

Terms of use of the data made available by Eurostat for Statistical Awards Programme – Web intelligence Deduplication challenge

For the purposes of the **Statistical Awards Programme – Web intelligence Deduplication challenge** ('WI Deduplication challenge'), the Statistical Authority of the European Union ('Eurostat') undertakes to provide a dataset containing Online job advertisements (OJA) data.

The dataset remains the exclusive property of Eurostat, which agreed to the use of the data for the WI Deduplication challenge provided that the conditions set out below are respected.

Nature of the data

An online job advertisement (OJA) is information posted online about a job opening targeted at potential job-seeking applicants. Online job advertisements are usually published on job posting websites, company websites and portals for job seekers. Online job advertisements contain a job description, information about the company looking to hire, job benefits, requirements for job seekers, etc.

The competition dataset contains 112k online job advertisements (OJA), retrieved from around 400 websites active in the European Union, which have been anonymised for the purposes of the competition (i.e. any reference to personal data have been removed, as well as the identification of the website(s) from where the OJA were retrieved). The competition organisers have taken unique real-world OJAs and created full, semantic, temporal and partial duplicates across different languages, thus creating a synthetic multilingual dataset for the competition.

The source of the original dataset is the European Web Intelligence Hub (WIH), where around 200 million online job advertisements have been collected and classified since July 2018. The WIH is a shared infrastructure of the European Statistical System, managed by Eurostat (the European statistical authority) and encompassing the statistical authorities of the European Union Member-States. The extraction and use of job advertisements for labour market and skills intelligence is done in collaboration with Cedefop, the European Centre for the Development of Vocational Training.

The data and dataset are to be treated as confidential and used solely for the purpose of the Statistical Awards Programme WI Deduplication challenge.

Safekeeping of the data

The confidential data, including any intermediate results allowing the extraction of record level data, must be stored in a secure environment. Access to the data must be restricted to the registered members of the teams participating in the WI Deduplication challenge, who explicitly accepted the terms of use set out in this document.

Duration of data access

Access to OJA's WI Deduplication challenge data (at record level) is granted only for the period of the WI Deduplication challenge and until the submission deadline for the Accuracy Award **on the 31 March 2023**.

Eurostat shall immediately block or terminate access to the data if the participant has not fully ensured compliance with the conditions and obligations referred to in these terms of use. In the event of non-compliance, Eurostat shall block the access and request the participant in writing to rectify the situation within one hour. In the absence of rectification, termination shall be effective on the subsequent day with Eurostat's confirming the termination through a registered letter to the team whose member breached the rules and conditions of the data use.

After expiry of data access

After expiry of the access to the OJA dataset (set for **31 March 2023**), the participant must destroy all local copies of the record level data.

The participants must not make further use of the information made available to them by Eurostat after the end of the WI Deduplication challenge or termination of the access to data. Failure to comply with this requirement shall result in liability to claims for damages and to penalties.

Agreement to the terms of use

By agreeing with these terms of use, you assert that you will be bound by all the conditions listed above and you will:

- (a) use the dataset only for the purposes of the WI Deduplication challenge and production of report based on the prototype there developed;
- (b) safeguard the dataset and any usernames and passwords associated with it;
- (f) preserve the full confidentiality of record level data provided for the purpose of WI Deduplication challenge;
- (g) destroy the copied record level data of the OJA dataset, or intermediate results allowing the extraction of record level data at the end of the data access period;
- (h) abide by any other conditions notified to you by Eurostat;
- (i) inform Eurostat immediately about any breach of the confidentiality rules laid down in these terms of use.

You will not:

- (a) allow non-authorized users to access the dataset (authorized users are named in the list of participants);
- (c) use the data in any manner that could cause harm to Eurostat or its data providers/sources of information, including attempts to identify individual persons, households and/or source websites represented by dataset records, or claims to this effect;
- (d) release or publish any information or results based on the dataset made available for the WI Deduplication challenge;

You certify that you have read all of the above clauses, you understand that you are accountable for correct and responsible use of the data and data access system, and you understand that, if you fail to comply with these clauses, your access to the dataset will be withdrawn and you will be liable to any other sanctions that may be specified in the applicable civil or penal law.

Team name_____

Signature_____

Print Name_____

Date_____