Online Job Advertisements

Landscaping Methodological Guide

Francesco Trentini

Web Intelligence Hub

Landscaping OJA Web data sources

Deliverable D1.1 – OJA landscaping methodological guide

Authors: Emilio Colombo, Anna Giabelli, Fabio Mercorio, Mario Mezzanzanica, Francesco Trentini

Approved by: Emilio Colombo, Mario Mezzanzanica

Version: 3

Date of Release: 2022-01-21

Conducted for Eurostat under Specific Contract No 2020.0399

Framework Contract 2020-FWC7-AO-DSL-VKVET-JBRAN-WIH-OJA002/20 between Cedefop and the Universitá Degli Studi di Milano-Bicocca, Burning Glass Europe S.R.L. and GOPA Luxembourg SARL - Towards the European Web Intelligence Hub - European system for collection and analysis of online job advertisement data (WIH-OJA)

Contents

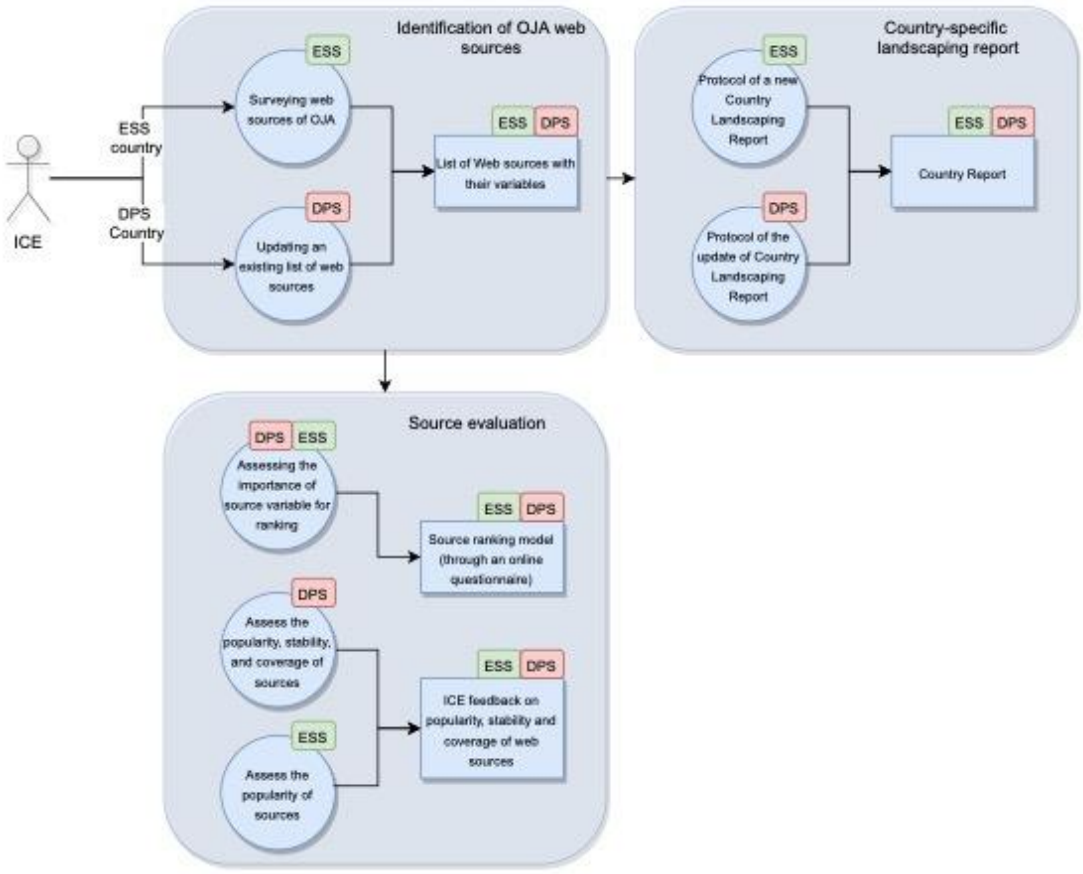
1. Introduction

This guide presents the rationale and operative tasks to perform the sources analysis of OJA and the identification of the characteristics of the OJA market in a country.

The procedure is divided into 4 main steps. In Section 2 we present the identification of OJA web sources. We discuss in detail the steps that International Country Experts (ICE) should implement to identify, select and assess new and existing sources. Section 2 covers the features of each country-specific report, defining a standardised protocol. Section 3 explains how the information gathered in previous steps is harmonized and centrally processed so to obtain a ranking model of sources based on multidimensional quantitative information.

Each step separates the specific tasks required to perform a new survey (in the specific case, the so-called ESS countries: Norway, Iceland, Switzerland and Lichtenstein) from those to perform an update of existing sources (for the so-called DPS countries).

Figure 1 Workflow of ICE interaction



2. Identification of OJA web sources

This step is aimed at identifying websites that advertise job vacancies. The outcome will be a list of sources and a detailed characterization of both the website characteristics and the OJA pages features. The activity differs for ESS countries that are running the survey for the first time, and DPS countries, which are surveying only new sources.

The activity is composed of 4 steps for DPS countries and 3 steps for ESS countries.

Step 1. Keyword translation in national languages. Keywords are provided in English to maximize the standardization of the search activity while leaving to the ICE the task of finding the translation that more accurately describe the same content in their national language, therefore increasing the consistency of the results. ICEs that report that the keywords shall be enriched by adding new terms in order to adapt to country specific features of the market are asked to provide a list of new keywords. These keywords are labelled as new and all websites emerging from queries based on such keywords are also labelled to identify which was the data generating process – the query on standard keywords or on additional keywords.

Step 2. List of job portals. Each keyword is used by the ICE to run a query on google.com and register all the search results that are produced by the search engine. The reason for the choice of google.com is manyfold. First, Google is almost a monopolist in web search: about 90%[1] of queries are processed by it. Since our aim is to simulate the typical search behaviour of a jobseeker, Google is likely to be the selected search engine by most of them. Second, this makes the standardization of the process of website listing free from any ex-post computations, which would be required in case of use of more than one search engine. Third, the calculation of popularity can take advantage of Google Trend which uses the same underlying algorithm that lists web sources, increasing the coherence of the procedure by minimizing the private data generating processes.

Step 3. New web sources. This step is not performed in ESS countries. Each website that is surveyed is processed to assess whether it is a new source for OJA, a known source or it is not a source of OJA. An algorithm evaluates whether a source is already present in the list of known sources and prompts a warning message. ICEs are provided the whole list of known sources and are asked to validate the output of the algorithm, approve the output or correct it. The ICE can also declare a website as spam if it does not advertise job vacancies (e.g., it advertises training courses, guides to write a CV and similar contents).

Step 4. Features of new sources. The websites that are categorized as new sources are analysed in depth by the ICE. This activity is needed to prevent the well-known *"garbage-in, garbage out"* phenomenon that might happen by selecting all the sources without assessing their quality, thus affecting the overall quality of the analysis of Web sources. The standardization of data gathering is ensured by a data validation procedure that constrains the possible values. A definition of each variable and its metadata is provided to the ICE, as well as the operative steps that are required to accomplish the activity, by means of contextual information and guidance.

[1] Source: https://gs.statcounter.com/. Last accessed: 8th October 2021

2.1. Surveying web sources of OJA: ESS countries protocol

This section describes the protocol for countries that are participating for the first time in surveying web sources of OJA. The procedure takes advantage of the previous iterations and is kept in line with the DPS countries procedure.

Annex 1 represents the operative tool that ICEs receive to perform their task. The deliverable is the completed Annex document itself. This document is divided into five parts:

1. **Readme**: an introductory sheet summarizes and explains the steps that the ICE must perform to complete the activity (Table 2). The use of hyperlinks let the ICE navigate the document and minimizes reference errors.

   Table 1 Section of the Readme sheet that describes in detail the operative steps.

Step 1	T1- keywords	1. Please provide a translation in your national language(s) of the suggested keywords.
Step 2	T2- new sources	2.1 Clean your web browser cache (step-by-step guide).
		2.2 Run a query on google.com with each of the translated keywords and transcribe the results from the first two pages that are job portals or aggregators.
Step 4	T4- variables	4. Collect information from each website and an instance of the OJA page. Metadata for each variable are available in separate sheets, each named after the variable it refers to, and are accessible via a hyperlink (just click on the variable name in the sheet). Click on the variable name in the glossary to move back to its column in sheet "T4- variables".

2. **T1 – keywords:** the sheet presents 2 main columns. The first column includes a number of suggested keywords in English, which the ICE is required to translate in her national language and transcribe in the second column. In countries with more than one official language, additional columns are added to host a separate translation for each of the official languages, so that the output is a column for each language. The keywords are:
   - "Job search"
   - "Job Offers"
   - "Online job search sites"
   - "Find a job"
   - "Job ads"
   - "Job recruiting websites"
3. **T2 – sources**: the sheet is organized in two columns, the first "Website name" to register the website name, the second "Website address" for the website address. The website address must be copied from the browser address bar and must be the complete address of the landing page of the job-portal.
4. **T4 – variables**: The sheet is used to guide the gathering of characteristics of both the websites and their OJA pages. For each website, website variables (W) and content variables (C) need to be filled according to metadata and operative instructions.

5. **Sheets W1-C9:** Each sheet explains a variable of the sheet "T4 – variables" by reporting the possible values and their definition. Moreover, a brief description of the process of information retrieval and interpretation is given. The following list gathers website variables (W) and content variables (C) with their values and description. Please refer to Annex 1 for complete reference.

   - **W1. Position in Google ranking**: The variable records whether the website was listed on the first or second page of Google Search results.

Value	Definition
first page	The website is listed on the first page of the Google Search results
second page	The website is listed on the second page of the Google Search results

   - W2. Type of job-portal:

Value	Definition
primary job-portal	Portals that advertise vacancies for which the user can apply on the portal itself and do not redirect to another website.
secondary job-portal	Portals that advertise vacancies that are published on other websites.
combination of primary and secondary functions	Portals that advertise both vacancies directly and vacancies ads published elsewhere.

   - W3. Type of operator:

Value	Definition
classified ads portal	General ads portals containing ads for job vacancies.
company website	The website is owned by a company that advertises internal jobs.
job search portal	Websites collecting OJA from different sources and presenting them in an integrated search engine (aggregators).
national newspaper	Job advertisement pages of online newspapers.
public employment service	The website of portals of national employment services.
recruitment agency	The website of a private employment agency that connects employers with jobseekers.

   - **W4. OJA Volume**: The ICE must report the number of advertised vacancies if the website displays it; otherwise, they leave it blank. The number of vacancies refers to the period in which the mapping of new sources is made, i.e. in the time window 29th September 2021 and 11th November 2021.
   - **W5. Geographical scope:** The ICE is informed that the focus is on the coverage of the website and not on the detail of the vacancy advertised in it.

Value	Definition
international	The source advertises job vacancies for more than one country.
national	The source advertises job vacancies inside a country.
regional	The source advertises job vacancies for a bounded area of a country, such as a province, a metropolitan area or a city.

- W6. Sectoral scope:

Value	Definition
one industry	The website advertises job vacancies related to a single industry. If so, please move to W6bis.
all industries	The website advertises job vacancies related to more than one industry. If so, please skip W6bis and move to W7.

- W6bis. Sector: If the variable "W6 - Sectoral scope" takes the value "one industry", the ICE assigns it to the most appropriate one. It is possible to input a free text to characterize the sector.

Value
A- Agriculture, hunting and forestry
B – Fishing
C- Mining and quarrying
D – Manufacturing
E- Electricity, gas and water supply
F – Construction
G- Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
H- Hotels and restaurants
I- Transport, storage and communication
J- Financial intermediation
L- Public administration and defence; compulsory social security
M – Education
N- Health and social work
O- Other community, social and personal service activities
P- Activities of households
Q- Extra-territorial organizations and bodies
Input the description in textual form.

- W7. Publication date: The ICE is asked to visit a few OJA pages and check whether a structured field for the publication date is present. In order to improve the clarity of the variable, it shall be renamed “W7. Type of publication date” and is recoded to “structured/not structured” respectively.

Value	Definition
present	A publication-date structured field is visible on the page.
not present	The publication date is not provided in a structured field.

- W8. Expiry date: The variable is coded with the same convention of "W7. Publication date". Analogously to improve the clarity of the variable, it shall be renamed “W8. Type of expiry date” and is recoded to “structured/not structured” respectively.
- W9. Update frequency: ICEs are advised that the focus of the variable is on the distribution of listed ads. They are required to visit a list of ads and check whether the time interval between entries is a day or more. The values are limited to two – “daily” or “other” –

because it is practically difficult for an observer to assess it for time spans that are larger than a day.

Value	Definition
daily	Entries in the ads list are posted at a time distance of a day.
other	Entries in the ads list are posted at a time distance of more than a day.

- W10. Languages: ICEs are asked to list (separated by a comma) all the languages that are used on the website to advertise vacancies. ICEs are also informed that usually portals advertise which are the languages of the ads they host. If this information is not available on the website, ICEs must look for OJA that may be posted in other languages, using their knowledge of the labour market in their Country.
- W11. Publishing option: The ICE is advised that the focus of the variable is on the type of service offered by the platform, which is usually presented in the "pricing" or "services" sections.

Value	Description
free advertisement	The user can advertise a job vacancies free of charge.
paid advertisement	The user pays a fee to advertise a job vacancy.
both options	The user may either post an advertisement free of charge or by paying a price (e.g. for additional services)

- C1. Type of the occupation:
  Please refer to the Annex to access a high-quality image of the example provided to the ICEs.

Value	Definition
structured	The information is displayed in a standard field. (highlighted in green in the example)
textual	The information is displayed in a text box which is not standardized. (highlighted in blue in the example)
both	The information is displayed both in a standard field and in a text box which is not standardized. (as in the example)
not available	The information is not displayed in the web page

- C2. Type of contract: The variable is coded with the same convention of "C1. Type of the occupation"
- C3. Working time: The variable is coded with the same convention of "C1. Type of the occupation
- C4. Sector: The variable is coded with the same convention of "C1. Type of the occupation
- C5. City: The variable is coded with the same convention of "C1. Type of the occupation
- C6. District: The variable is coded with the same convention of "C1. Type of the occupation
- C7. Region: The variable is coded with the same convention of "C1. Type of the occupation
- C8. Qualification level: The variable is coded with the same convention of "C1. Type of the occupation
- C9. Wage: The variable is coded with the same convention of "C1. Type of the occupation

2.2. Updating an existing list of web sources: DPS countries protocol

Annex 1 represent the operative tool that ICEs receive to perform their task. The deliverable is the filled Annex document itself. This document is divided in 6 parts:

1. **Readme**: an introductory sheet summarizes and explains the steps that the ICE must perform to complete the activity (Table 2). The use of hyperlinks let the ICE navigate the document and minimizes reference errors.

Table 2 Section of the Readme sheet that describes in detail the operative steps.

Step 1	T1 - keywords	1. Please provide a translation in your national language(s) of the suggested keywords.
Step 2	T2 - new sources	2.1 Clean your web browser cache (step-by-step guide).
		2.2 Run a query on google.com with each of the translated keywords and transcribe the results from the first 2 pages that are job portals or aggregators.
Step 3	T2 - new sources / T3 - known sources	3.1 Compare the new list with the list of websites already surveyed in the running version, presented in the sheet "T3 - known sources". The message "May be present in know sources" is displayed next to the source if it is found in "T3".
	T2 - new sources	3.2 In Sheet "T2 - new sources", flag the variable "New source" with "Yes" if the source is not listed in "T3", "No" if it does or "Spam" if the source does not advertise job vacancies (e.g. it advertises training courses, guides to write a CV and similar contents).
Step 4	T4 - variables	4. Collect information from each website and an instance of OJA page. Metadata for each variable are available in separate sheets, each named after the variable it refers to, and are accessible via hyperlink (just click on the variable name in the sheet). Click on the variable name in the glossary to move back to its column in sheet "T4 - variables".

2. T1 – keywords: the sheet presents 2 main columns. The first column includes a number of suggested keywords in English, which the ICEs are required to translate in their national language and transcribe in the second column. In countries with more than one official language, additional columns are added to host a separate translation for each of the official languages, so that the output is a column for each language. The keywords are:
    - "Job search"
    - "Online job search sites"
    - "Find job"
    - "Job ads"
    - "Job recruiting websites"
3. T2 – sources: the sheet is organized in three columns, the first "Website name" to register the website name, the second "Website address" for the website address and the third for a flag

"New source". The website address must be copied from the browser address bar and must be the complete address of the landing page of the job-portal. Once the websites are listed, a lookup procedure automatically scan the sheet "T3 – known sources" check. If it is the case a warning message is prompted "May be present in known sources". Then the ICE compares the sources and fills the third column, "New source", to flag the type of source with three admitted values: "Yes" (if the source is not listed in "T3"), "No" (if it does) and "Spam" (if the source does not advertise job vacancies (e.g. it advertises training courses, guides to write a CV and similar contents)).

4. T3 – known sources: the sheet reports a list of sources already surveyed in the previous landscaping exercise and currently in use.
5. T4 – variables: The sheet is used to guide the gathering of characteristics of both the websites and their OJA pages. For each new website, website variables (W) and content variables (C) need to be filled according to metadata and operative instructions – refer to the next section for a detailed account of each variable.
6. Sheets W1-C9: Each sheet explains a variable of the sheet "T4 – variables" by reporting the possible values and their definition. Moreover, a brief description of the process of information retrieval and interpretation is given. The following list gathers website variables (W) and content variables (C) with their values and description – refer to ESS countries section for a detailed account of each variable. Please refer to Annex 2 for the complete reference.

2.3. Training and support

The required actions to perform the activity are explained in a kick-off online meeting during which an example of the activity has been presented. The meeting has been registered and shared as a support material, together with the slides used. Support is provided by CRISP staff members via email or video calls.

2.4. Transmission of the documents and performance monitoring

The documents are delivered to the ICE by uploading them on a cloud drive service – in this specific case they are hosted in a GSuite Drive, provided by University of Milano-Bicocca – and ICE are required to fill the spreadsheet online (using Google Sheets). This arrangement has several advantages: transmission issues and compatibility of editors are taken care of; monitoring of the progress can be done live by the administrators, by means of dashboards and analytics tool.

2.5. Feedbacks and iterations

Feedback constitutes an important mean of improvement of the procedure that may require adjustments based on some specific features of each Country, which may have not been included in the first place. The use of cloud services allows to implement changes in the procedure on the go, e.g., modifying the structure of the deliverable while preserving the filled content.

A changelog document (Annex 3) is registered, including the following information: date, countries involved, section that is addressed, description of the content of the change.

3. Country-specific landscaping report

Source selection is complemented by an analysis of the Online Job-vacancies market. Each ICE is asked to write/update a landscaping report based on a standardised template, complemented by data analysis provided by CRISP and performed centrally.

3.1. Protocol of a new Country Landscaping Report: ESS countries

The following protocol presents how the country-specific landscaping report is produced. The report for the new countries shares the same structure of the previously produced reports to promote a high-level standardization.

Each ICE will be provided with a document that includes desk research and data analysis produced by CRISP and guided comments to specific sections that need to be updated. A detailed template is available in Appendix 4.

More specifically CRISP has collected and elaborated data from the following sources

- LFS microdata (Employment trends, employment distribution by sector/occupations, numbers of new hires etc.)
- Information society indicator [isoc_ci_ac_i], from Eurostat. (Internet usage, internet access etc.)
- Digital Economy and Society Index (DESI). Individual country reports and dataset

These will create a rich set of data and statistics which will be made available to each ICE. The template, presented in Annex 4, reports the structure of the previous landscaping report that ICEs are expected to follow. In the template two elements are inserted. One is a data section that describes the data (from the sources above) that will be provided to ICE who will be required to comment upon. The other element is a comment section that contains a set of questions that are designed to guide ICEs in drafting the country report.

3.2. Protocol of the update of Country Landscaping Report: DPS countries

DPS countries have already produced a landscaping report in 2018 which is updated.

The 2021 update needs to pivot on the main strength of the previous country landscaping report and to address the existing inconsistencies, while using the OJA data of the current version of to inform the current landscaping phase. The availability of new data on OJA sources allows to use some of the available information as feedback to ICE who can use it to improve the report.

These goals operatively translate into keeping the main existing structure of the reports and introducing some novel elements with a high level of standardization. At the same time, it is crucial to keep in focus the role of the ICE, namely the role of validating evidence to produce valuable knowledge and providing qualitative insight.

ICEs are provided a document analogous to the one provided for ESS countries including desk research and data analysis produced by CRISP. This document will include the same sources previously identified (LFS microdata, Digital Economy and Society data, DESI reports). The document will be further integrated with data on overall characteristics of OJA as derived from the available dataset (N. of sources, N. of sources by type, market concentration, trends in OJA by sector and occupation etc.) ICEs

will be provided guided comments to specific sections that need to be updated. A detailed template is available as Annex 5.

The template reports the structure of the previous landscaping report that ICEs are expected to critically assess and update. In the template two elements are inserted. One is a data section that describes the data that will be provided to ICE who will be required to comment upon. The other element is a comment section that contains a set of questions that are designed to guide ICEs in commenting and updating the country report.

4. Source evaluation

The field work conducted by the ICEs produces information on every source and a general overview of the OJA market in the Country. The first constitutes the initial inputs for the raw data extraction, while the second is a contextual information that allows us to interpret the scope and completeness of the survey conducted on websites. After this assessment, the next stage is the evaluation of the sources itself. This activity is divided in three parts:

- The first part is the encoding of categorical website variables into numerical ones (3.1);
- The second part is **the synthesis of the ranking model** that considers each website variable having equal importance;
- The third part is the **assessment of website characteristics (3.3) based on three criteria**: **popularity** (Section 4.3.1), **stability** (Section 4.3.2) and **coverage** (Section 4.3.3). While the study of popularity can be produced for both countries that run the landscaping procedure for the first time (EES countries) and for countries that update it (DPS countries), the two latter evaluations require some historical data on which to be performed (DPS countries only).

4.1. Transform categorical and ordinal variables in numerical variables

The goal of this step is to transform the categorical variables identified in the protocol described in Section 2 to numerical values. The two criteria that we want to satisfy are the following:

- Numerical values are assigned following the relative importance that each value bears with respect to the other.
- Include the preferences of all the involved stakeholders.

Notice that the AHP model presents two distinct features useful in our context:

1. First, it highlights inconsistencies in evaluating preferences to participants, enabling them to revise their judgments accordingly[2];
2. Second, it allows computing the consensus among participants, at each level of the taxonomy. This means the decision maker can have a fine-grained analysis of the degree of agreement for each variable value identified, and to decide accordingly.

With these two characteristics in mind, we want to use a model that can integrate different point of views on the relative importance of every value of a variable in a transparent and accountable way. The

[2] Intuitively, as AHP deals with independent criteria, this means that if criterion A is better than B, and B is better than C, we expect a consistent judgment requires A to be better than C (transitivity).

Analytic Hierarchy Process (AHP) model is designed to handle such a scenario. The use of an AHP model and user settings guarantees trust and explainability of the model results.

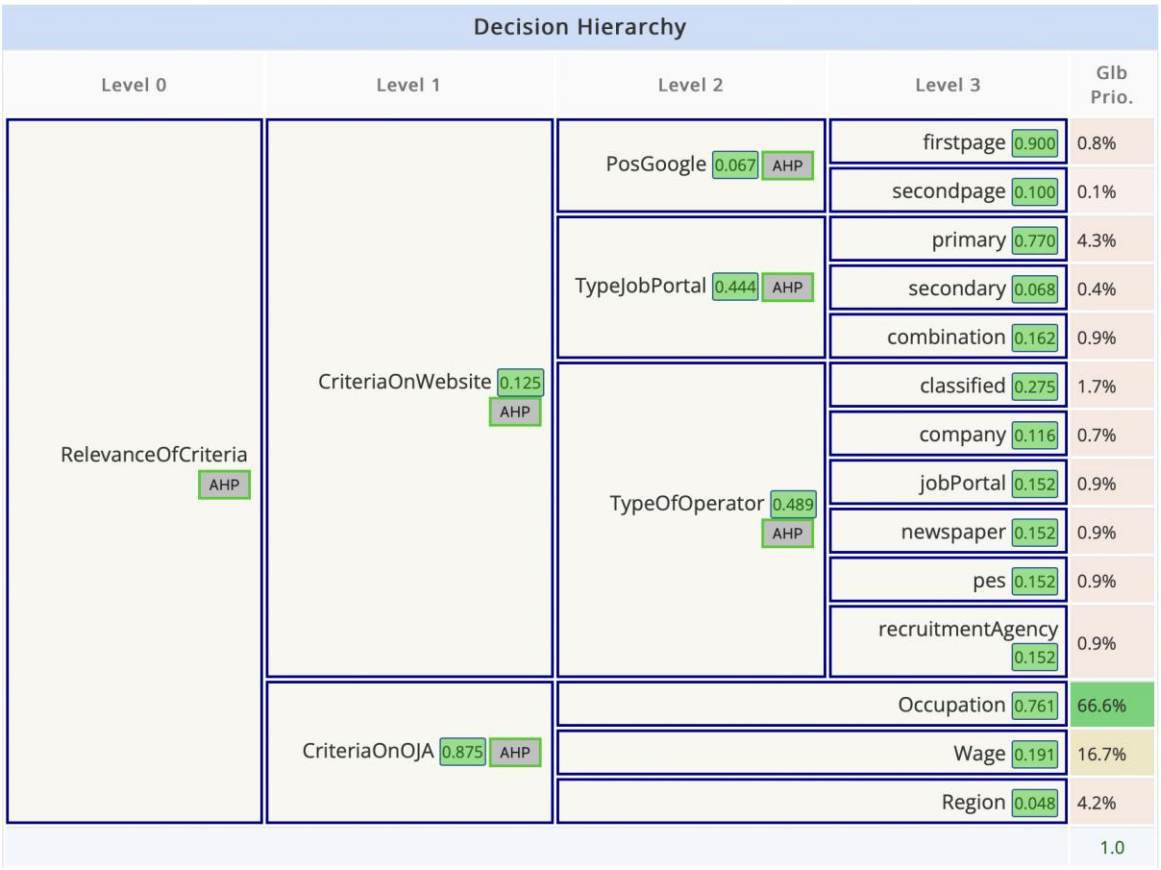
The AHP[3] is an effective technique for dealing with multi-criteria decision-making problems that allow decision-makers to set priorities to variables integrating the preferences of many stakeholders. By reducing complex decisions to a series of pairwise comparisons and then synthesising the results, the AHP helps to capture both subjective and objective aspects of a decision. The AHP is a very flexible and powerful tool because the scores attributed to variables' categorical values are obtained based on the pairwise relative evaluations of both the criteria and the options provided by the user. Moreover, the AHP can be considered as a tool that is able to translate the evaluations (both qualitative and quantitative) made by many decision-makers into a single score and the process can be repeated at higher levels of the structure and assigning a score to variables and to group of variables.

An example is shown in the Figure 2, where the hierarchy of criteria – group of variables, variables and variable values – are evaluated by a stakeholder. The procedure requires each stakeholder to pairwise compare all the element belonging to a level of the hierarchy, which are then translated in a score – such that the sum of the scores for each level is equal to 1. Moreover, the procedure checks the consistency of preferences for each level – namely that the preferences expressed by each etakeholder must satisfy the axiom of transitivity. The column "Level 3" shows the scores of the categorical values of each variable. Beside the attribution of numerical values to categorical variables, this method allows to have feedback on variables or group of variables importance from the point of view of each stakeholder.[4]

[3] Saaty, Thomas L. "What is the analytic hierarchy process?" Mathematical models for decision support. Springer, Berlin, Heidelberg, 1988. 109-121.

[4] The specific tool that is used in this case is: Goepel, K.D. (2018). Implementation of an Online Software Tool for the Analytic Hierarchy Process (AHP-OS). *International Journal of the Analytic Hierarchy Process*, Vol. 10 Issue 3 2018, pp 469-487, https://doi.org/10.13033/ijahp.v10i3.590

Figure 2 AHP model on the complete hierarch. Level 3 corresponds to variable values.



4.2. Ranking model

The list of web sources provided by ICEs will allow CRISP members to draw a ranking of web sources.

The goal of the ranking model is twofold:

1. To identify sources that are not eligible for being included in the final list of sources, due to low quality score (e.g., based on distribution of sources scores)
2. To prioritise sources that should be engaged as first through an ad-hoc agreement

4.3. Centralised assessment of web sources and ICEs validation

This part of the procedure aims at producing information concerning sources based on criteria of popularity, stability and coverage. The step produces evidence on some structural and technical characteristics of the source that are not available via inspecting the sources directly but can be discovered through an analysis of the data downloaded from them. Statistics on popularity, stability and coverage will be computed by CRISP. ICEs are then asked to evaluate the importance to each source, considering the evidence provided by CRISP and their personal knowledge of the labour market in their country. For instance, ICE may decide to flag a source as relevant even if it is not stable or does not rank high in popularity because of her personal judgement, based on evidence and knowledge of the country labour market. In such a case, ICEs are required to motivate their personal judgment. Moreover, ICEs are asked to report specific features of the websites that can be useful to identify features of the website such as implicit variables.

4.3.1. Popularity

Popularity aims at measuring how popular are individual sources by assessing the frequency of web searches that are referred to them. This is done through Google Trends provided by CRISP, which produces an index of search interest based on the volume of search normalized by major region (NUTS1) and time range. Each query produces three main detailed outputs: an index by basic regions (NUTS2), by week, and a list of the top 25 closest queries. The service allows one to compare up to 5 keywords and to extract the relative interest in these words. In this case, the index is additionally standardized in relation to one of the provided keywords.

The output produced by Google Trend is computed on a sample of all the queries that are conducted through the Google search engine. The representativeness of the sample is granted by the large number of queries administered by the provider. Individual searches data anonymized, categorized and aggregated. Following Google documentation, "[s]earch results are normalized to the time and location of a query by the following process:

1. Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, items with the most search volume would always be ranked highest.
2. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics.
3. Different regions that show the same search interest for a term don't always have the same total search volumes."

Another feature of the service is the availability of categories to which each keyword can be associated. This feature is important to address the ambiguity of keywords and and refine searches: for example, searches for "Indeed" would incorrectly count queries for the English adverb among the relevant searches of the job portal. Among the available categories, "960 – Job listings" is the one of interest in the case of LMI.

We have extended an existing prototype in Python that automatically generates the ranking of sources based on Google Trends. The current prototype uses Pytrends, an unofficial open-source API for Google Trends that allows a direct download of the data from Google Trends in an organized structure, i.e. a pandas data frame. Pytrends allows specifying all the parameters available through the GUI of Google Trend, namely localization (Country), time interval (daily, from 2004 to 36 hours before current time), categories (topic of searches) and type of search (Google Search, Google Images, Google News, Google Shopping and YouTube search).

Below we report an example of how to use Pytrends to get the trends of top OJAs providers in Italy. The name of the Source[5] has been used as keyword for searches and Google Trends measure, which is weekly, has been averaged to obtain the median value over the year.

The Figure below shows the ranking and the relative popularity of all the sources for Italy in 2019, comparing the results without a specified category parameter.

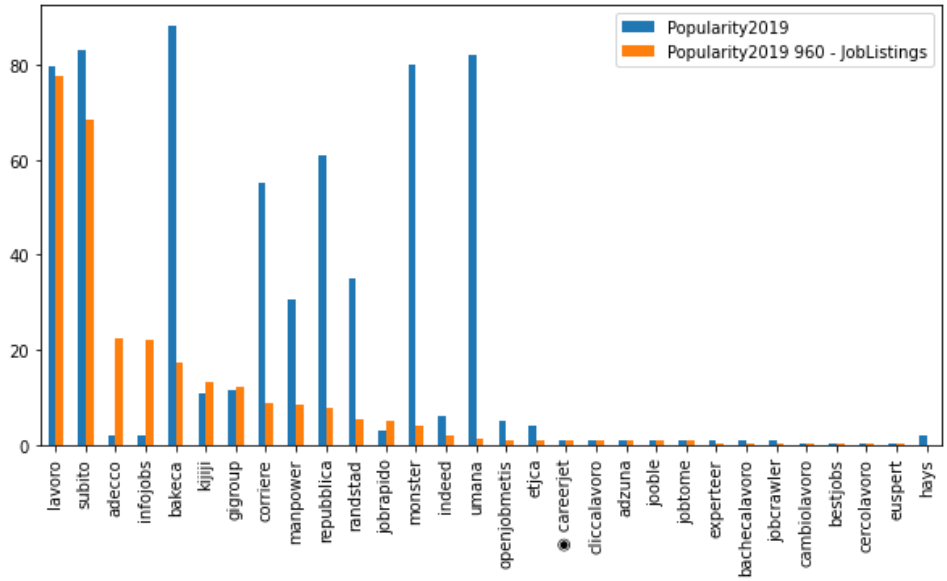
[5] Google guidelines stress the importance of punctuation to refine searches. Avoiding the use of punctuation allows including associated queries implicitly (for example searching for *Careerjet* automatically includes *Careerjet jobs in Milan* and all other associated queries; on the contrary searching for *"Careerjet"* would exclude all other associated queries). The drawback of this method is that "[n]o misspellings, spelling variations, synonyms, plural, or singular versions of your terms are included".

Evaluating the popularity of a web source

The proposed approach allows evaluating the popularity of a web source (or a term) in a specific time window.

To allow for a fair comparison among Google Trend Terms, either a specific source is be used as a benchmark, or a measure from the entire Google trend distribution (i.e. mean, median).

Figure 1 Popularity of keywords, with and without specification of category = 960 "Job listings". Median overall weeks of 2019, Italy. Benchmark ◉.



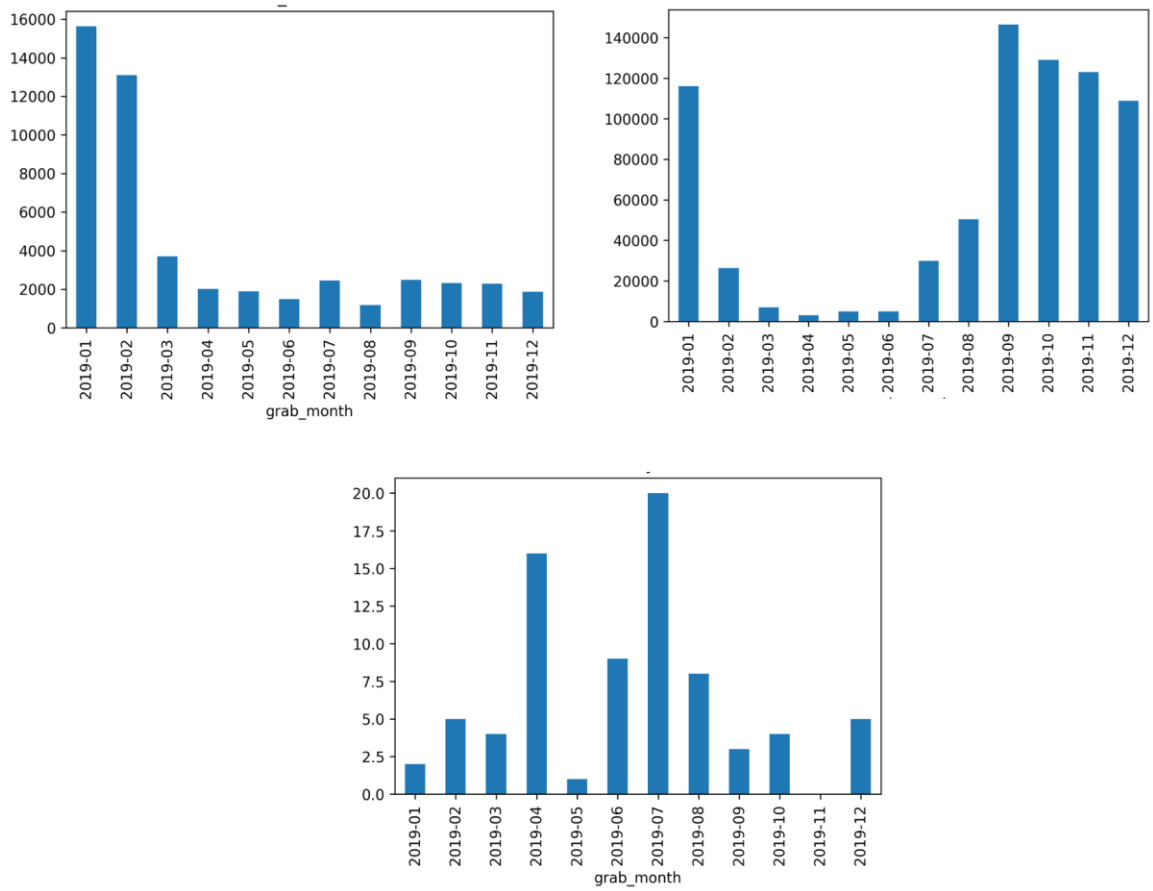
The two popularity rankings, as the Figure also suggests, differ in many respects, some websites, such as in the example *bakeka* and *subito*, offer services other than job listings, and once the category parameter (960 "Job Listings") is set, their popularity decreases. On the contrary, for other websites, namely *Adecco* and *infojobs,* their popularity increases once the parameter is set. The correlation between the two rankings is 0.66. In the case of equal scores, an additional criterion related to popularity may be needed.

4.3.2. Stability assessment

This step is aimed at keeping updated the list of sources over time, provided by CRISP. It can be performed after an initial block of data has been collected from the selected Source by evaluating each Country based on two criteria: Stability and Coverage. Therefore, this study can be performed on DPS countries only.

This activity makes use of the time series already available from the previous phase for 27+1 Countries to estimate source stability over time. The goal is to identify sources that provide stable information over time and sources that are unstable. Source stability is a common criterion used to evaluate the reliability and trustworthiness of a data source.

Figure 2 Example of three different Country's source time series.



The stability assessment needs to be evaluated following a set of criteria:

1. Percentage of missing OJAs in the time series;
2. Unavailability of OJAs for at least three months in the time series (even not consecutive)
3. The *diffrange* criterion is the ratio (max – min) / max of the number of OJA for each time point.[6]
4. The mean of the variation for each time point of the series;
5. The relative standard deviation of the time series;
6. The presence of outliers, computed by using the interquartile range (Q1 − 3 · IQR, Q3 + 3 · IQR)

For each criterion, a threshold will be identified and fine-tuned by following optimality criteria.

Below we provide some examples to explain the approach.

The Source in the upper right panel of Figure 1 is marked as unstable as it is above the threshold identified for the *diffrange* criterion (item no. 3) set to 90% (above 96% for this Source).

The upper left panel shows the case of a source that is marked as unstable due to (i) the presence of an outlier (two out of one allowed) and (ii) the value of the relative standard deviation 1.14 above the threshold, i.e., 1.

[6] Notice we plan to use the second min, and max value as this does not penalise sources having a one-time issue with missing a (or very low) number of OJAs.

Finally, the lower centred panel shows the case of a source that is classified as unstable because of the high mean of the variation, 67% against a threshold of 45%. In the above exercise setting the threshold is a crucial aspect. We rely on the stability criteria and threshold that are now in production in WIH.

Benefits of Country Stability

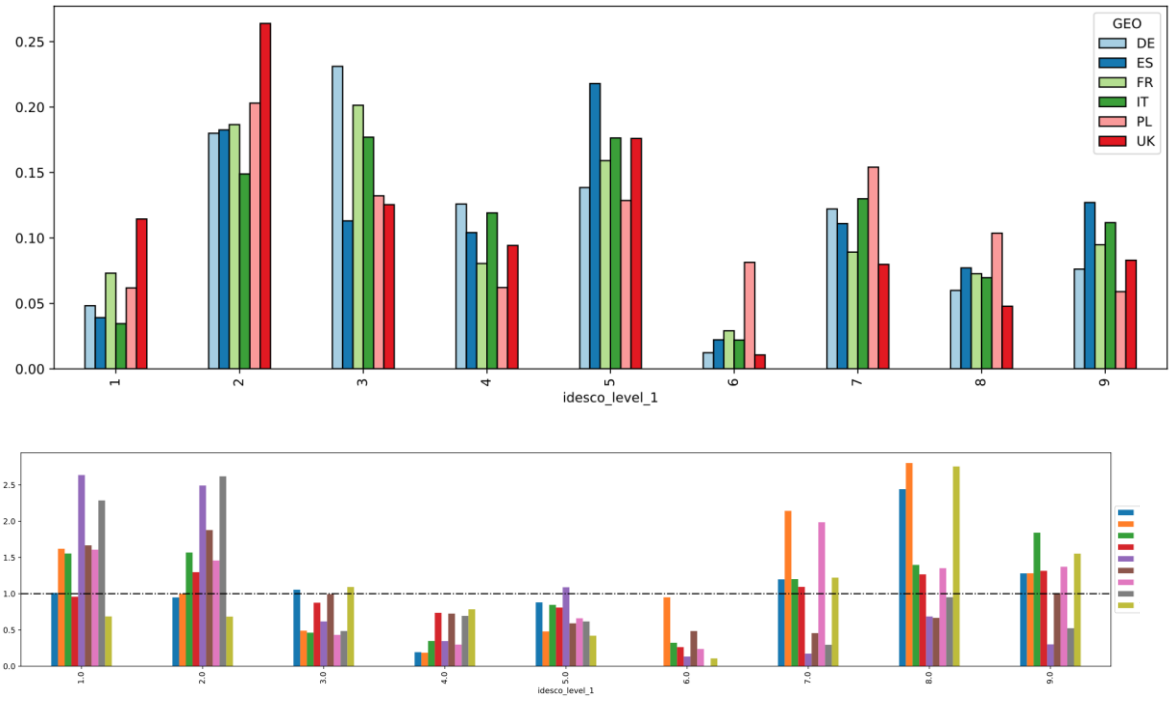
The stability assessment is crucial as it allows monitoring if – and to what extent – a source continues providing a strong enough time-series.

4.3.3. Coverage assessment

Coverage is defined as the ability of the Source to cover all the occupations of a country. From a "data quality" perspective, this can be seen as a proxy of the completeness of the Source.

Using information from data collected from the previous project, each Country's source coverage will be evaluated based on the distribution of OJAs classified over the first ISCO occupation digit. As a first step, the Eurostat LFS data might be used as a reference term of the distribution over the first ISCO occupation digit (see the Figure below).

Figure 3 (top) Eurostat LFS coverage over ISCO first digit for a selection of Countries. (bottom) Distribution of OJA over Sources for Germany. Source names have been omitted



The comparison between the source occupation distribution and Eurostat's will be made at the Country level to assess if – and to what extent – the list of sources selected can replicate the LFS benchmark.

The Figure above provides an example of this exercise by showing the distribution of OJAs for selected sources in Germany. The figure reports on the x-axis are the ISCO I digit, while on the y-axis, the distance

between the relative frequency of the Source and that of the benchmark. For ease of interpretation, the benchmark has been set to 1. Therefore, values above 1 imply that the Source over-represents a given occupation code, while values below 1 imply under-representation.

The example provided by the Figure shows that it is possible to distinguish:

1. Over-represented occupations. This refers to occupations that are over-represented concerning the LFS benchmark.
2. Under-represented occupations. This refers to occupations that are under-represented concerning the LFS benchmark.
3. LFS-aligned occupations. This refers to occupations whose distribution is aligned with the LFS benchmark.

Notice that the coverage assessment will be performed only on stable sources. Note also that coverage is a concept related to Source Representativeness, although there is a fundamental difference between the two. Coverage assesses whether the selected sources "cover" some well-defined domains (in the example above, the distribution of occupations). This is important for assessing the validity of the list of sources. Representativeness is a statistical concept that analyses whether the information on certain domains extracted from the selected sources is representative of the benchmark population. Regarding the LMI, the source representativeness has been addressed in a specific research report.

Benefits of Coverage Stability

The coverage stability would highlight those countries whose list of sources is not aligned with the LFS (i.e., it over/under represents some occupations).