparency, minimization of burden, and consideration of relevant statistical needs, with the aim of ensuring methodological soundness of the WIH data and derived statistics
- **Types of Agreements**: Includes content retrieval agreements and cooperation agreements for regular data transfers and common projects seeking higher quality data or metadata production
- **Criteria for establishing agreements**: defined by relevance of the websites and the impact of the retrieval activities on websites, with specific conditions for each type of agreements

WIH Access to Content and Data (ACD)

- **Purpose**: Guides stakeholders in accessing and using retrieved web content and data, ensuring compliance with ethical and legal standards.
- **Scope**: Directed at all ESS partners, website owners, and potential data users.
- **Principles**: Use of data for statistical purposes, data confidentiality; and wide access to data and equal treatment of users.
- **Data Types and Domains**: Distinguishes between raw and curated content, microdata, aggregated data, and monitoring data. Identifies three access domains: confidential, restricted, and public.
- **Access Roles**: Defines roles for producers, official statisticians, scientific researchers, analysts, and the general public, each with specific access rights and restrictions.
- **Dissemination Channels**: Includes use files, dedicated databases, and data lab access and data use constraints. The licensing terms vary by access domain.
- **Communication**: Ensures public dissemination of methodologies and data access rules.

Key quality concerns and challenges

- ❑ Maintaining high **accuracy** for content that is mostly “crowd-sourced”
- ❑ Assuring **relevance** of data produced from content of which generation was not originally designed for statistical purposes
- ❑ Maintain **punctuality** despite unexpected changes and uncertain availability of websites
- ❑ Ensure **comparability** despite lack of standardization across websites in structure, terminology, update frequency and change practices
- ❑ Limit the **burden** on website owners, given the potential large amount of content retrieved from their websites
- ❑ Despite the public nature of most web content, assure statistical web scraping does not jeopardise legitimate commercial interests of website owners (**data confidentiality**)

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

Maintaining high **accuracy** for content that is mostly “crowd-sourced”

Coverage errors

Landscaping activities ensures that the selection of websites from which content is retrieved is as thorough as possible [WCR]

Transparency and communication of a set of rules and procedures to foster trust and engagement of as many website owners as possible [WCR, WSA, ACD]

Measurement and Processing Errors

Selection of websites based on the assessment of the quality of their content [WCR]

Data provision agreements allowing the transmission of more structured data, not requiring algorithm-based data extraction [WSA]

Non-response Errors (completeness of content retrieved)

Data provision agreements with website owners ensuring more stable data access, reducing the risk of missed content while scraping and potentially access to content held by website owner but not available in the website [WSA]

Assuring **relevance** of data produced from content of which generation was not originally designed for statistical purposes

Regular assessment and selection of relevant websites via landscaping ensure data collected is pertinent to current statistical needs, reflecting user priorities [WCR]

Cooperation agreements with web sources for a good understanding of the content / data generation process [WSA]

Making data widely available allows for a broad range of user inputs, helping to monitor and improve the relevance and value of the data [ACD]

Distinguishing between different types of data (content, microdata, aggregated data) ensures that data is fit for different purposes and user needs [ACD]

Maintain **punctuality** despite unexpected changes and uncertain availability of websites

Continuous assessment and updating of website selections ensure the WIH focuses on sources that provide timely data, enhancing the overall timeliness of the collected data [WCR]

Data provision agreements with website owners that include specific data access or transmission formats independent of the design of the website [WSA]

Agreements that include provisions for alternative content retrieval methods, like file transfers or direct database access, ensure that data can be accessed even if website is temporarily unavailable [WSA]

Communication with website owners about content retrieval schedules and any changes ensures coordination, maintaining punctual data delivery [WSA]

Ensure **comparability** despite lack of standardization across websites in structure, terminology, update frequency and change practices

Regularly assessing and selecting websites based on quality criteria ensures that data sources are reliable and consistent, enhancing coherence and comparability [WCR]

Including provisions for standardized data formats in agreements with website owners ensures that data collected is comparable, regardless of the source [WSA]

Limit the **burden** on website owners, given the potential large amount of content retrieved from their websites

Limit the load on website servers, such as limiting the crawl rate, retrieving content during off-peak times, and optimizing retrieval strategies [WCR]

Use efficient crawling methods, such as leveraging sitemaps and indexing pages, to identify and retrieve relevant content without overloading the server [WCR]

Maintain communication with website owners to help address concerns regarding retrieval activities, enabling WIH to adjust practices to minimize burden [WCR]

Agreements to include specific conditions to reduce the burden on website owners, such as scheduling retrieval during low-traffic periods [WSA]

Use APIs or direct feeds with largest websites to ensure controlled and efficient content retrieval, reducing the server load and minimizing disruption [WSA]

Despite public nature of most web content, assure statistical web scraping does not jeopardise legitimate commercial interests of website owners (**data confidentiality**)

Clear identification of the WIH's web crawlers and transparent communication about data collection activities help build trust and ensure ethical content and data use [WCR]

Establishing formal agreements with website owners ensures that data retrieval is conducted under mutually agreed terms, protecting both parties' interests [WSA]

Identifying and appropriately handling personal data, statistically confidential data, and content protected under intellectual property legislation ensures that sensitive information is safeguarded [ACD]

Conclusions

Web Intelligence Hub (WIH) Overview

- Foundation: Essential for extracting quality data from web content for official statistics.
- Collaborative Effort: Involves Eurostat, NSIs, other statistical authorities, and partners.
- Use Cases: Online Job Advertisements, Online Based Enterprise Characteristics, Multinational Enterprises.

Rules and Procedures

- Modular Rules: Adaptable for future needs (retrieval, agreements, access).
- Transparency & Collaboration: Ensures trust and engagement with website owners.

How are Quality Concerns and Challenges addressed

- Accuracy: Thorough website selection, data provision agreements.
- Relevance: Regular assessment, cooperation agreements.
- Punctuality: Continuous updates, alternative retrieval methods.
- Comparability: Standardized formats, quality criteria.
- Minimal Burden: Optimized retrieval strategies, API use.
- Data confidentiality: Recognise all types of sensitive data, formal agreements.

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union