WIHOJA_OJAES_Q_4D_2023_0004

Reference Metadata in Euro SDMX Metadata Structure (ESMS)

Compiling agency: Eurostat, the statistical office of the European Union

Eurostat metadata

Reference metadata

For any question on data and metadata, please contact: Eurostat user support

1. Contact

Top

1.1. Contact organisation	Eurostat, the statistical office of the European Union
1.2. Contact organisation unit	A.5: Methodology; Innovation in official statistics
1.5. Contact mail address	European Commission - Eurostat - L-2920 LUXEMBOURG

2. Metadata update

Top

2.1. Metadata last certified	11/12/2023
2.2. Metadata last posted	11/12/2023

2.3. Metadata last update	11/12/2023

3. Statistical presentation

Top

3.1. Data description

The WIH-OJA is a micro-data set of classified Online Job Advertisements (OJA) produced by the Web Intelligence Hub (WIH) of Eurostat. Online Job Advertisements are retrieved from the web and classified into several variables by various algorithms (ontologies, Machine Learning models, fuzzy-matching techniques among others).

The WIH-OJA micro-data is at the moment at experimental stage.

The micro-data includes the following variables:

- occupation4d
- occupation3d
- occupation2d
- occupation1d
- skill
- city
- nuts3
- nuts2
- nuts1
- country
- first_active_date
- last_active_date

3.2. Classification system

The following classifications and breakdowns are used in the WIH-OJA microdata set:

- ISCO-08 – International Standard Classification of Occupations, used to classify occupations found in OJA
- ESCO - Multilingual classification of European Skills, Competences, and Occupations, used to classify Skills found in OJA (including Skill reusability level).
- LAU - Local Administrative Units. The LAUs are the building blocks of the NUTS and comprise the municipalities and communes of the European Union.
- NUTS 2021 levels 1-3 - The Nomenclature of territorial units for statistics, abbreviated NUTS (from the French version Nomenclature des Unités territoriales statistiques) is a geographical nomenclature subdividing the economic territory of the European Union (EU) into regions at three different levels (NUTS 1, 2 and 3 respectively, moving from larger to smaller territorial units). A Member State consists of one or several NUTS regions.

The following standard is used for the time related variables:

Unix epoch format for First active date and Last active date. Unix epoch format is defined as the number of non-leap seconds which have passed since 00:00:00 UTC on Thursday, 1 January 1970. Unix time is typically encoded as a signed integer.

3.3. Coverage - sector

All sectors of enterprises that post job advertisements in the World Wide Web, through one of the job portals included in the WIH.

3.4. Statistical concepts and definitions

The concept of the statistical unit is:

- Online job advertisements (OJA): advertisements published on the World Wide Web, revealing an employer’s interest in recruiting workers with certain characteristics for performing certain work. This could be motivated by the employer’s need to fill a current vacancy, by an exploration of potential opportunities, or other reasons. These advertisements normally include data on the characteristics of the job (e.g. occupation and location), characteristics of the employer (e.g. economic activity) and job requirements (e.g. education and skills). Part of this information is available only as natural language textual data.

The following concepts are used in the WIH-OJA data:

- Occupation (ISCO levels 1-4): the ISCO occupation that can be best identified in the online job advertissement. Occupations are closely related to job titles that are used in job postings to help attracting the right candidates and facilitate easier comparison of applicants' previous jobs and experiences. The list of variables associated to this concept are:
    - occupation4d
    - occupation3d
    - occupation2d
    - occupation1d
- Skill (ESCO levels 3): the skill, as defiined in ESCO ,identified in the online job advertisement. Skill is defined as "the ability to carry out the tasks and duties of a given job". The list of variables associated to this concept are:
    - Skill
- Skill reusability level: How widely a knowledge, skill or competence concept can be applied, according to ESCO. List of variables associated to this concept:
    - esco_v0101_reusetype
- ICT/Digital skill: indicates whether a skill is considered part of the ICT (Information and Communication Technology), or digital skills set. The variable associated to this concept:
    - esco_v0101_ict
- Place of work: city (LAU unit), region (NUTS 1-3) and country (Alpha-2 ISO 3166) of place of work identified in the text of the OJA. List of variables:
    - City
    - Nuts3
    - Nuts2
    - Nuts1
    - Country
- Advertisement active time: The period of time during which the job advertisement is available on the web. This includes the following two concepts:
    - First active date: the date when the advertisement was first published online. The first active date is determined by the publication date, if provided and extracted from the advertisement, or by the date when the advertisement was first detected in instances where the publication date is absent.
    - Last active date: the date when the advertisement is no longer available on the web. The last active date is determined by the date when the posting is not found in any of the sources, or its expiration date if mentioned in the advertisement. The list of variables associated to this concept are:
        - first_active_date
        - last_active_date

3.5. Statistical unit

Online job advertisement

3.6. Statistical population

Online job advertisements for jobs located in the reference area

3.7. Reference area

EU27 Member States, EFTA countries, The United Kingdom

3.8. Coverage - Time

From 01 January 2019 onwards

3.9. Base period

Not applicable

4. Unit of measure

Top

All micro-data variables are categorical. Statistics on number and percentages of job advertisements can be computed for the several classes of the categorical variables.

5. Reference Period

Top

Two reference periods are available in the micro-data, the first active date and the last active date. See statistical concepts and definitions for more information.

6. Institutional Mandate

Top

6.1. Institutional Mandate - legal acts and other agreements

The Web Intelligence Hub (WIH) is an integral part of the Trusted Smart Statistics (TSS) initiative within the European Statistical System (ESS). The creation and development of the WIH is aligned with the broader objectives of TSS.

As of now, there is no specific legal mandate or formal agreements that assign explicit responsibility and authority to the WIH for the collection, processing, and dissemination of statistics. This also applies to the WIH-OJA data.

The WIH is established within the ESS in response to various strategic initiatives, including the Scheveningen Memorandum, the Big Data Action Plan, and Roadmap. These initiatives aim to explore the advantages and challenges of using alternative data sources and data science techniques in official statistics.

The European Statistical System Committee (ESSC) adopted the Bucharest Memorandum on "Official Statistics in a datafied society (Trusted Smart Statistics)" in 2018. This marked a significant milestone in recognizing the importance of integrating new data sources and data science techniques into official statistics.

TSS is integrated into the European statistical programme, highlighting its strategic significance in advancing official statistics in the region.

Cedefop, an EU agency, playes a crucial role in exploring the potential of online job advertisements (OJA) data for statistics on skills. The ESS closely follows the development of the system and is engaged in collaborative efforts, including the European Big Data Hackathon, to exploit OJA data for policy-related questions. A Memorandum of Understanding between Eurostat and Cedefop was signed in 2020 with the objective to develop the above use case of OJA data. Cedefop, as one of the main users of the WIH-OJA data, with high expertise on the domain, continues to work together with Eurostat to further develop and improve the quality of the WIH-OJA data.

Eurostat is in the process of establishing formal Rules and Conditions for the operation of the WIH. These rules will ensure that web content retrieval activities align with established guidelines and comply with relevant regulations, including data protection laws, copyright regulations, and the European Statistics Code of Practice.

The WIH Rules and Conditions document is currently in draft form and is undergoing consultation with various stakeholders within the ESS. Once formal approval and adoption are completed, these rules and conditions will provide a clear framework for the WIH's activities, ensuring compliance with legal and ethical standards while facilitating the responsible use of web data for official statistics within the European Statistical System.

Upon formal approval and adoption, the WIH Rules and Conditions document will be published here and/or on the Eurostat website.

6.2. Institutional Mandate - data sharing

Cedefop and the Web intelligence Hub (WIH) operate as shared ressources within the European Statistical System (ESS). The WIH serves as a centralized repository, enabling National Statistical Institutes (NSIs) to access and utilize the same data resources. NSIs make use of this shared resource to dissseminatinate and use data at a national level, ensuring consistency and harmonization across statistical information shared within the ESS network.

7. Confidentiality

Top

7.1. Confidentiality - policy

Personal data and Automated Retrieval

The retrieval of personal data from websites is unavoidable when automated crawling and scraping is used. Only after the retrieval process it is possible to determine if personal data is present on a webpage or not. Examples of personal data that can be found on websites are names, personal addresses and email addresses.

Personal data stored and processed by the WIH, even if publicly available on websites, is treated as personal data according to European legislation on personal data protection, the General Data Protection Regulation (GDPR), or Regulation (EU) No 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices.

Data protected by statistical confidentiality

Data protected by statistical confidentiality is ruled by Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics. It protects all the data which allow a natural person, a household, an economic operator or other undertakings to be identified, either directly or indirectly.

For simplicity, we refer to data protected by statistical confidentiality as statistically confidential data.

Statistically confidential data can include both content retrieved from websites and the data derived from that content. Statistically confidential data may be found in the content retrieved from the websites, when it includes references to names or other identifying information. When such identifying information is kept in the data derived from the content, the derived data is statistically confidential data.

The identity of the source website when linked to the content retrieved or data derived is considered statistically confidential data.

This data is flagged internally by the WIH and is accessible only to official statistics authorities, in particular to the staff which require access to such data in the context of their responsibilities in the production of European and national official statistics.

Access to statistically confidential data is possible for scientific purposes as is done for other statistical sources, as stipulated in the Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics and Regulation (EU) No 557/2013 on access to confidential data for scientific purposes and for microdata access.

The statistical aggregates should not allow the indirect disclosure of statistically confidential data.

7.2. Confidentiality - data treatment

The WIH considers that data are protected by statistical confidentiality. For this reason confidential variables are only available for specific users.

8. Release policy

Top

8.1. Release calendar

Data are released between one and two months after the end of the reference period.

8.2. Release calendar access

Accessing the Release Calendar for OJA data is facilitated through our WIH wiki platform at the following page: WIH.OJA Data releases.

To ensure access to this resource, users are required to request rights to view this page. The Release Calendar provides an overview of scheduled OJA data releases, serving as a crucial reference point for stakeholders seeking information on upcoming data dissemination within the Web Intelligence Hub.

8.3. Release policy - user access

Confidential domain

The confidential domain includes all the content, data and metadata necessary for the activities of the WIH or that are the result of those activities. The activities include the retrieval of content from websites and the processing of that content for the extraction of information in the form of standardised statistical variables.

The confidential domain includes both non-sensitive and sensitive data of any type.

Access to this domain depends on the conditions determined for the producers, official statisticians and scientific researchers roles as defined in the rules and conditions.

Restricted domain

Content retrieved and data produced in the WIH may be made accessible to users for specific uses, under certain conditions in the restricted domain.

The restricted domain includes standard datasets which are available upon request, and tailored datasets following agreement between the WIH and the user.

The following uses are considered:

- Scientific research;
- Research and development activities in the interest of the production of official statistics based on content retrieved from the web;
- Analytical purposes of public interest, e.g. by public organisations; The data must not be used for purposes of taxation, investigation, regulation, or any other purpose having negative impacts on the website owner;

Besides non-sensitive data, the restricted domain includes the following two types of sensitive content or data:

- Content or data protected under intellectual property legislation;
- Content or data with potential negative impact on the website owner;

The restricted domain includes standard datasets, which are available upon request, and tailored datasets following agreement between the WIH and the user.

Data in the restricted domain is accessible by:

- Organisations that belong to the ESS, as listed in Eurostat website, and third parties or intermediaries acting on behalf of those organisations;
- Public organisations without a supervision role in the sector of the website owners;
- Members of the research community;
- Private organisations which access to the data do not pose a risk to the owners of the websites from where the content was retrieved;

Public domain

Although the goal of the WIH is the production of an intermediate product (microdata) to produce official statistics, some data may be released in the public domain when it has a general public interest.

Only non-sensitive data is available in the public domain, and it includes both microdata and aggregated data.

9. Frequency of dissemination

Top

Data are released on quarterly and annual basis, depending on the dataflow.

10. Accessibility and clarity

Top

10.1. Dissemination format - News release

Currently no press releases were planned for the OJA data.

10.2. Dissemination format - Publications

Currently the OJA data have been published in the following working paper:
Competition in urban hiring markets: evidence from online job advertisements — 2021 edition - Products Statistical working papers - Eurostat (europa.eu)

OJA aggregated data from the WIH have been recently published under the Statistics Explained article of Job Vacancy Statistics: Job vacancy statistics - Statistics Explained (europa.eu)

10.3. Dissemination format - online database

The OJA data are currently not available in a public database. A user needs to make a request before accessing the data.

To request access to the OJA databases, users may contact: ESTAT-WIH@ec.europa.eu.

10.4. Dissemination format - microdata access

The OJA microdata are available and accessible to users as described in concept 8.3.

10.5. Dissemination format - other

Skills-OVATE dashboards offer detailed information on the jobs and skills employers demand based on online job advertisements (OJAs) in 28 European countries.

Skills-OVATE are powered by Cedefop's and Eurostat's joint work in the context of the Web Intelligence Hub. Please visit the project page to find out the latest news and publications in the OJA work.

Skills OVATE dashboards can be found at: Skills-OVATE | CEDEFOP (europa.eu)

10.6. Documentation on methodology

Currently all methodological notes on the collection process of the OJA data is available in the online Web Intelligence Hub wiki. To access this wiki, users must send a request to ESTAT-WIH@ec.europa.eu.

10.7. Quality management - documentation

As of the current update, specific quality reports are not yet accessible. Future documentation will include quality reports, studies, and descriptions of quality procedures. The same applies to national quality reports.

11. Quality management

Top

11.1. Quality assurance

A data validation procedure is developed and runs every month, with various checks on data consistency, credibility (plausibility), distribution stability and outlier detection, among others.

A validation report is produced, highlighting the result of all validation rules (Warning, Error), as well as any previous Exceptions (Warnings that are considered plausible) on the data. Any Warnings or Errors in the validation report are followed up by taking corrective measures if necessary, or by adding an Exception to the validation rule.

Data annotation exercises are carried out, labelling the classified OJA data, which will in turn help with the assessment and evaluation of the accuracy of the OJA data.

The Web Intelligence Network, a community created around the Web Intelligence Hub to set up, animate and transfer knowledge on statistical methods and tools related to the use of web data in official statistics, supports the development of a quality framework and methodology of the OJA data, including review of definitions of the OJA data, landscaping of the data sources.

11.2. Quality management - assessment

Evaluation of the quality of the OJA data is an ongoing process. Eurostat plans to introduce quality indicators in the future to evaluate and monitor the overall quality of the OJA data.

In the context of quality assessment for algorithms classifying OJA data, an evaluation procedure is implemented, primarily focusing on the occupation variable.

The objective was to optimize the accuracy of the data, by synergizing human and machine intelligence, to enhance the efficiency of the classifier and to supporte human tasks through machine learning.

To further improve the quality of the applied algorithms, a gold standard dataset for WIH-OJA data is currently being developed.

Currently, labeled data collection has been executed specifically for the occupation variable. However, there are plans to routinely collect labeled data for other pertinent statistical variables in the future. This ongoing data labelling collection strategy reflects a commitment to continuous quality assessment and improvement in the classification algorithms used in OJA.

12. Relevance

Top

12.1. Relevance - User Needs

The main users of OJA data are the following:

- CEDEFOP, for policy making on skills and Vocational Training
- National Statistical Authorities, namely those participating to the Web Intelligence Network
- Eurostat Job Vacancy Statistics and ICT statistics, for producing statistics and indicators on occupations and skills
- Researchers and universities, for various research purposes

12.2. Relevance - User Satisfaction

No formal user satisfaction survey has been carried out so far. However, Eurostat collects feedback on user needs via direct consultations with key users of the OJA data very frequently. Based on this feedback and consultations, Eurostat tries to provide users with new indicators or with statistical variables that are relevant to them.

12.3. Completeness

For reference period of Q3 2023 the completeness of the collected variables are as follows:

Variable	Completeness
Occupation	100 %
Time (first_active_date, last_active_date)	100 %
Skills	97,8
NUTS1, NUTS2	79%, 71% respectively
Country	100 %

13. Accuracy

Top

13.1. Accuracy - overall

The OJA data may encounter some data accuracy issues, especially due to the use of different classifiers per language and country, but also due to the complexity and the amount of data that has to be classified in the data processing phase. More precisely, miss-classification may be encountered regarding the occupation (ESCO), skills (ESCO), geo localisation (NUTS), economic activities (NACE), as well consistency issues between classification levels or data releases.

Gold standard

Evaluation procedures of algorithms used to classify OJA are addressed, mainly for the occupation variable. These procedures combine human and machine intelligence to maximize accuracy, to assist human tasks with machine learning to increase classifiers efficiency. The creation of a gold standard for OJA data is one way to address this need for evaluation and quality improvement of algorithms.

Along with labelling of occupation variable, future collections of labelled data for other relevant statistical variables are planned.

13.2. Sampling error

Not applicable. No sampling is used in the production of OJA data.

13.3. Non-sampling error

Coverage error

The target population of job portals advertising potential jobs in the European labour markets may not be completely identified.

Two landscaping studies are carried out to create an inventory of all the websites (job portals, etc.) in scope for the OJA use case.

Undercoverage error may be present due to this inventory of sources being incomplete. Although possible, undercoverage should not be a significant source of error. The economic viability of job portals depend on their visibility and notability. Eurostat involved national labour market experts and used a harmonised methodology across countries, to discover and identify all relevant job portals, such as web search engines. One possible exception is very specialised job portals, which are active and notable only within a restricted community of experts.

A certain amount of undercoverage error, may be mitigated by the fact that many online job advertisements not retrieved from one source are anyway collected from other websites. This can happen when large job portals (aggregators) cover job advertisements from other (smaller or less relevant) job portals which are not included in the inventory of sources.

Overcoverage of OJA data may arise due to the fact that more than one source (website) is publishing the same advertisement, which would result to double-counting or multiple listing of certain OJAs in the data. The table below presents some indicators on double-counting of OJAs, for example the average number of sources an advertisement has been reported by more than one source.

Quarterly indicators exceeding the rule's threshold	**2019Q1**	**2020Q1**	**2020Q3**	**2022Q1**
Rate of change between the average numbers of sources per ad in two subsequent quarters	0.0378	0.041	0.0256	0.0161

Quarterly indicators exceeding the rule's threshold	**2019Q1**	**2020Q1**	**2020Q3**	**2022Q1**
Rate of change between the share of ads with more than one source in two subsequent quarters	0.1229	0.0583	0.0109	0.0263
Average number of sources per ad in the earlier of the two quarters	1.0395	1.0345	1.0361	1.0263
Share of ads with more than one source in the earlier of the two quarters	2.0676	2.0985	2.0936	2.1167

Measurement error

Web scraping of job vacancies on job portals faces measurement errors, that are mainly the result of scraping errors (scraper may download incorrect data from the web page), errors on the web page or

incorrect data on the web page (e.g. employers may upload incorrect data).

Non-Response error

In data from web content, like the WIH-OJA data, non-response errors can happen when one source of data (a job portal website) may either be "down", i.e. not operating, or may block the access to the data, by either block a crawler from scraping a website, or the block the API). In such cases the nonresponse error can be understood as an "unit non response" errors, where no data is collected from a unit (job portal).

Such events may happen frequently in a usually small number of sources in the WIH-OJA data. Details in such unit non-response errrors are documented and published with the methodological notes of the data release.

Processing error

Dealing with unstructured data text extracted from the web is challenging and errors may ocur in the processing and data preparation. More specifically, processing errors may arise in the OJA data due to the following processes: deduplication, merging, spam filtering, text processing, data cleaning, date detection, language detection.

Deduplication of job advertisements involves the identification and removal of duplicate entries within the dataset. Potential errors may arise during the deduplication of OJA data, specifically in the semantic deduplication stage of the pre-processing phase, which comprises two primary steps:

1. Metadata Matching: Utilizes job portal metadata (e.g., reference ID, page URL) to eliminate duplicate job ads on aggregator websites, following physical deduplication principles.
2. Logical Deduplication (Fuzzy Matching): Examines the content (description) of job ads, deeming them duplicates if the description and location closely match existing ads in the database, as part of logical deduplication.

Afterwards, a third deduplication phase compares structured fields from the information extraction stage. In the final comparison during the processing phase, ads are assessed based on publication date, occupation, place, skills, industry, and other distinctive characteristics. If ads share identical values in these fields, even if from different sources or phrased differently, they are considered duplicates and merged into a single advertisement. This step enhances the identification of duplicates with identical content present in classified variables.

Errors in **detecting the language** used in each OJA may introduce inaccuracies in the subsequent text processing and summarizing phases.

Errors may occur in **date detection** (first_active_date and last_active_date) using pattern matching techniques if the regular expressions used are not effecively customized.

Errors may arise during **noise filtering** of the downloaded web content. To detect true job advertisements in online advertisements of job portals from training courses or other non-related job-ad content, the system currently employs two distinct noise filters:

- No Vacancy Filter: designed to exclude content unrelated to job vacancies.
- Spam Filter: aimed at excluding postings that do not contain work opportunities, such as training courses.

These filters are language-dependent and undergo training on different datasets.

Sometimes incoherent classification between the label and the code of a given skill may arise due to processing.

Model assumption error

Various models are used in the WIH-OJA, e.g. to classify the text found on online advertisements to a particular position in a classification (occupation, skill, location).

Ontology-based models are used to classify advertisements according to the terms in the corresponding ontology. These models are used to search for an exact or similar match of the terms in the ontology and the natural language text in the title or description of the online job advertisement or of the metadata text.

Machine Learning models (classifiers): If the ontology models provides no results, a pre-trained machine-learning algorithm is used to classify the advertisement.

The most adopted machine-learning model is Naïve Bayes, which is trained on datasets using unsupervised learning techniques (for example, Word2Vec). Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from a finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle.

The process must be applied separately for each attribute (variable) considered and for each language, as each model must be trained to best fit the characteristic of the domain and the language.

The process starts with the creation of the training set. Training sets are defined separately for each language pipeline, depending on the available vacancies, the size of the country and the composition of vacancies. In general, each training set includes over 70,000 vacancies.

The training set is then divided into three parts:

- Training (60 %)
- Testing (20%)
- Evaluation, or quality measure (20%)

Training and testing can also be carried out on different kind of models to compare results and choose the best-suited model on a case-by-case basis. In all OJA cases, Naïve Bayes models have proved to be the most suitable, best-performing option.

Once trained, the machine-learning model can be used to classify vacancies in the processing pipeline.

14. Timeliness and punctuality

Top

14.1. Timeliness

The web scraping (data collection) of the Online Job Advertisements from the web sources takes place almost every day, depending on the crawler/website.

The web content is then processed and validated and the micro-data are then subsequently released at the earliest **one month after the end of the reference period**.

14.2. Punctuality

In exceptional cases there can be small delays in the release of the data.

15. Coherence and comparability

Top

15.1. Comparability - geographical

In the context of ensuring geographical comparability of data for online job ads, there exist two geographical aspects: the country/region where the job advertismeent is posted versus the country (or labor market) targeted by the job advertisements or potential employees. For example, an advertisement for a Data Scientist post in Luxembourg, may be advertised in a job portal in Germany.

The methodology of OJA data is highly comparable across countries. The landscaping and selection of the sources (web portals) is carried out using the same harmonised methodology across countries (i.e. the same keywords were used across countries to identify the sources) and over time.

The processing of the data (crawling, extraction of information, classification methods) are identical and consistent across all countries and all languages.

Nevertheless, classifiers tailored for distinct languages could introduce a potential challenge of non-comparable data across countries. The variances in cultural norms between counries might contribute to data incomparability. For instance, a job advertisement in certain countries may explicitly outline a particular skill, whereas in other countries, the same skill might be only implicitly implied.

Finally, technological changes affect comparability between countries because technological evolution is not uniformly distributed across countries. Depending on the level of adoption of technology, some countries may use more traditional ways to advertise their job postings, e.g. via small ads in newspapers or word-of-mouth, while some others may use online job advertissements as the main source of job posting.

15.2. Comparability - over time

As the raw content is stored, any potential changes of ontologies are systematically applied accross the entire time window of the data.

Moreover, the terminology used by employers in job advertisements is subject to change over time. Classifiers that proved accurate for older text may not be suitable for more recent job descriptions.

Similarly, algorithms fine-tuned for a given language on recent job advertisements may not be optimal for the same language used in older job advertisements.

Furthermore, the transition between versions of the international classification system, such as for skills (from ESCO V1.0.8 to V1.1.1), could impact the comparability over time of OJA data.

Posting job advertisements online has become a common practice in recent years, following the widespread adoption of the internet. However, it's important to note that not all job listings are found online. Certain industries or markets still rely on traditional methods such as newspaper ads, radio and TV spots, and even word-of-mouth to announce job openings. Over time, we can expect to see a growing number of job postings on online job portals. This increase is likely due to more companies recognizing the benefits of online recruitment, including the use of social media platforms for hiring purposes.

15.3. Coherence - cross domain

Online job advertisements (OJA) and Job Vacancies are two different but linked concepts and indicators.

Eurostat defines a Job Vacancy as follows (please see Job Vacancy Statistics Metadata):

A 'job vacancy' is defined as a paid post that is newly created, unoccupied, or about to become vacant:

(a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and

(b) which the employer intends to fill either immediately or within a specific period of time.

'Active steps to find a suitable candidate' include:

- notifying the job vacancy to the public employment services,

- contacting a private employment agency/head hunters,

- advertising the vacancy in the media (for example internet, newspapers, magazines),

- advertising the vacancy on a public notice board,

- approaching, interviewing or selecting possible candidates/potential recruits directly,

- approaching employees and/or personal contacts,

- using internships.

'Specific period of time' refers to the maximum time the vacancy is open and intended to be filled. That period shall be unlimited; all vacancies for which active steps are continuing on the reference date shall be reported.

Job advertisements and job vacancies, despite similarities, cover different populations and may differ in their definitions of a job opening. A job advertisement might not necessarily correspond to a vacant post and the company's intent to hire remains uncertain.

Thus, it is important to make the distinction between an online job advertisement and a job vacancy. In particular, it cannot be excluded that one single online job advertisement refers to several vacancies at the same time. Therefore it is considered that online job advertisements provide specific insights on the labour side of the labour market.

15.4. Coherence - internal

Not applicable

16. Cost and Burden

Top

Cost

The production of OJA data involve the following costs, taking as reference 2022. Direct running cost: 600 person days per year (2.7 FTE) of senior and junior specialists.

Burden

There is no direct burden on the respondents because content is retrieved directly from their websites. In some cases, website owners require the direct provision of the content, instead of having their websites content downloaded.

There is an indirect burden posed by the potential heavy load on the web servers when significant amount of content is downloaded. This is minimised by restricting the download to only new content.

Burden – number of websites: 413 [2022]

Burden – number of job advertisements: 65 million per year [2022]

17. Data revision

Top

17.1. Data revision - policy

Data are revised every quarter if errors are found (due to processing, or misclassification). Also the WIH-OJA data may be revised when a new version of a classification is introduced (e.g. from ESCO 1.0.8 to ESCO 1.1.1). These revisions are applied to all historical data.

Any revision of the data due to methodological changes, is explained with the release of data affected by such changes with explanatory notes.

17.2. Data revision - practice

Data undergo revisions for various reasons throughout the data pipeline. These revisions are crucial to maintaining the accuracy and reliability of the OJA data. The primary reasons for data revisions include:

1. **Error Correction:** Data is revised monthly to address any errors discovered during processing, such as deduplication issues, code errors, or misclassifications. These revisions are applied retroactively to ensure consistency in historical data.
2. **Classification System Changes**: Revisions occur when there are changes in the classification system. This involves updating the taxonomy version and potentially modifying models, whether machine learning (ML) or rule-based, to align with the new classification standards. This ensures that the dataset remains up-to-date with the latest classification methodologies.
3. **Methodological Improvements**: Data revisions may be prompted by enhancements in methodologies to achieve more accurate results. This typically involves refining models, whether through rule adjustments or ML model updates, to improve the overall quality of the dataset.
4. **Misclassifications and Model Errors**: Corrections are made in response to misclassifications or errors identified during model debugging sessions. This process involves improving variables or the dataset itself to enhance the performance and accuracy of the models.
5. **Code Errors**: Revisions are made when code errors are detected during debugging sessions in the data pipeline. This includes correcting errors in the source code of the data pipeline itself, as well as changes in the code used for training models.

All changes are meticulously tracked in Git repositories. These can include:

- Changes in the source code of the data pipeline
- Updates to the training set
- Modifications to the rules governing classification
- Adjustments to the taxonomy
- Changes in the code used to train ML models

The version control provided by the repository allows for a thorough review of changes, enabling the navigation through the history of revisions and ensuring transparency in the data revision process.

18. Statistical processing

Top

18.1. Source data

The source of OJA data is web content retrieved from job portals. Job portals were selected according to a landscaping study. The landscaping study identified 1000 job portals, from which 500 job portals were selected.

18.2. Frequency of data collection

Web content is collected continously. Job portals are visited daily.

18.3. Data collection

The data collection process of the OJA data can be described in the following steps, as can be seen in the following image:

Landscaping

The purpose of the WIH OJA landscaping exercise was to obtain a comprehensive and representative list of job portals with OJA data of good quality.

The sources (job portals) selection strategy has been implemented prior the start of the crawling activities, based on work and reports collected from International Country Experts.

To explore the available data on the use of online job-portals in recruitment and job-search in the 28 European Countries, International Country Experts (ICEs) were engaged in the Landscaping Activity.

This process is composed of 4 main steps:

1. Source selection in landscaping
2. Augmentation of source list
3. Agreements with source data providers
4. Coverage assurance

At the end of the source selection process all the sources have been tested to identify potential issues related to the content of the site, to possible technical problems or to the unavailability of the site.

Data Ingestion

This phase includes all the tasks needed to obtain and import raw data from web portals and to store them into a database. The key point in the ingestion phase is to ensure a stable data flow preventing potential loss of data due to harvesting issues (i.e. due to incomplete or inaccessible information). The data ingestion phase deals with both structured sources, which store information in structured fields, and unstructured sources, where information is largely extracted from large chunks of text.

The data ingestion activity includes the collection of data through:

1. Scraping
2. Crawling
3. Direct access (e.g. API)

Deduplication

Deduplication is a fundamental step of the OJA data processing: the ability to detect duplicate postings is necessary to obtain a high-quality set of results. In order to improve deduplication performance, this task is split in three distinct parts performed during the whole data processing pipeline:

- Physical deduplication
- Semantic deduplication
- Logic deduplication

The data ingestion phase is in charge of the first part (physical deduplication): postings coming from the same URL are considered as duplicates and not passed to subsequent phases. Metadata matching using metadata deriving from job portals to remove job advertisements duplicates on aggregators' websites (e.g. reference id, page url). The two remaining phases will be described while depicting the whole process.

The pre-processing phase is in charge of the semantic deduplication: this type of deduplication is based on content and is mainly composed of two phases:

- Logical deduplication or fuzzy matching carried out on the description (or content) part of the job advertisement. An OJA is considered a duplicate if the description (and the job location) is the same (or very similar) of another job ad already present in the database.
- Metadata matching using metadata deriving from job portals to remove job advertisements duplicates on aggregators' websites (e.g. reference id, page URL)

After the physical deduplication step (based on the URL or the unique id of the vacancy) and the logical deduplication (based on text similarity) a third deduplication step has been added to compare structured fields resulting from information extraction phase. At the end of the processing phase resulting advertisements are compared by:

- Publication date
- Occupation
- Place
- Skills
- Industry

Distinct advertisements from the previous 2 phases with the same values for these fields are considered as duplicates and merged into a single advertisement. This additional step allows the identification of duplicates posted on different sources, written in different ways but all containing the same detailed content after the classification phase.

Pre-processing

The pre-processing is a critical step of the data processing pipeline. This phase must face some critical issues related to the particular type of data collected:

- Presence of noise in the source (information not directly related to OJA)
- Uniform coverage among all countries and occupations despite different languages and vocabularies
- Non-existence of a common standard to submit information

- Unstructured data (text)

Moreover, to obtain usable results the following challenges must be faced:

- How to measure the quality of collected data.
- How to increase the quality of the data.
- Which variables affect quality of data.
- Duplicated job ads
- Availability of metadata
- Timeliness and periodicity of the data
- Complexity of data
- How to keep track of quality of data.

The number of words collected is not considered an interesting measure as it depends mostly on the language characteristics and on the way the OJAs are written. The number and the quality of features extracted is considered much more important and will be monitored and verified step by step.

The pre-processing phase can be divided into 3 steps:

- Merging
- Cleaning
- Text processing and summarising

OJA Information Extraction

The information extraction process is defined through a set of processing pipelines. A pipeline can be defined as a part of information extraction dealing with a specific content (attribute or variable) and with a particular language. Several pipelines can be combined to define the whole information extraction process. During the processing of each pipeline, advertisements are analysed to classify the contents of the pipeline (contract, occupation…) according to the specified language.

In practice, the information retrieval process is composed of:

- One pipeline for each language considered
- One pipeline for each attribute (statistical variable) to be classified:
- occupation
- skills
- etc.

18.4. Data validation

To check that the data does not display implausible values that are unlikely to reflect genuine statistical variation, a data validation process has been set up for OJA data.

Data validation is defined as an “activity verifying whether or not a combination of values is a member of a set of acceptable combinations”, but this activity must be intended as a dynamic process (Eurostat’s Methodology for data validation manual).

Every time that a data release is planned, statisticians check data compliance with the validation rules. When the data does not conform with the rules, one (or more) of three outcomes can occur:

- Data is revised and improved

- Information on the reasons for anomalous data behaviour is collected, contributing to the improvement of meta-data
- The rule is modified or dropped

Therefore, while the validation process is a fundamental tool for data improvement and data quality control, its rule set is always amenable to change. For this reason, feedback on the validation rules and on the validation process is always welcome from statisticians and data users.

Validation rules

The validation rules currently applied to the OJA dataset covers a variety of statistical properties of the data. The most common (i.e., the ones that apply to the largest number of OJA variables) are:

- Consistency with Eurostat's official code lists (e.g. occupation codes should follows the ISCO08 classification)
- Consistency within hierarchical classifications (e.g. the NUTS3 region of an ad must be contained in the NUTS2 region in which it is classified)
- The distribution of ads within categories of a classification should be reasonably stable over time and across data releases (this is a plausibility rule, meaning that its violation raises a 'Warning', which is investigated further)
- When a variable follows a certain classification, the distribution of ads across the categories should not be too unbalanced (e.g. if 50% of all ads were found in a single 4-digit ISCO occupational category, or if no data is found for dozens of occupational categories, this raises a 'Warning' which is then investigated further)
- For some variables (e.g. date and country), there should be no missing data
- For string variables, some formatting requirements are usually imposed (e.g. there should be at least one alphabetic character in every company name)

Besides these rules (and other minor ones) operating at the level of the data record, or of the distribution, there are some others that check the consistency of the whole database against basic logical or numeric requirements. These rules (also called structural validation rules) include for example checking for correct naming of datasets and variables, absence of empty fields, etc.

Furthermore, country experts regularly check, on a random set of OJAs, if the text found on OJAs is correctly allocated to the right position in the various classifications used (ISCO for occupations, ESCO for skills, NUTS for regions, etc..).

Additional information on the validation procedure, the validation rules or validation reports the can be found on the following wiki page of the Eurostat's WIH: Validation rules. The access to this page is nevertheless restricted to a specific community of users. An access grant can be provided upon request by sending an email to the functional mailbox ESTAT-WIH@ec.europa.eu.

18.5. Data compilation

Not applicabe.

18.6. Adjustment

Not applicable.

19. Comment

Top

Not applicable.

Related metadata

Top

Annexes

Top

Footnotes

Top