ESSnet Trusted Smart Statistcs – Web Intelligence Network

Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 4
- Methodology and Quality

Deliverable 4.5: Quality Guidelines for acquiring and using web scraped data

Final version, 10 January 2025

Prepared by:

1. **WP leader: Alexander Kowarik (STAT, Austria, alexander.kowarik@statistik.gv.at)**
2. Magdalena Six (STAT, Austria)
3. Manveer Mangat (STAT, Austria)
4. Johannes Gusenbauer (STAT, Austria)

Funded by
the European Union

This deliverable was funded by the European Union.

Content

Funded by the European Union

Funded by the European Union

1. Introduction

In this deliverable we present the quality work done in WP4. The document is structured along the phases of the production process – from selecting the web data sources to ingesting the web data, to processing it and to producing output with it.

Figure 1 shows the different phases of statistical production with web data, lists included processes and depicts the end product of each phase, which is in turn the input for the next phase.

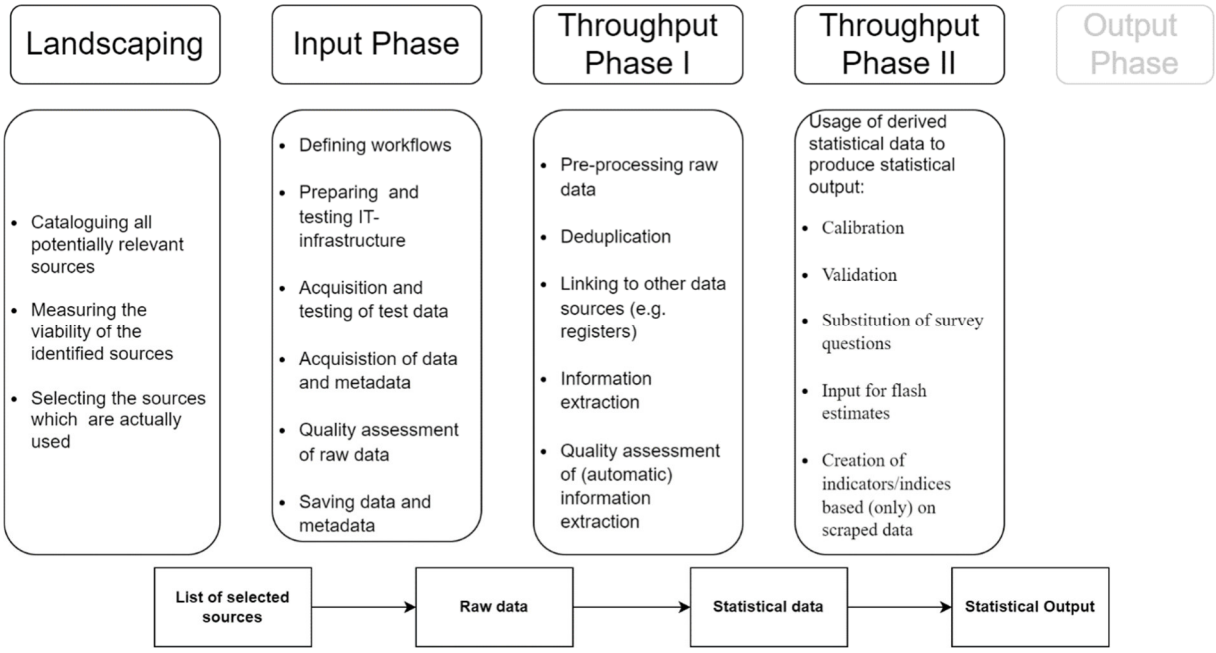


Figure 1 Quality-relevant processes along the production process

It was part of the tasks of WP4 to collect the quality aspects from the different use cases of WP2 and WP3 and to present them - potentially in generalized form.

We have to emphasize that the different phases of the production process are covered in varying detail, depending on the work in the different use cases in WP2 and WP3 as well as on the specific decisions on which aspects to focus in WP4. The members of the WP4 quality task made these decisions based on the observation of areas where quality input was lacking. This applies in particular to the detailed framework developed for landscaping (see Chapter 2) as well as the theoretical foundation for hierarchical classifications (see Chapter 6.1).

WP4 also contributed intensively to the quality assessment of the automatic information extraction models in WP2. This primarily concerns use case OJA. The results of these quality-related efforts are largely presented in Deliverable 4.8 (Quality Assessment for the statistical use of webscraped data) and Deliverable 4.6 (Methodology report on using webscraped data). General aspects about the annotation exercises can also be found in this document (see Chapter 6.2).

The document is structured as follows.

Chapter 2 introduces a comprehensive framework for mapping web data sources, presenting a general approach with examples from the Essnet WIN and other web data applications.

Chapter 3 addresses quality considerations for the web data acquisition process, followed by Chapter 4, which offers guidelines and indicators specific to various use cases.

Funded by the European Union

Chapter 5 discusses the general procedures for processing scraped web data, while Chapter 6 focuses on evaluating the quality of automatic information extraction models using annotations. Chapter 7 provides specific guidelines for processing ingested web data according to the different use cases.

Chapter 8 outlines quality standards for using derived statistical data in the production of output statistics, acknowledging that most use cases aim to generate statistical data, leading to a more concise treatment in the respective chapter.

Finally, Chapter 9 presents guidelines on establishing a centralized web data infrastructure, drawing insights from the "lessons learned" with the Web Intelligence Platform (WIP), which served as the central infrastructure in Essnet WIN.

This report underwent two rounds of review by Antoniade-Ciprian Alexandru-Caragea. The first review was conducted approximately one year prior to the final deadline on a partial draft, while the second review focused on the complete version shortly before the submission deadline. We extend our gratitude to the reviewer for his positive feedback and constructive comments, which significantly contributed to enhancing the quality of this document.

Funded by
the European Union

2. Landscaping of websites for webscraping with focus on selection models

2.1. Introduction

The experiences from the ongoing work in the ESSNet WIN show: No matter how well the process steps of data ingestion and data processing are done, the quality of the output very much depends on the quality of the source. As typical for Official Statistics, "quality of the source" is a multi-dimensional concept.

It can refer to several different aspects, e.g.:

- the stability of the access to the website,
- the availability of the most important information for the topic of interest,
- the trustworthiness of the website owner or the market share of the website,
- etc.

In this chapter, we therefore focus on the process steps of finding out which web sources are available as input for a specific topic of interest and how to – if necessary – select the ones which will lead to the highest quality of the statistical product.

Systematically identifying, evaluating, and selecting web sources that are most relevant and suitable for a specific statistical topic or interest is performed through landscaping, thus we first try to provide a definition of the term Landscaping, and subdivide Landscaping into three subprocesses "Catalogue", "Measure" and "Select". We analyse each subprocess and show that the complexity of each subprocess is very use-case dependent. In the last subchapter we focus on those cases where a selection model is needed. We first categorize groups of information upon which the selection model relies. We then present two selection models of varying complexity which were developed in WP2 and WP3.

The proposed actions might depend on national legalization. Whether the NSI is allowed or not to perform a certain action is not within the scope of this document.

2.2. Definition of Landscaping

Within a company or organisation, the term "landscaping" refers to cataloguing and measurement of all the data in the company or organisation[1].

Similarly, in the world of web-based data, landscaping could be understood as cataloguing and measurement of all web-based data sources relevant for the topic of interest.

It is worth noting that no general definition of landscaping in case of web-based data for Official Statistics has emerged yet. There seems to be a common understanding that "landscaping" refers to the process(es) before the actual ingestion of data from the websites starts. **Fehler! Verweisquelle konnte nicht gefunden werden.**, which depicts the data pipeline for Online Job Advertisements (OJAs) scraped at the European wide Web Intelligence Hub (WIH), illustrates this idea.

Informally speaking, "landscaping" can be interpreted as "getting an overview of the relevant sources". Once one knows about all relevant or all potentially relevant sources, one can gather

[1] See https://euler.net/landscape-analysis-data-assets-data-processes/: "What is landscape analysis? Simply put, it is the cataloguing and measurement of all the data in your company or organisation."

Funded by the European Union

information in a further step about these websites. Based on this information one can select the sources out of the potentially relevant sources, which are afterwards actually used for web-scraping.

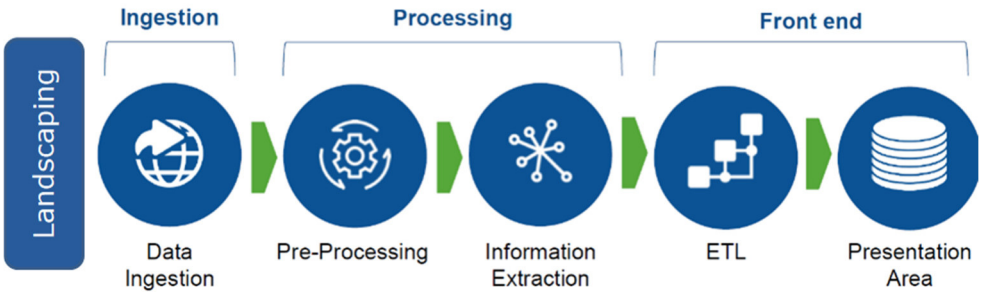


Figure 2 Data Pipeline for online job advertisement data, taken from the WIH for OJA data

We are not aware of a precise definition clarifying if the term "landscaping" refers only to the first step, namely the cataloguing of potential sources, or if it also comprises the measurement of the sources and based on this measurement, the selection of sources.

Following examples such as the data pipeline for online job vacancies (see **Fehler! Verweisquelle konnte nicht gefunden werden.**), where landscaping seems to include all processes before the data ingestion starts, we propose our own definition as follows:

Definition: **Landscaping** comprises all process steps necessary to **catalogue** all relevant sources for a specific topic of interest, to **measure** the quality and technical viability of the catalogued sources and to **select** the sources, which are actually used, based on the measured criteria.

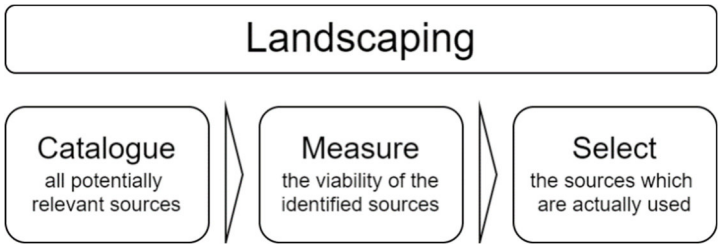


Figure 3 Landscaping and included sub-processes, according to our Definition

Whereas it seems natural that the sub-processes cataloguing, measurement and selection of sources build on each other and therefore must happen one after another (see **Fehler! Verweisquelle konnte nicht gefunden werden.**), this does not always have to be the case. In reality, this can be an iterative process. E.g., the term "relevant" in the subprocess "catalogue" implies that some form of measurement and/or pre-selection is already needed to decide if a source is relevant enough to be catalogued. Quite often, the use of a search engine such as Google or Bing implies some form of pre-selection, because only the first x-ranked search results are catalogued and afterwards examined in more detail. Another example for an iterative process is that you find out during the subprocess "Selection" that you need another selection criterion which has not been collected (measured) so far, so you have to go back to the subprocess "Measure". Further, the subprocesses are not always completely distinct.

Funded by the European Union

Therefore, **Fehler! Verweisquelle konnte nicht gefunden werden.** with overlapping subprocesses, which do not have to happen sequentially, one after the other, might draw a more realistic picture.

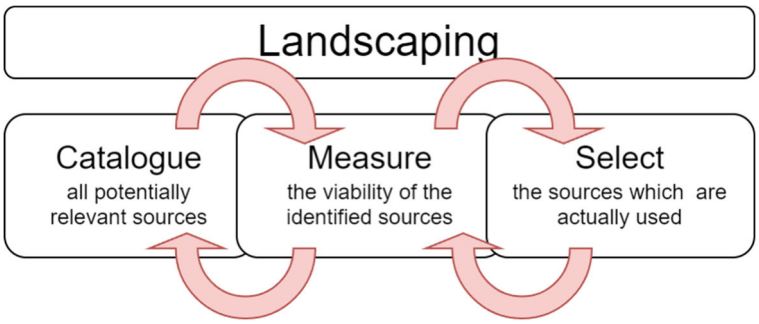


Figure 4 Overlapping, iterative subprocesses of landscaping

It is also worth mentioning that landscaping is not an isolated process, but is itself already a subprocess of the whole statistical production process of a certain web-based statistics. With this in mind, it is quite logical that other process steps outside of landscaping can also influence the landscaping process itself. For example, if certain quality problems are identified in the statistical output this could lead to the introduction of a new, or the adaption of, an existing selection criterion.

2.3. Varying complexity of subprocesses for varying topics of interest

Depending on the topic of interest, the difficulty of the landscaping exercise can vary enormously. Similarly, the effort needed for each of the subprocesses varies.

Some examples to illustrate this observation:

- For specific services or goods such as household energy or mobile phone tariffs, there exist price comparison portals in many countries. These portals give a good overview of all relevant prices on the market. If your topic of interest are prices in a market segment such as household energy or mobile phone tariffs, the subprocesses of cataloguing and selecting of sources simplify a lot. So the remaining challenge is only to have to measure the quality and technical viability of the price comparison portal.
- In other cases, such as online based enterprise characteristics[2] or retrieving all web-based information about the European drone industry [DA22], one of the main challenges is to find (catalogue) the relevant and technical viable homepages and to develop a selection mechanism which selects those which belong to the target population (e.g. decide if a catalogued website belongs indeed to a company active in the drone industry).
- In case of online job advertisements (OJA), real estate prices or other price statistics a great extent of websites – with varying market share, varying trustworthiness, varying legal and technical characteristics etc. – exist. The goal is not to scrape all of them, but to select and scrape the most relevant websites. A successful selection makes it necessary to have a catalogue of relevant websites, and the application of a selection model based on the measurement of the website.

These examples show: The landscaping process very much depends on the following questions:

[2] For more information about Online Based Enterprise Characteristics, see here: https://cros-legacy.ec.europa.eu/content/WPC_Enterprise_characteristics_en

Funded by the European Union

1. Does your topic of interest require to find **all websites** or only some **representatives** which fulfil certain criteria?
2. Do you have any information as input for landscaping, apart from the information on the website itself (**additional information**) or not? An example for additional information is a list of statistical units, such as enterprises from the SBR, whose websites you want to scrape.

Fehler! Verweisquelle konnte nicht gefunden werden. shows examples for each of the possible answers.

Table 1 Examples for situations with/without additional information and for all vs representative websites

Examples	**Find all websites**	**Find representatives**
Only information from website available	Case I: Identify the population of businesses active in a specific sector (e.g. green industry, drone industry)	Case II: Online Job Advertisements in case of centralised scraping by Eurostat[3]
Additional information (not from the website) available	Case III: Enhancing Business Registers *Additional information: Names of businesses from business register*	Case IV: Prices of clothes on online-shopping websites *Additional information: Companies with highest turnover from business register*

The actual form of the subprocess "Selection" very much depends on the question if you are looking for a certain "representative" subset of catalogued URLs and therefore on the question if you want to find all or only representative websites for a specific topic of interest:

In those cases where you want to find all websites belonging to a target population, after the cataloguing process you have a list of catalogued URLs, where potentially, some of the URLs belong to companies or other website-owners, which are not part of your target population. The aim of the Selection subprocess is to select those URLs, which belong to your target population – in case of additional information this might include some linking procedure with the statistical business register (SBR) or some other existing list of enterprises. This form of landscaping can be seen as an analogy to frame or sample (if a subset is selected) creation in a classical survey or administrative data based statistical production process. So, in statistical terminology you might think of getting rid of frame errors such as over-coverage and under-coverage. By mistake websites can be left out of the catalogue, they can be wrongly classified as being in scope and they can occur more than once in the catalogue (e.g. multiple URLs forwarding to the same website).

In cases where you look for representative websites, you have the advantage that the quantity of catalogued websites is not that large, and you can check manually if the catalogued URLs belong to the target population and are distinct websites. In this case the Selection subprocess is about

[3] In case of OJA there is in general also additional information available. For example, rankings or assessment analysis of job portals from recruiting agencies or other institutions like "deutschlandsbestejobportale.de", "cross-job-guide.com" or "online-recruiting.net" (see [KO16] p.11ff). To the best of our knowledge, this additional information was not used by Eurostat for the landscaping of OJA sources.

Funded by the European Union

selecting some specially viable websites from a catalogue of URLs (all part of your target population), based on a set of rules or based on a selection model.

Guidelines

- Define your target population.
- Define how you decide if a catalogued URL belongs to a unit in your target population

As the term representative in statistics is not well defined mathematically, we referred to representativity as a concept. The specific way this concept is actually operationalized, e.g. with a random or cut-off sample, must be determined based on the use-case.

2.4. Landscaping for all websites

2.4.1. Cataloguing all websites with no additional information (Case I)

This case is typical for situations, where you want to explore an enterprise website for industry analyses such as “green goods” or “drone industry”, but where you have no list of companies active in the respective field. It can also be the aim of the project to identify the population of enterprises active in the respective sector, based on their enterprise websites.

In this situation you have to search for **keywords** instead of enterprise names, which will generally lead to a higher variety of search results.

The authors of [DA22] list the following approaches to find all companies in a specific industry sector such as the drone industry:

1. Searching for websites that provide an overview of drone companies
2. Searching for websites of individual drone companies
3. Finding drone websites via pictures of drones
4. Finding drone websites via a company’s imprint page[4]

The authors list pros and cons for each approach, and they name some country specifics. In the end, they implemented the first and the second approach.

Guidelines

Since in this case the focus lies on finding *all* relevant company websites, it might pay off to test several approaches instead of simply searching via keywords for company websites.

- Be aware that different approaches might lead to results which would not have been detected by only adopting one approach.
- Make a list of possible search approaches and test each possible approach systematically.

Guidelines

According to [DA22], the following parameters can affect your search results when using search engines such as Google, Bing or DuckDuckGo

[4] The authors of [DA22] list also a fifth approach based on additional information such as a list of companies active in the drone industry from the chamber of commerce.

Funded by
the European Union

- the location (IP address) of the searching user,
- the user's previous search history (cookies),
- the country extension of the search engine used (e.g. using Google.nl vs. Google.ie), and
- the User-agent of the 'browser' program used.

The authors of [DA22] list the following ways to reduce these effects:

- using a VPN connection,
- using a browser that has no search history or searching the web via an anonymised (incognito) browser, and
- using a search engine that is specific to the country under study.

Guidelines

- Ideally, use a search engine which is not tracking your submitted search queries, alternatively search requests could be distributed to different search engines.
- Independent search engines should be favoured in comparison to commercial search engines. The European Open Web Index (OWI) is currently under development and will soon be the basis for independent and European search engines.

2.4.2. Cataloguing all websites with additional information (Case III)

In all the examples considered the goal was to gather information from company websites.

Thus, it seems natural that the most important source for relevant additional information is the statistical business register (SBR). But of course, "additional information" can have a different source: E.g., lists with companies active in a specific field connected to your topic of interest might also be available at the Chamber of Commerce or at other institutions.

If you have the company's names you can use them specifically in your online search, which will lead to more specific results.

2.4.3. Selection of websites belonging to a target population

In both cases – with or without additional information about the target population - the result of the cataloguing process is a list of URLs which can, but do not have to belong to the respective statistical unit / topic of interest. Due to the sheer number of URLs visiting the website "manually" to decide if the URL belongs to the target population is not an option. You therefore need an automated mechanism, which estimates for each URL in your catalogued URLs if it belongs to the target population.

The only information available for this automated mechanism is found on the websites itself. Therefore, this information needs to be scraped from the website (process "Data Ingestion") before you can proceed with the selection of websites. This complicates the distinction which subprocesses belong to the Landscaping process.

Figure 5 shows an illustration of involved subprocesses. Empty boxes in Figure 5 illustrate that the Figure does not show an exhaustive taxation of possible subprocesses of the respective processes "Data Ingestion" and "Processing".

Funded by
the European Union

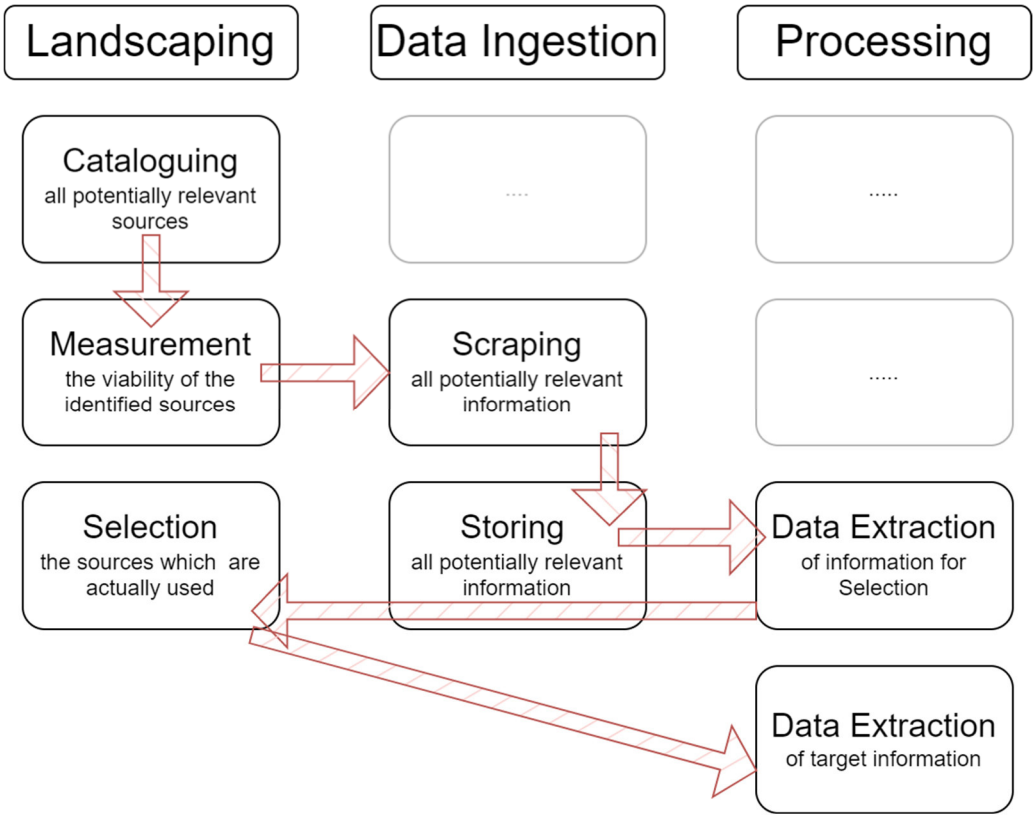


Figure 5 Involved subprocesses for Selection of websites belonging to a target population

2.4.3.1. Selection of websites with additional information

In this case you have a list of enterprises $e_i, i = 1, \dots, N$, e.g. from the Statistical Business Register (SBR), for which you want to find information on the web. On the other hand, you have a catalogue of retrieved websites after searching for the respective enterprises, which might or might not belong to the respective enterprises in the SBR.

Cataloguing for each e_i a list of potential websites and selecting the valid ones via linking procedures can be summarised as follows[5]:

1. Search in a search engine the name and/or address of $e_i$to retrieve a list of possible websites $u_{1(i)}, \dots, u_{n(i)}$ to match with.
2. Save besides the list of websites also additional information from the search engine result like the displayed text snippet or position of search result.
3. Further process the list of websites by removing duplicates and keeping country-coded folders for regionalized website
4. Use a crawler to scrape each of the websites.
5. During the scraping process the crawler scrapes the main page and up to N subpages on the same domain, while respecting the robots.txt exclusion protocol. Choosing N between 20 and 25 can be a good starting point.
6. The crawler can be instructed to look for certain subpages, like “imprint”, “contact”, “impressum”, which might contain the name and contact information of the website owner.

[5] More information and country specific reports can be found in ESSnet WIN WP3’s Interim technical report [ST22]

Funded by the European Union

In some countries also direct identifiers such as the VAT or commercial register numbers (CRN) of the website owner can be found on the website.

7. Process the scraped text from each website and try to extract the enterprise and or address from the scraped text as well as any direct identifiers. In this step the use of string similarity metrices can also be a viable choice.
8. Use the information gathered from Step 2. and Step 7. to either directly link an enterprise to a website or to build a model to predict such a link. The second approach depends on a pre-labelled data set.

At Statistics Austria the scraping procedure involves scraping up to 25 subpages for website. In addition, direct linkage between website and enterprise is viable due to national digital law. Given the ICT population a VAT or CRN number which can be linked to an enterprise was found in roughly 40 % of the scraped websites.

The outlined procedure clearly demonstrates that the "Measurement" subprocess includes scraping and storing URLs, as well as extracting information from the stored URLs to be used for the Selection subprocess. The Selection subprocess involves then a two-step linking procedure with the SBR.

2.4.3.2. Selection of websites without additional information

In this case selection means classifying websites which belong to your target population (e.g. companies active in the drone industry), with no additional information.

The authors of [DA22] list the following requirements which should a classification method should fulfil (for their use case "drone industry)

1. The classification method is able to discern between a drone and a non-drone website as accurately as possible,
2. The classification method has a high precision and a high recall, and
3. The classification method should be generalizable for other countries/languages.

The authors of [DA22] studied, applied and described three methods to discern between drone and non-drone websites. Their decision to apply these three methods out from a whole realm of binary classification options depended on their choice of Python and on the available data as well as on their experiences. The three tested methods (see [DA22] for more details) are

1. Word-based: Determine and count the occurrences of words expected to be specific for a drone company website.
2. PUlearning: Develop a semi-supervised machine learning approach using a set of labelled positive and unlabelled examples. This kind of approach is known as Positive-Unlabelled learning (see [PU21]) and has the advantage that only a labelled set of positive examples, e.g. drone websites, needs to be available.
3. Supervised ML: Develop a machine learning approach using a set of labelled positive and negative examples. This is an example of supervised machine learning. This approach requires a set of labelled positive and negative examples.

Funded by
the European Union

Guidelines

- Be aware of the sensitivity of the classification models on the structure, quality and amount of labelled data.
- If you need a sample for generating a labelled training data set, the sampling design should follow state-of-the-art research in this area.

2.5. Landscaping for representative websites

2.5.1. Cataloguing representatives with no additional information (Case II)

Cataloguing only a (small enough) number of representatives has the advantage that you can check manually if the first x search results fulfil your requirements. Since there is no need to find all respective websites, it is less important to invest in different search approaches.

For the use case of WP2, OJA, a further level of complexity arose because different partners with different languages worked together and agreed on a common, standardized protocol.

There needs to be some kind of assessment of the representativity of the selection.

Guidelines

If several parties, potentially with different national languages, all perform the subprocess "catalogue" and want to achieve a high degree of standardisation, the following points can help:

- Agree in advance on the kind of website you want to scrape the information from, be as specific as possible.
- Find out in advance about country-specific features of the respective markets and if these features might influence the results (e.g. in one country the job portals market might be more concentrated than in others, in some countries the most important job portal might be state-owned, for some countries the most important job portal might belong to a newspaper etc).
- The partners have to agree on the keywords, which are searched for, and translate them into the respective languages.
- Each partner should clean the web browser cache before starting the search
- All partners should use the same search engine to increase consistency (in case of OJA, all partners used google.com)

2.5.2. Cataloguing representatives with additional information (Case IV)

In this case the complexity of the landscaping process is small. You only have to find representative websites and the information based on which you select websites does not come from the website itself.

One possible method is to design a cut-off sample: a typical example is scraping online prices of specific consumer goods for the Consumer Price Index (CPI). The more difficult questions "Which goods are indeed sold often online?" and "Whose online prices are representative for all prices?" are out of scope.

In Austria, the chosen selection criteria, which companies are important in the specific sector, is the turnover from the business register. The focus of the cataloguing process is searching for the e-

Funded by
the European Union

commerce websites of the companies with the highest turnover in the respective sectors. Since the e-commerce websites of companies with the highest turnover in the sector are expected to be easily found, this is done quickly.

2.6. Selection of websites as Multi-Criteria Decision Making problem

When it is necessary to select representatives from all catalogued websites, it is of utter importance that this selection process does not happen arbitrarily. Especially, when the comparability between different countries has to be guaranteed, the need for a standard tool for the assessment of websites becomes evident.

A generic answer to the question "Which websites should be selected?" would probably involve answers such as: "The most important ones", "The ones with the highest quality", "The most representative ones" or all of the beforementioned. But quantifying importance or quality can be rather tricky.

Table 2 illustrates the problem: Even with a small number of job portals and a small number of characteristics (criteria) to be considered, it is not clear, which job portal in this hypothetical example is the "best one". It becomes obvious that the ranking of job portals depends on the importance of the considered criteria.

	Popularity	Reliability of Owner/Operator	Number of needed variables available in structured form	Original Ads?	…
Job portal 1	High	High	0	Yes	
Job portal 2	Medium	Low	4	Yes	
Job portal 3	High	Low	2	No	

Table 2 Example of several websites of job portals with different characteristics

Displaying the selection problem in this form shows that it is actually a M**ulti-Criteria Decision Making (MCDM) problem**. MSCM is well known as subdiscipline of Operations Research. The field of MSCM offers manifold tools and methods for determining the best alternative by considering more than one criterion in the selection process.

MCDM problems are characterised by three ingredients:

1) the **alternatives** which should be ranked (the websites)
2) the **criteria** based on which the alternatives should be ranked (the characteristics of the job portals)
3) the **model,** which determines - based on the criteria and the values of each alternative - the ranking of the alternatives (the selection model)

We will present two selection models used for different use cases in the ESSnet WIN, both of which fit well in the framework of MCDM models.

Before that, we introduce a useful classification of criteria to be used in the models.

Funded by
the European Union

We differentiate between three groups of criteria:

- **Information from the website itself (Inf1)**
  This includes all sort of technical and content-wise information that can be found on the website itself. It does not include historic information from previous scraping rounds.

- **Information about the website (Inf2)**
  This includes meta-information, which is not available on the website itself. This can be information about the market share of the website, or the popularity of the offered services/products. Also, information about the trustworthiness of the website owner (e.g. if the website is very trustworthy due to its public ownership like public job portals) or if an NSI and the website owner have a long-term agreement about the data access fall into this category.

- **Information from test runs or from experiences about scraping a website in the past (Inf3)**
  If an NSI has already scraped in the past and is re-evaluating the selected websites to scrape, it can build on the experiences from previous scraping rounds. Thereby, especially information about the stability of the access to a website (e.g. how often did the scraping processes fail in the past?) and information about the content-wise stability of the scraped information play an important role.
  Extensive testing of scraping of websites can partially replace non-existing experience from past scraping rounds.

The inclusion of all three categories of information might be costly but leads to the most trustworthy selection of websites suitable for scraping.

In the ESSnet WIN, the selection of representative websites played a role for several use cases in WP2 and WP3. A specific selection model was developed for the use case Online Job Advertisements in WP2, whereas the members of WP3 developed a more generic and common selection model for several use cases of WP3.

2.6.1. Selection model for OJA

The use case Online Job Advertisements (OJA) in WP2 differs from the other use cases because of the central role Eurostat plays in it: Eurostat provides the technical infrastructure (Web Intelligence Hub, WIH) for centrally scraping, processing and storing the OJA data for all ESS countries. Thereby, Eurostat is not only responsible for centrally scraping online job portals, also the selection of websites is in the hands of Eurostat[6].

The applied selection model consists of two building blocks: **a quantitative assessment of adherence of each website to the desired characteristics** and a **qualitative assessment of the sources' relevance** in OJA markets [TR22].

The first building block based on the desired characteristics concerns the website itself (Inf1) as well as information about the website (and the website owner) (Inf2), whereas the second building block considers the sources' relevance relying only on information from categories Inf2 and Inf3.

[6] Eurostat works partly with private subcontractors, the process steps of landscaping and selection of websites is mainly done by subcontractors, who in turn cooperate with international country experts to have access to country-specific knowledge.

Funded by
the European Union

The first building block involved variables such as:

- the type of the job-portal (primary job portal, secondary job portal or mixed),
- the type of the operator (classified ads portal, company websites, national newspaper, recruitment agency, …),
- the OJA volume displayed on the website,
- the sectoral scope (one or more),
- the displayed form (structural field, text or mixed) for variables such as "Type of Occupation", Type of Contract", "Working Time" etc,
- further variables from the website

These categorical variables have to be transformed into numerical values and combined to allow for an overall quantitative assessment of the website. This process must satisfy two criteria (see [Tr22]): First, numerical values are assigned following the relative importance that each value bears with respect to the other. Second, the preferences of all the involved stakeholders can be included.

In [Tr22], the authors explain the usage of **an Analytic Hierarchy Process (AHP)** for the creation of an **AHP-score** for the quantitative assessment of a website as follows: "The AHP 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."

The resulting **AHP-score** assigns **websites with the highest adherence** to the desired characteristics the **lowest score**.

The second building block is based on three dimensions: the **popularity** of the website, its **stability** and the **coverage of the scraped information** for each website.

Popularity was measured by the websites' relative interest as produced by Google Trends (see [Tr22]). It clearly falls into the category of information about the website (Inf2).

The dimensions "**stability**" and "**coverage**" fall into the category "information from previous scraping rounds" (Inf3). Of course, these dimensions can only be considered for countries who update already existing lists of scraped websites.

Stability involved several criteria, affecting the **stability of the access to the website** as well as the **stability of the time series** based on the scraped data.

Coverage refers to the question if the scraped OJAs **cover all classes belonging to a classification of interest** such as ISCO or NUTS in **a similar way as comparable known data**. More specifically, the distribution of scraped OJAs with respect to ISCO first digit is calculated for each source. Then, the

Funded by
the European Union

calculated distribution is compared to the distribution known from the Labour Force Survey[7] (see [CO21]).

The website with the most similar distribution of the respective variable gets the lowest value. This holds also for stability and popularity: the more stable and the more popular, the lower the value.

Combining the three dimensions "popularity", "stability" and "coverage" into one rank leads to the so-called **ICE-rank**.

Combining the two building blocks - the AHP score and the ICEs' ranks – leads then to **a final score**. This step involves the mapping of the AHP score and ICEs' ranks to the quartile of belonging in the respective distribution of values. Then five groups were defined to consider the joint distribution of AHP score and ICEs' ranks and mapped according to the scheme provided in Table 3.

Table 3 Score definition, taken from [CO21]

Score	Definition	Cases (AHP score quartile, ICE rank quartile)
1	Sources with position in Q1 of ICE rank and Q1 of AHP score.	(Q1,Q1)
2	Sources with position in Q1 or Q2 of ICE rank and Q1 or Q2 of AHP score. Exclude the case (Q1,Q1).	(Q1,Q2),(Q2,Q1) and (Q2,Q2)
3	Sources with position in Q2 or Q3 of ICE rank and Q2 or Q3 of AHP score. Exclude the case (Q2,Q2).	(Q2,Q3),(Q3,Q2) and (Q3,Q3)
4	Sources with position in Q3 or Q4 of ICE rank and Q3 or Q4 of AHP score. Exclude the case (Q3,Q3).	(Q3,Q4),(Q4,Q3) and (Q4,Q4)
5	Sources with distance between position in ICE rank distribution and AHP score distribution larger than 1 quartile.	All the others, e.g. (Q1,Q4), (Q3,Q1)

The **final decision which websites to scrape** is then based solely on the calculated final score: the smaller the score, the higher the probability that the website is indeed scraped. It is not feasible to establish a specificthreshold for the final score, below which all websites are scraped. This threshold is country-specific and varies also for countries who participate for the first time; for most countries' websites with a score smaller or equal to 2 or 3 are scraped.

2.6.2. Selection model for WP3

The partners of Work Package 3 "New use cases" developed a common checklist for assessing the information from web data sources, designed specifically for the purpose of systematically selecting websites in a coordinated way.
Contrary to the selection model for OJA, the selection model developed by the WP3 partners involved different use cases. Thus, it could not refer to use-case specific variables on the websites to be scraped, but had to be more generic.

A further difference to the OJA selection model is the focus on the information from the websites themselves (Inf1). Since it was the first scraping round for all use cases, no experience from previous

[7] The provided documents do not give further explanation why the LFS was chosen for comparability purposes. The LFS includes information about the distribution of the stock of non-vacant jobs and not the distribution of the vacant jobs. A distribution of the vacant jobs from the Job Vacancy Survey (JVS) would probably be better suited for comparability purposes.

scraping rounds (Inf3) could be taken account. No reason was given in the technical reports why no meta-information about the websites (Inf2) was incorporated in the selection model.

The developed checklist includes a list of **necessary characteristics** of the website and its content and a list of **optional characteristics** of the website and its content. Whereas the non-fulfilment of the necessary conditions leads automatically to the exclusion of the respective website from the list of websites potentially suitable for scraping, the existence of the optional beneficial characteristics is the basis for the calculation of a score according to which the potentially suitable websites are ranked.
The list of necessary characteristics / conditions is subdivided into:

- **Stop criteria** (if one of the **stop criteria is fulfilled**, the **website is immediately rejected**). Examples for these mostly technical stop criteria are: whether a website uses captcha or whether a website blocks robots.

- **Minimal criteria** (if **not all of the minimal criteria are fulfilled, the website is rejected**) Examples for these minimal criteria can be of technical nature (e.g. whether a web source offers a content filtering functionality relevant for the use case, whether the web source has new content published within the last month), as well as of content-wise nature (e.g. whether the number of ads on the web source is greater than a certain number)

- **Mandatory variables** (if not all listed mandatory variables can be found and scraped on the website, the website is rejected). Mandatory variables can be both content-wise and technical. Content-wise mandatory variables are those that are required to make the website informative and relevant to the user. For example, the price of a property must be included in all real estate ads, and the name of a product must be included in all product ads. Technical mandatory variables are those that are required for the scraping process, for example, all ads must have a unique ID and a URL.

The list of optional characteristics / conditions, whose fulfilment on the website is checked can be further described as follows:

- Existence of **additional criteria**, which are not mandatory but which **increase the viability** of the website
- Existence of **optional variables** to be found on the website, which provide **additional useful information**

One way **to derive a score** is to simply add up all the fulfilled optional beneficial characteristics. If one of the necessary characteristics is not fulfilled, the score is automatically set to zero.

Whereas the generic description of the mandatory and optional variables allows for a lot of freedom and the application to many different use cases, the model to actually combine all the collected information and transform it into a score or a rank is rather simple. Introducing weights for different variables would allow the model to differentiate between more important and less important criteria.

All websites in question are then **ranked according to the calculated score**.
Depending on the use case either the x highest ranked websites are scraped, or all websites above a fixed threshold are scraped.

Funded by
the European Union

2.7. Conclusion

Landscaping web-based sources is a relatively new part in the statistical production process. In this chapter, we aim to systematize this concept by presenting and describing the three subprocesses Catalogue, Measure and Select, which collectively constitute the Landscaping process. Additionally, we differentiate cases based on whether the goal is to scrape all websites or only representative ones, and whether there is supplementary information beyond the website itself. This differentiation enables us to group use cases with similar subprocesses, facilitating the description of relatively homogeneous subprocesses for the use cases within these groups.
For selecting representative sources, we show that the field of Multi-Criteria Decision Making (MCDM) provides a broad and flexible range of selection models. These models can effectively address the challenge of choosing specific sources from a large pool based on a diverse set of decision criteria.

Funded by
the European Union

3. Input Phase – General aspects

The landscaping phase as presented in the previous chapter ends with a list of selected websites to ingest web data from. The ingestion process typically takes the form of web scraping, an API connection or data delivery by the web data owner.

3.1. General aspects about the inclusion of web data in the statistical production

The following questions about the statistical production should be considered:

- What are the exact usages of the web data? Is the web data used as (sole) input to produce statistical output or to complement or check other data sources?
- What is the intended timeframe?
- Is it planned to generate Experimental Statistics, or should the web data be used directly for the production of Official Statistics?
- Is it planned to produce statistical output in parallel, both in the traditional way and with the inclusion of the new web data, for testing and observation purposes?
- What are the implications of using web data? What trade-offs are to be made? For example, we may obtain more granular indicators but with an unknown coverage bias.
- Should the website owners be informed about (repeated) scraping by the NSI? Would a different access to the web data such as an API connection or direct data delivery be helpful and how could it be negotiated?

The following questions about risks beside the statistical production should be considered:

- Could the outputs involved become vulnerable, especially due to blocked websites?
- Could there be any consequences to the reputation and the trustworthiness of the statistical office?
- Which legal aspects have to be considered?
- What risk mitigation strategies can the statistical office develop?

Data access has to be taken into account: Long-term access has to be guaranteed.

3.2. Acquisition and testing of test data

The acquired test data has to be tested thoroughly.

During this stage it should be clarified:

- which (main) processes - technical and statistical - are necessary to use the new data source,
- whether the skills necessary to process the data are available in the statistical office,
- whether the available tools of the statistical office can adequately deal with the data. Particular attention should be given to the IT-issues of storage and computing capacities.

The forensic investigation of the test data involves:

Funded by
the European Union

- all the known steps involved in data cleaning,
- the production of aggregate statistics and the production of outputs,
- linking of the test data with existing data.
- checking if specific information, e.g., territorial unit or industrial sector, can be extracted from the website
- monitoring if the information on the website is up to date and if changes over time can be identified (in longer time series).

3.3. General guidelines for acquiring web data

- Keep the data acquisition and recording tools ("scrapers") updated. Web-scraping, text processing and machine learning tools have to be agile to follow the necessary changes of the data source. For example, if the website (e.g. a job vacancy portal) changes its structure, a person at the NSI responsible for web-scraping has to change the web-scraper to record the appropriate data. In other words, to scrape the data in a long time series, we need to monitor changes on the website and quickly modify web-scrapers.
- Ensure that each data set will have a corresponding metadata set. Use a unified format for data and metadata storage.
- When collecting the data, ensure that the needed information to derive reliable attributes, that can be used to link to other data (e.g., geolocation, NACE, etc.), are present in the respective web source.
- If possible, allow to access the raw data with a unified interface, i.e. the same name of fields for a specific dimension, e.g. company-id, NACE.
- If there are any methodological differences in the interpretation of the same dimension, e.g. job vacancy vs. job offer, please save them in the metadata.
- Ensure that all data is stored in a secure way and try to create different groups of users, e.g. external users vs. internal users to allow limited access to the data.
- Try to estimate the target population size, if possible, and use metadata to store this information.
- Use similar classifications, if possible, or at least create the transition key to encode/decode the list of possible values from one data source to another, i.e. level of education, recode lower secondary and upper secondary to secondary.
- Store the data in machine readable format, which can be processed directly by a computer.
- If possible, allow to access raw data in standard formats like JSON or CSV, to be easily loaded into most common data science environments.
- Since the web as data source is very volatile, time stamps need to be saved with collected data. Replication and possibility of reproducing each data generation after the raw data is crucial as in statistical production in general.
- The scraping process should be set up in a way that automatic checks are included throughout the whole process. So if the website has a changed structure, is not reachable at a certain point in time or has other problems, warnings should be logged in the system. The responsibility of maintaining and updating scrapers should be clearly assigned.

3.4. ESS-web-scraping policy

For web scraping, follow the document "ESS web-scraping policy" prepared by ESSnet Big Data WPC [CON2019].

This document provides the following Principles and Practices:

Funded by
the European Union

Principles

Web scraping will be performed in adherence with the principles of the European Statistics Code of Practice, and in compliance with intellectual property legislation (national copyright laws and the Database directive) (6).

The members of the ESS should use web scraped data solely for statistical purposes as laid down in regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics and the applicable national statistical legislation.

The principles guiding web scraping activities should be to maximise the benefits while minimising any burden, risks or potential impacts arising from these activities.

To this end, members of the ESS should:

- seek to minimise burden on the website owners;
- identify themselves to the scraped website;
- protect all personal data according to the GDPR;
- abide by all applicable EU and national legislation;
- respect the website scraping policies, being in agreement with the statistical principles laid down in regulation 223/2009 on European statistics;
- scrape for the purpose of creating new statistical information from the data;
- ensure transparency concerning the tools, and the methods and processes that are used for web scraping as well as the methods used for producing the relevant statistics;
- ensure that statistical information is produced in an unbiased manner, e.g when preparing training data sets for machine learning.

Practices (Implementation guidelines)

When web scraping, members of the ESS should:

- Respect the robots.txt exclusion protocol and only follow links to the extent necessary for maintaining the quality of statistics;
- Respect the wishes of website owners as set in terms and conditions insofar as practical to check those terms, and scraping is not essential to maintaining the quality of statistics as implied by statistical law;
- Identify themselves in the user-agent string and provide contact channels. This could include a link to a web-page explaining the scraper's purpose and what data it collects, responsible team contact details, and information on how to opt-out and request that extracted data be deleted;
- Follow internet standard conventions for scraping, such as standards established by the W3C consortium[8];
- Be transparent about its web scraping activities, possibly by providing information on the associated website;
- Inform website owners if a considerable amount of data is extracted on a regular basis. This would not be the case if a website is scraped with a low frequency and is not scraped in full depth;
- Seek to minimize burden on web servers, by:
    - adding idle time between requests,

[8] See : https://www.w3.org/TR/

Funded by the European Union

    - scraping at a time of day during which the web server is not expected to be under heavy load,
    - optimise the scraping strategy for minimising the number of requests to a domain;
- Only scrape data within the scope of the statistical office's legal mandate, and do not re-use or distribute the raw data;
- Handle web scraped data securely;
- Avoid web scraping in using public APIs or other data provision options where available.

Funded by
the European Union

4. Input phase – Use case specific quality guidelines

4.1. Quality guidelines for OJA

4.1.1. Introduction

The centralised webscraping of OJA data (ESSnet WIN, WP2) on the Web Intelligence Platform (WIP) represents the most significant input for ESSnet WIN WP4, primarily due to its extended duration and substantial investments in personnel and infrastructure.

Many guidelines in this section are framed from the perspective of data users. Since Eurostat produces the OJA data centrally, NSIs focus on leveraging this data to generate valuable statistical outputs. However, the experience gained here is equally valuable for future scenarios, where NSIs either produce their own data or actively participate in its production within the WIP. This use case can serve as an exercise in communication with data users, laying the base for engaging with national data users at a later stage.

To the best of our knowledge, until the end of the project, the following OJA indicators will be published for several countries:

- Number of active OJA towards the Last Days of Quarters
- Index of active OJA towards the Last Days of Quarters
- Share of active OJA towards the Last Days of Quarters

The indicators are calculated by the following variables:

- Macro Sectors
- Economic Sections
- Occupation, level 1 and level 2
- Skill level 0 and level 1
- NUTS 2

There are more actors / roles than usual: On the one hand, the landscaping of sources, including the selection of sources as described in Chapter **Fehler! Verweisquelle konnte nicht gefunden werden.**, the data acquisition phase as well as the first part of the data processing phase happens on the WIP and fall within the area of responsibility of Eurostat. On the other hand, the producers of statistical output such as the indicators listed above work only with pre-processed statistical data in tabular form, downloaded from the WIP. The typical quality reporting as it is done via the EHQMR covers the whole production process. So far it is unclear how users should complete such a quality report when significant parts of the production process happen on the WIP, where the users have limited insight.

A second issue is that the quality guidelines as provided in the EHQMR are mostly too generic to actually capture (all) quality issues in case of OJA data.

In this chapter, we therefore focus on a very specific aspect of the whole production process – the **relevance** and **the stability of the selected sources** (jobportals). We do this from the **perspective of a user** of the pre-processed data from the WIP, who does not have any insight in Eurostat's landscaping decisions and advanced selection models. We believe that this is a very meaningful exercise – it shows if the producers' efforts in landscaping OJA sources fulfil indeed the users' expectations about the relevance and the stability of the sources.

Funded by
the European Union

We are aware that for the actual assessment of the stability and relevance of the sources, the results of throughput phase I (the pre-processed data on the WIP, downloadable for the users) are needed. Therefore, this chapter could also be listed under the Chapter Throughput Phase I, but since we explicitly want to measure the success of the source selection, we decided to leave it under the Chapter Input Phase.

The following quality guidelines were actually implemented and tested in a centralised way for a whole list of countries. A summary of the outcomes is presented in "Deliverable 4.8: Quality Assessment for the Statistical Use of Web Scraped Data". The automatically produced country reports including the guidelines can be found on the internal ESSnet WIN's Wiki[9].

4.1.2. Guidelines for relevance of selected sources

The landscaping and source selection process is done centrally by Eurostat. Although experts with country specific knowledge were consulted, it is a good idea to check if the included sources on the WIH are indeed the sources the domain experts at the NSIs considers as the most important one

Proposed guidelines:

- If your NSI scrapes OJA data itself, compare the included sources from your own scraping processes with the included sources on the WIP
- If your NSI does not scrape, consult the labour market experts in your NSI and ask them to name the x most important job portals in your country and compare this list with the sources on the WIP for your country

4.1.3. Guidelines for the stability of existence of the included sources

The general goal when working with OJA data is to capture dynamics in the labour market. The number of vacant positions advertised online via job portals has the potential to be a good indicator for dynamics in the labour market. Due to restrictions (scraping possibilities, quality of sources), it is impossible to scrape *every existing* job portal (source) per country. Already in the landscaping process, specific sources (websites) per country were selected by Eurostat. Unfortunately, the selected sources do not stay the same over the whole available time series, some sources fall away, other, new sources are additionally included in the list of scraped sources.

Creating a time series by simply adding up all unique OJAs over an instable number of sources can capture effects with respect to the included sources instead of capturing dynamics in the labour market. E.g. if a formerly included source falls away and the number of all scraped OJAs falls, one does not know if this decrease is due to the excluded source or due to a decrease of advertised open positions. If the number of existing sources is instable, more advanced methodological tools such as chaining need to be considered to construct a meaningful time series over the aggregated sources. As a first step, it is important to get an overview over the stability of the existence of the sources.

Proposed quality guidelines to measure the stability of the existence of the included sources

- Determine if it is always the same sources in the course of the time span considered

[9] https://webgate.ec.europa.eu/fpfis/wikis/display/WIN/Quality+assessment+of+OJA+sources

Funded by
the European Union

- Determine for several points in time (e.g. at the beginning of the time series, the middle and the end of the time series) the x (e.g. 5 or 10) most important sources w.r.t. to the volume of OJAs scraped of each of the sources. Are the thereby found important sources included in the list of scraped sources over the whole time series?

4.1.4. Guidelines for the stability of the popularity of the included sources

Even if the number of scraped sources stays stable over time and all of the most important sources are included in the whole time series, the following can happen: The popularity of one source increases, leading to more OJAs on this portal, uncorrelated with a general increase of vacant positions. It is for sure not possible to say for sure if an observed increase in the number of OJAs happens due to an increase in vacant positions or due to an increase in the popularity of an included source. But the following information can give you hints:

Proposed guidelines to measure changes in the popularity of the sources:

- Calculate the ranking of the most important sources w.r.t the OJA volume and observe this ranking over the course of time
- Determine the number of OJAs per source and check (e.g. via a plot of the individual time series) if the dynamics of the individual time series per source are similar

4.1.5. Guidelines for the stability of sources over different versions of data

When a new version of OJA data at the WIP is available, the data should not change for time intervals which were already covered by older versions of available data on the WIP.

Most important, this is true for the sources - It can be annoying if a source which was included in former years disappears for the present year. But it completely makes your analysis unusable if the sources in former years disappear for former years in a new version of data.

Further, the microdata for former years should stay the same for new versions. If the microdata changes, this sort of revision needs to be announced and the changes need to be explained.

Proposed quality guidelines to measure the stability of sources over different versions of data

- Load an old version of OJA data as well as the most recent version from the WIP. For the overlapping years, calculate for relevant sources the number of OJAs per year for the old data version and the most recent data version. Calculate the difference in absolute numbers as well as well as in relative numbers.
- Of course, you can do this for several versions of old data.

4.2. Quality guidelines based on WP3 use cases

4.2.1. Introduction

In this subchapter we present quality guidelines for the input phase as well as for the resulting raw data based on the experiences from the use cases in WP3. Contrary to WP2 OJA, where the selection of sources and the scraping happened both centrally via the WIP, the members of the use cases of WP3 were able to independently decide which data sources they wanted to select. Also the scraping

Funded by
the European Union

process itself took place in a decentralized manner at the premises of the NSIs and not at a centralized platform[10].

As WP4, we extracted relevant quality aspects and potential quality guidelines from the Technical reports from WP3 (see [ST22], [ST23]). We present them here in a generalised form and give examples from the use cases of WP3 to illustrate how the respective guidelines were implemented.

4.2.2. Stability of the access to the sources and breaks in time series

The continuous access to the selected web data sources in the chosen ingestion period is essential. While access to data from traditional sources was largely in the hands of national statistical institutes (NSIs), access to web data can be suddenly blocked if site operators prevent scraping. Similarly, technical difficulties can cause scrapers to fail (temporarily).

In order to understand the sources of potential time series breaks or missing values, it is necessary to know the exact data sources, the types of data access, the intended data ingestion periods as well as the actual data ingestion periods.

Proposed guidelines:

- Name the sources you have access to

- Name the kind of access you have to the before named sources (webscraping, API, contract with the website owner) and describe – if existing - technical challenges

  Example WP3 UC1 France: INSEE has contracted through a partnership the provision of data with the owner of the main portal on real estate advertisements in France. As a consequence, data acquisition does not pose any problem as the data transmission comes under the form of a regular transfer using a secured platform, involving no web scraping procedure of any kind.

  Example WP3 UC3 SWEDEN: A technical issue: The critical point to raise is that even though there are several options to get a relatively fast solution to start using web scraping as a source of data, there is always a risk of data loss due to changes online. These changes are of such nature as the webpages changing, thus making it complicated to keep a script going over time without some maintenance work. Another challenge has to do with the page itself and how it operates. Other languages than HTML sometimes need to be managed when gathering data online. This could be JavaScript scripts that make web scraping more intricate.

- Describe the period in which you had access to each source

- If you did not have permant access to all of the selected sources over the whole time period in question, describe the reasons of the lost access (technical problems on side of the NSI, technical changes on the side of the source, being blocked from the source, changes in the contract,…) and potential solutions.

[10] The members of the use cases of WP3 agreed on a common selection model as described in Chapter 2.6.2

Funded by
the European Union

Example WP3 UC3 Sweden: one of the companies we were collecting data from got bought and changed its name at the end of the year. This caused all the article numbers to change, making it almost impossible to maintain a continuous timeline

- Do you have a contact person at the source in case of problems?

  Example WP3 UC1 Finland: Data acquisition continues directly from the provider in 2024. Occasionally there have been some issues with the received data tables, for example observations have been missing. Fortunately, these issues were fixed by the provider. Contact with the data provider is set by a dedicated email and the responses are usually solved within a week.

4.2.3. Stability of the content per source

The continuous monitoring of the input data per source is an important task to ensure stable quality of outputs. By monitoring basic characteristics per source, erroneous sources can be removed, replaced or imputed early in the production process and are not causing unpredictable errors in later stages or in statistical outputs.

Fluctuations in the number of observations can lead to fluctuations in the derived indicator. This is especially true if one expects that the number of observations itself to be a good indicator for the phenomenon to describe.

Fluctuations in the observations can indicate that the popularity/market shares of the sources changes, which again can distort derived indicators about the phenomenon to measure.

Small number of observations in certain regions (or w.r.t to other domain variables) can also indicate coverage problems.

Proposed guidelines:

- Calculate the number of observations per source and per time interval in question

  Example WP3 UC1 Finland: There are no volatile fluctuations in the indicators nor numbers of observations between consecutive time periods

  Example WP3 UC5 Netherlands: Number of URLs per quarter in data file of DataProvider

  Example WP3 UC5 Finland: Number of URLs vary by new additions and expiration of ones that are not renewed. Data is administrative and highly regulated for it's use in directing internet traffic.

- If you want to calculate indicators w.r.t. domain variables (e.g. NUTS region, NACE, age groups, rental/selling, urban/rural…), calculate the average number of observations per source and per time interval in question for each category/domain.

  Example WP3 UC1 Germany: there are considerably more offerings in major cities like Berlin, Potsdam, and Cottbus, compared to rural areas. In more remote rural regions, especially those far from Berlin, real estate portals list far fewer properties, especially rentals. This may be due to limited availability or a preference for traditional, offline sales channels in these

regions. Such disparity suggests that areas with sparse listings might not yield reliable data for the project's scope.
Remarkably, the data shows a relatively small number of rental properties in Berlin. This could be due to the extremely high demand for apartments, leading to properties being rented out before they are even listed online; this is an aspect that requires further investigation.

Example WP3 UC1 France: Another challenge was the geographical coverage, which may be quite poor depending on the areas considered. It has had a significant impact on the derived indicators, especially in the stratification design for the computation of indices, and also the breakdowns of such indicators at a local level.

4.2.4. Missing values

In surveys, we know the concept of unit non response and item non response, both causing missing values. Also in case of ingested web data, missing values can occur. The concepts behind these missing values are similar to the ones in case of survey data, but differ slightly.

Missing units on source level

Missing values can result from the complete failure of a data source. This means that, potentially for a limited time, no data is available for one or more data sources. The failure of a data source is most comparable to the concept of unit non-response in the case of surveys. Reasons for the failure of the source may include blocked data access by the site operator or technical issues on the scraper's side. The guidelines addressing this type of missing data have already been addressed in Chapter 4.2.2

Missing items

It is possible that not all predefined variables of interest exist on all the selected web sources. We call missing variables on specific data sources structural missing items. On the other hand, a certain variable, e.g., the year of construction of a building, can be available on a portal in general, but missing for some apartments.

Even more complex, it happened for use cases in WP3 that variables of interest existed on a selected source, but did not continuously exist all the time.

Guidelines:

Do all the predefined variables of interest exist all the time on all the sources?
If not, describe which variables were missing on which source and the respective time interval.

Example WP3 UC4 : Regular web scraping of 2 portals was conducted in 2023: Expatisan.com and Numbeo.com, which contain data on the cost of living at the city level. During the period under review, the availability of cities varied according to the sources. Expatistan.com consistently provided a stable number of cities with a 10% increase over the period, whereas numbeo.com showed fluctuations in the number of cities, with increases and decreases ranging up to 50%.

- Are there missing values in your target variable(s)?
  If yes, how do you proceed?

Calculate the absolute and relative number of missings for each variable of interest per source and per time interval.

Example WP3 UC1 Germany: Another major challenge is missing data. It has been decided (for Hesse) not to concentrate on complete cases only as this would mean a large loss of offers and may neglect the risk of bias. Therefore, we keep offers with missing data even if the missing data appears for "mandatory variables" (e.g. price, size, number of rooms).

- Are there missing values in variables which contain information that defines the population? If yes, how do you proceed?
  Calculate the absolute and relative number

  Example WP3 UC2 Sweden: Missing information in the variable "year of construction" leads to problems of defining the population of newly constructed buildings.
  For 84% of all offers the variable „construction year" is completed

  Example WP3 UC2 Germany: For some ads with unknown type of offer from the advertisement itself, the type can be derived from the URL of the offer or from the price (since there should be no objects for sale with a price less than 10.000 euro). But still, for some offers of specific portals, the URL has not been saved as well as the price is missing, so this information cannot be derived.

  Example WP3 UC5 Netherlands: missing values in the identification variables of data set with URLs from DataProvider and in the Statistical Business Register. That hinders to achieve a high proportion of linked URLs

  Example WP3 UC5 Finland: owner identification variable can be national or foreign business or personal ID, not all registered domains can be linked due to lack of foreign IDs

- Are there missing values in variables you use for deduplication? If yes, how do you proceed?
  Calculate the absolute and relative number

  Example WP3 UC2 Germany: The crux of duplicate identification lies in having precise address details. Regrettably, in our dataset, complete addresses are available for only about 23 % of the listings. This limitation stems from various factors. For instance, addresses are often provided by real estate agents only upon request, and in the case of new constructions, addresses may not yet exist, especially in newly developed residential areas where house numbers and even street names are often being established during or even after construction.

Funded by
the European Union

5. Throughput phase I – General aspects

The end product of the Input Phase, as described in the previous chapter, is the raw data. The Throughput Phase I focuses on the process steps through which the ingested raw data is transformed into statistical data (see Figure 1).

We have decided to include the use case specific quality assessment of the scraped data in Chapter 4. Thus, we have already addressed quality aspects such as the stability of the input and missing values in the input phase. Some quality aspects such as Coverage or Linking are relevant in several phases, thus they are also mentioned and described several times.

We dedicated an extra subchapter to the most prominent process step in the ESSnet, the extraction of information from scraped web data, more specifically the extraction of specific classifications, see Chapter 6.

5.1. Linking

Identify information enabling linking
So as to link scraped information to statistical units or concepts, you first need to identify which statistical concepts can theoretically be extracted from the text. This might involve dates, (business) names, business-ids, addresses, but also statistical classifications such as territorial units or economic sectors. Focus if possible on linking information according to official statistical classifications, such as NUTS for regional territories or NACE for economic activities.

Be aware about uncertainties of derived linking information and document it
Be aware that the result of information extraction such as Natural Language Processing (NLP) models is subject to uncertainty. Especially if the target information is not available in structured form but has to be derived through models, the extracted information is only "the most probable" one.
It can be helpful for further processing to also store the probabilities of the derived information. In this way, uncertainties can be documented and considered throughout the whole production process.

5.2. Coverage

Establish the population of interest.
The definition and study of coverage errors require the definition of the target population that should be explicitly identified in terms of type, time and place.
Identifying the relevant unit and, subsequently, the population of interest is often not a trivial task and can have far reaching consequences. For instance, defining the target population as „all online job advertisements" or „all enterprises posting job advertisements online" or even „all enterprises that post job announcements on job portals" can potentially yield very different results with respect to the estimated number of online job advertisements. Similarly, when looking at estimates with respect to enterprise websites it makes a difference, if your target population contains „all enterprises in the business register" or „all enterprises with a homepage" or „all enterprise websites", since the relationship between enterprises and websites is an n:m relationship.

Describe the statistical variables
Define and describe briefly the main statistical variables that have been observed or derived.
Indicate discrepancies, if any, from variables which were previously collected in a different way (e.g. via surveys).

Funded by
the European Union

Representativeness - Try to estimate the population size and compare with traditional data. For example, when you are scraping enterprise characteristics, try to count the number of websites that are accessible and can be used for web-scraping. Compare this number with the data from your business register.

5.3. Comparability over time

Closely monitor the structure of the data.
Check each data generation on structural changes in comparison to the previous one.

Fit an appropriate statistical methodology for producing the output.
According to the Analyse Stage of data generating process by AAPOR (2015), apply a statistical method not sensitive to extreme data and define statistical tools for smoothing the break in the time series related to structural changes of the data source or coverage changes over time.

Check if the modification/update date can be extracted from the website.
When web-scraping specific information from the website (e.g., job vacancies), try to extract the data of publishing this information. If the website is not up to date, it is unlikely to detect enterprise activity in longer time series. Since scraped data can disappear or change over time, the risk of missing data is high: monitoring and maintaining scrapers up to date should be considered as a priority.

5.4. Measurement errors

Establish the target information.
The definition and study of measurement errors require the definition of the target variable of interest.

Research on measurement errors.
If possible, measurement errors should be evaluated (on a small sub-sample) with an appropriate method, e.g., manual reviewing or comparison with other data.

Track changes need to be observed.
If values are changed or imputed because of detected errors or implausibilities, these changes should be tracked.

Verify, if the webdata fits the definition from official statistics.
It should be noted that sometimes the same variables may have a different definition.

5.5. Model errors/ process errors

Estimating the quality of models is of great importance. The most prominent process step in the ESSnet WIN which involved models, is the extraction of structured information from unstructured text. We thus dedicated a separate chapter (see Chapter 6) to this topic.

Apply appropriate model selection and evaluation criteria.
Techniques like cross validation, out-of-sample tests, etc. should be applied wherever possible to assess the model quality and possible errors.

Compare multiple machine/statistical learning methods.

Funded by
the European Union

Since it is not always straightforward to choose the right tool for the job, different methods should be tested and evaluated.

Evaluate the bias of the training data set.
In supervised learning, an unbiased training data is very important to not estimate based on a biased model.

6. Throughput phase I - Classifications

6.1. Assessing the quality of hierarchical classification models for web data

In many applications of web data in official statistics production, classifying mostly text data into categories is a challenging part of the process. Moreover, most classifications encountered in official statistics have a hierarchical structure (e.g. NACE/ISIC, ISCO, ISCED, etc.). Traditionally, the national statistical institutes often conduct the assignment of these classifications manually, which requires significant resources and time and is just not possible in some cases, e.g., for huge amount of online job advertisements. Hence, NSIs have been working on building hierarchical classification models using mainly machine learning methods.

There are three main approaches to conducting a hierarchical classification 1) the flat classification approach, 2) the local classifier approach, and 3) the global classifier approach. The flat classification approach ignores the hierarchical structure and constructs a classifier that distinguishes between all the leaf nodes at the desired level of detail. While this method provides consistent predictions, its main disadvantage is that - depending on the regarded use case - it might have to differentiate between a significant number of classes.

The local classifier can be categorised into three subgroups 1) the local classifier per node, 2) the local classifier per level, and 3) the local classifier per parent node. While the local classifier per node constructs a Boolean classifier for each node in the tree, the local classifier per level trains one multiclass classifier for each level. However, both these classifiers can produce inconsistent predictions. In contrast, the local classifier per parent node approach only trains each parent node to distinguish between its child nodes, hence providing a consistent prediction by construction. However, an evident disadvantage of this method is that a misclassification at any level will inevitably be propagated downwards.

Lastly, the global classifier approach trains a single classification model on the entire hierarchical structure in a single run. While this method provides consistent predictions and does not require the computation of multiple classifiers, its main drawback is its complexity.

For a detailed description of these different hierarchical classification approaches, see the survey by [SIL11].

Whether a proposed model can be used to produce statistical output or at least assist the manual classification process will obviously depend on its quality. Hence, suitable measures must be used to examine the quality of hierarchical classification models.

Standard Evaluation Measures (Flat Metrics)

- Precision: Measures the proportion of true positive predictions in the total predicted positives.
- Recall: Measures the proportion of true positive predictions out of the total actual positives.
- Accuracy: Measures the proportion of true predictions (both positive and negative) in the total dataset.

Funded by
the European Union

The drawback of using standard measures such as precision, recall and accuracy to evaluate hierarchical machine learning models is their inability to account for the relationship or closeness between the true and predicted classes.

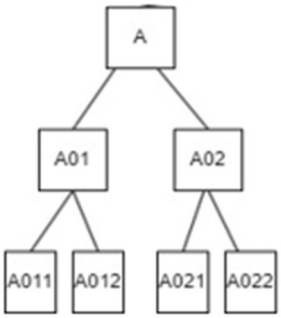


Figure 1: Example of a hierarchical structure of labels

Consider the following example: suppose that A022 gives the actual third-level class of an entity, and model 1 predicts A021 and model 2 predicts A012; then, with respect to the standard evaluation metrics, both models perform equally poorly as they are both incorrect. However, the prediction of model 1 coincides with the actual code of the entity up to the second level (A02), hence providing a "closer" and, therefore, better estimate of the true level 3 code A022 (see Figure 1). Therefore, model 1 should not be penalised to the same extent as Model 2, as the degree of alignment between the true value and the predicted value provided by Model 1 is greater than that of Model 2. Accordingly, Sun and Lima (2001) proposed adjusted versions of the standard evaluation metrics, which account for the class relationships in a hierarchical structure, to provide a more accurate evaluation method for hierarchical models.

To quantify the closeness between two classes in a hierarchical structure, we can compute the 1) category similarity or the 2) category distance.

- In order to compute the category similarity, the most common measure of choice is the cosine similarity. The cosine similarity of two classes is defined as the cosine angle of their respective feature vectors. Thereby, the cosine similarity is solely computed between classes on the same level, as the aim is to compute the extent of similarity between the predicted and true class on each level separately. The higher the cosine similarity, the higher the degree of alignment between the respective two classes.
- The category distance is defined by the shortest path between two classes in the hierarchical structure. Similarly, the shorter the distance between two classes, the higher the degree of alignment between the respective two classes.

Once the closeness between the predicted and actual class is quantified, it is incorporated into the computation of the standard evaluation metrics (see [SUN01]), which can then be used to assess hierarchical models.

For example, a misclassification that is 'closer' in the hierarchy (e.g., misclassifying a 'sparrow' as a 'robin' rather than a 'shark') is less penalized.

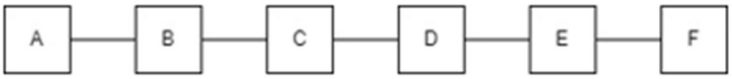
The adjusted standard evaluation measure based on category distance requires a hierarchical structure of the labels. Indeed, if there were no hierarchy, there would be only one level. Accordingly, the distance between the predicted and true class would either be zero (true and predicted class align) or one (true and predicted class do not align).

Funded by
the European Union

In contrast, the adjusted standard evaluation measure based on category similarity provides useful results even without a hierarchical structure of the labels. Indeed, as the category similarity is defined for each level of the hierarchy separately, it can also be computed for a flat classification structure which is a hierarchical structure with a single level. Therefore, it can be used in addition to the standard evaluation measures in flat classification problems.

An overview of the evaluation measures that can be used for the different hierarchical structures of labels is provided below:

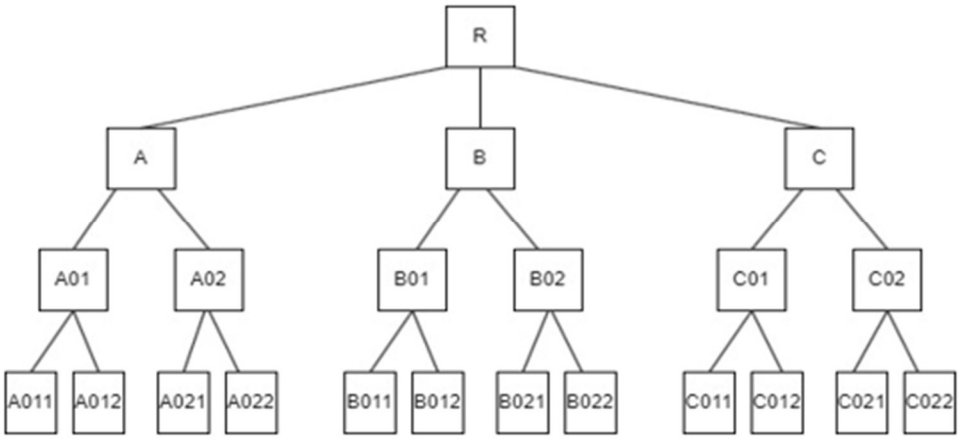
6.1.1. Flat classification



Guideline:

- For a flat classification the standard evaluation measures (precision, recall, accuracy) or adjusted standard evaluation measure based on category similarity can be used.

6.1.2. Hierarchical classification



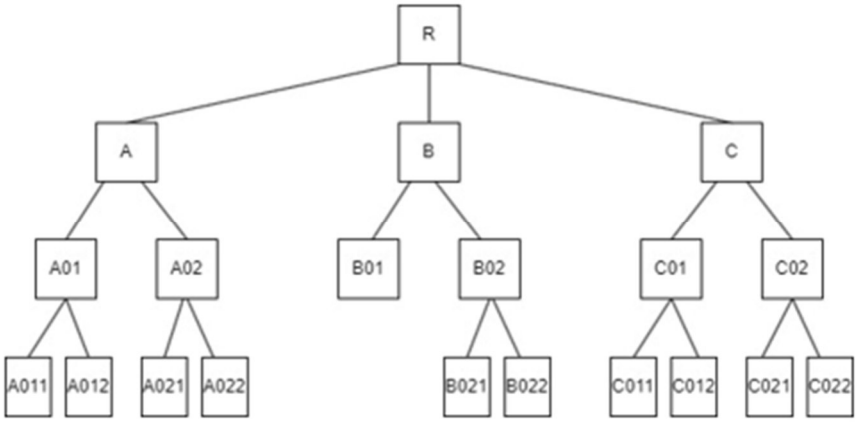
The Hierarchical structure of labels: Rooted tree, where all leaf nodes are on the same level

Guideline:

- Use one of these suggested evaluation measures:
  - Adjusted standard evaluation measure based on category similarity
  - Adjusted standard evaluation measure based on category distance
  - Hierarchy-based evaluation measures

Funded by
the European Union

The Hierarchical structure of labels: Rooted tree where all leaf nodes are on the same level



Guideline:

- Use one of these suggested evaluation measures:
  - Adjusted standard evaluation measure based on category similarity
  - Adjusted standard evaluation measure based on category distance
  - Hierarchy-based evaluation measures

6.1.2.1. Tailored Hierarchical Performance Measure

The measures considered so far, all give equal weights to each class when computing the quality of the model. However, for many applications the importance of correctly classifying units into the different classes might not be the same for all classes.

As an example, from classifying the economic activity of enterprises, it would be an option to give more weight to classes, to which large enterprises tend to be assigned. The reasoning for this is that a misclassified large enterprise would introduce a higher bias in a business statistic than a small one. Consider the following oversimplified illustration:

Suppose our population consists of ten enterprises and there are only 2 possible classes A and B. Five of these enterprises belong to class A and have a turnover of 1 Mio., 2 Mio., 3 Mio, 500.000, 4 Mio. euros, respectively, and the other five with a turnover of 1 Mio., 700,000, 1. Mio., 500.000, 1.5 Mio. euros, respectively, belong to class B. Then the total turnover for class A amounts to 10,5 Mio. and the total turnover for class B is given by 4,7 Mio. If our model were to wrongly classify the enterprise with a turnover of 4 Mio from class A to class B, then this would detrimentally distort the total turnover for both. In particular this would lead to the wrong conclusion that enterprises belonging to class B are responsible for the majority of the turnover, namely 8.7 Mio., while the enterprises in class A only account for 6.5 Mio. of the turnover in the population. If on the other hand our model were to wrongly classify the enterprise with a turnover of 500,000 Mio. from class A to class B, it would only introduce a minimal bias without violating the overall true distribution of the turnover among both classes.

Hence it is important to consider evaluation measures that weigh classes according to their importance for the statistical output.

Funded by
the European Union

The concrete proposal is to use a weighted variant of the evaluation measures:

$$\rho_l = \frac{1}{\sum_{i=1}^{n_l} s_{li}} \sum_{i=1}^{n_l} s_{li}\, p_{li}$$

$p_{li} \in \{Pr,\ Re,\ Ac\}$ evaluation measure at level $l \in \{1, \dots,\ 5\}$ for the class $i \in \{1, \dots, n_l\}$,

s_{li} number of employees -> weighted evaluation measure available for each level l ->

Take (weighted) average for an overall evaluation value

Guideline:

- Consider the development of a tailored hierarchical performance measure for your specific application.

6.1.2.2. Quality Guidelines

Applicability: The choice of metric and its adaptation should be aligned with the specific context and objectives of the hierarchical classification task.

Complexity: Adapting these metrics for hierarchical data can add complexity to the evaluation process.

Interpretability: Maintaining the interpretability of these metrics after adaptation is crucial for them to be useful in practical scenarios.

Understanding and applying these adapted metrics can significantly enhance the evaluation process in hierarchical classification systems, providing more nuanced and context-aware insights into model performance.

6.2. Assessing the quality of a specific classification by annotating a sample (OJA)

Annotation is the process of manually labelling or classifying data to use it later either as training data for machine learning algorithms or to validate results from such an algorithm. Given the volume of the data the annotation is usually done on a selected sub sample of the whle available data. Ideally, the annotation is done by subject matter experts with experience in the specific classification. To further improve the quality of the annotated data multiple experts should label the same data set, making it feasible to measure the certainity/uncertainity of human judgement to assign a specific class.

In the case that a labelled data set is already available, additional annotation might not be necessary as the traditional approach of splitting the availabe data into training, validation and test data might be sufficient to train the algorithm and measure its quality.

In the case of web data, annotation is almost exclusively the labelling of text according to a classification. However, in our application it can also including labelling sound, images, classifiying sentiments, etc. .

Funded by
the European Union

If the annotation is performed to get a labelled training data set, the quality of the annotation directly impacts the quality of the machine learning model. Accurate and consistent annotations will lead to improved models, while poor annotations can lead to inaccuracies and biases in the classified variables.

Guidelines:

Design the sample according to the needs
Ensure that the sampling plan is optimized for the specific annotation exercise. E.g., if the annotation exercise focuses on evaluating the accuracy of classification variables in web-scraped data. The sampling should aim to provide a representative and comprehensive sample to evaluate the quality.

Determine the necessary sample design
Determine the size and coverage of the sample. This should be based on the number of classes to be assessed in the classification, the volume of data, but it is limited by available resources (time and/or personnel) for annotation. The sample should be large enough to provide reliable results but manageable for the annotators.
Consider stratifying the sample to ensure representation across different categories or classes. For instance, in the case of OJA data, stratification could be based on industry sectors, job types, or geographical regions.

Define time horizon
Web data in some areas might change rapidly, so it's important to define if a specific period should be evaluated (cross-cutting evaluation) or if the sample should be continuously distributed over time. Another option would be to sample for more than one period a representative sample to be able to compare the quality for several points in time.

Establish annotation guidelines
If multiple persons function as annotators, it is important that they annotate in a similar way. To ensure that, guidelines for annotation should be developed in a clear, comprehensive way. This includes especially the process for handling ambiguous cases.

7. Throughput phase I – Use case specific guidelines

7.1. Quality guidelines based on WP3 use cases

In this subchapter we present quality guidelines for the throughput phase based on the experiences from the use cases in WP3.

7.1.1. Target population

As described in Chapter 5.2, the establishment of the target population is not trivial. In some cases, the target population is defined by a source which is independent from the web data source, such as the business register. In this case the task is more to find mechanisms to link e.g. a website of an enterprise to the enterprise in the business registers. In other cases, it is information from the web which defines the population. An example is the target population of new constructions, which you want to establish by scraping real estate portals. In the latter case missing values, measurement errors or models to extract the information if an object on a real estate portal is a new construction become more relevant.

Funded by
the European Union

For guidelines about missing values see Input Phase, Chapter 4.2.4

Proposed guidelines

- Define the target population of the statistical units for which information is sought.
  Is your target population defined by information from the source? Describe your rules or models deciding if an observation belongs to your population of interest or not

  Example WP3 UC2 Germany: Combinations of key words ("Baujahr", "Neubau", "Erstbezug") have to be set: In general, it has been decided to use a strong / narrow definition of "newly constructed building". This means that the year may be missing or must refer to the year of reference but an offer must meet the condition of "Erstbezug" ("First Occupancy").
  This strict definition may lead to false negatives but this seems to be preferable to the risk of including numerous false positives / duplicates.

  Example WP3 UC2 Sweden: For identifying new constructions among the sold objects we currently use a naïve rule based on the assumption that an object that was sold before or during its year of construction should be considered a new construction.

7.1.2. Coverage across domains

It is one of the major challenges to find out if a smaller number of scraped units w.r.t. a specific domain origins from a smaller coverage or from a smaller target population. For example, a smaller number of rental offers in rural areas can stem from a smaller number of houses to be rent, but also from a smaller coverage. Similarly, a smaller number of online job advertisements in a certain NACE domain can stem from a smaller number of open jobs, but also from a smaller coverage since most companies in this NACE category might prefer to search for new employees not on online job portals.

Proposed guidelines

- Is there a reason to assume that coverage across domains is different?
  If yes, what is your explanation for it?
  If yes, does this affect the calculated indicator?

  Example WP3 UC1 Germany: We observed a relatively small number of rental properties in Berlin. This could be due to the extremely high demand for apartments, leading to properties being rented out before they are even listed online.

  Example WP3 UC2 Germany: The more rural NUTS-3 areas have the lowest mean monthly numbers of advertised objects. These pictures are quite plausible since it could be assumed that newly constructed buildings in rural areas are primarily used by their builders and may appear on real estate portals years after being constructed

7.1.3. Linking web data to a known statistical population

Proposed guidelines for linking web data to the Statistical Business Register (SBR)

Funded by
the European Union

- Describe the linking process of the web data and the statistical register. Be specific about the units you want to link, the variables you use from both data sets for the linking process as well as the linking method you use.

  Example WP3 UC5 Netherlands: Statistics Netherlands (CBS) links the URLs from the DataProvider (DP) to legal units in the Dutch SBR. This linkage method works as follows: first CBS applies some basic text cleaning to standardize variables of both DP and SBR. Next all entries from both DP and SBR are compared. Only a selection of variables (see Table UC-5-NL-1, in [ST23]*) is used in this comparison. Each pair of identification variables is assigned a weight, that is added to the total linkage score when the values of the pair match. This total score is subsequently used as input for a function to generate a confidence score between 0 and 1 (linkage probability function). The confidence score is an estimate for the probability that a 'DP-URL – legal unit' pair is a true link or not.*

  Example WP3 UC5 Finland: Linkage probability of URLs to Business Registry legal units and establishments

- Describe how you evaluated the linking method. When you drew a sample for manual annotation, also describe the sample design.

  Example WP3 UC5 Netherlands[11]: In 2022, CBS has evaluated the quality of the original weights and of the linkage probability function. To that end, four samples were drawn, which were annotated by experts of Statistic Netherlands:
  - *400 URLs from DP that could not yet be linked to the SBR;*
  - *400 legal units from SBR, from legal units for which no DP URL was linked;*
  - *400 legal unit - DP URL links with a linkage probability of at least 50%;*
  - *400 legal unit - DP URL links with a linkage probability smaller than 50%;*

7.1.4. Validation and imputation of NACE codes in the Statistical Business Registers

Texts of business websites can be used for the enhancement of the NACE codes in the Statistical Business Registers (SBR), as was the focus of UC5 in WP3. NACE-classification models can help to detect misclassifications as well as to predict NACE codes for missing values in the SBR, or differently said to impute missing or wrong values.

Proposed guidelines:

- Compare the estimated NACE codes based on a prediction model and the NACE codes from your SBR.
- Compute measures for the prediction performance of the models such as the F1 score or the number of false negatives or the proportion of true positives that are actually found (=recall of the true class)
- Compute number of false positives or the proportion of true negatives that are actually found (=recall of the negative class)

[11] *For more details on the linking process as well as the annotation and its results, see the 2nd Technical report, "Update on linkage process at CBS*

Funded by the European Union

Examples:
WP3 UC5 Hesse: F1 scores of URL finding
WP3 UC5 Netherlands Austria: F1 scores of NACE prediction
WP3 UC5 Netherlands: TPR, TNR of NACE misclassifications
WP3 UC5 Hesse: number of false negative of retrieval of emails and of names of managers of businesses

7.1.5. Duplicates

Duplicated units lead to overcoverage. Unfortunately, duplicates are a prominent phenomenon when web scraping and are often hard to detect. Even within one source, there are cases where it is hard to decide if scraped information such as online job advertisements or real state advertisements on a real estate portal refer to the same or different jobs / apartments. This differentiation becomes even trickier when combining several sources.

Proposed guidelines for duplicates within one source

- Do you have duplicates of observations within a source which you need to remove (deduplicate)?

  Example WP3 UC1 France: One of the main issues has been the treatment of ads' duplication, as a rental offer may be published by several realtors, with the consequence of this offer to appear as many times as it has been published.

  Example WP3 UC1 Germany: Duplicates are sometimes hard to detect, in larger buildings, without full information, distinct objects may appear as duplicates

- Describe your deduplication strategies for duplicates within a source

  Example WP3 UC2 Germany: Deduplication within a portal uses the portal's own ID for an ad or object: only the last / newest ad is kept. Same objects can have different Ids, thus as second step we check address and other characteristics

- Calculate a duplication rate per source, describe also the time interval in question.
  If you tried several deduplication strategies, give a duplication rate for each strategy, or describe the upper bound of the duplication rate for the strictest set of deduplication rules and a lower bound for the least strict set of deduplication rules.
  You can also differentiate between observations you have for sure identifies as duplicates and observations which might be duplicates but where the identification is not 100% sure.

  Example WP3 UC2 Germany: duplication rate 20%

Proposed guidelines for duplicates across/between sources

- Do you have duplicates of observations between different sources which you need to remove (deduplicate)?

- Describe your deduplication strategies for duplicates across sources. Name the variables involved in the deduplication procedure.

Describe if there are cases where it is impossible to determine for sure if an observation is a duplicate or not.

Example WP3 UC2 Germany: treat all offers at same address as duplicates gives lower bound, deduplication needs heuristic approaches because despite the use of satellite data, it is not deterministically to decide always if two offers are duplicates

- If you tried several deduplication strategies, give a duplication rate for each strategy, or describe the upper bound of the duplication rate for the strictest set of deduplication rules and a lower bound for the least strict set of deduplication rules

8. Throughput phase II

8.1. Introduction

This chapter builds on the previously developed "Quality Guidelines for the Acquisition and Usage of Big Data.", see [KO20] . Within the scope of the Web Intelligence Network project, the experience gained during the final step of statistical production—Throughput Phase II, which involves generating statistical outputs—was relatively limited. Consequently, the updates to this section are minimal and primarily adapt the existing "new data source agnostic" guidelines to the current context.

8.2. Replacement of questions from surveys

Example: Survey on ICT usage and e-Commerce in Enterprises Guidelines for this example

Compare information coverage.
First, it is important to compare the coverage of the traditional survey with the possibilities of the big data source. Coverage is one of the most important aspects. Sometimes, for example in Online Job Vacancies data, the definition of job vacancy in the traditional survey may be different than the one used in the big data source (online job vacancies).

Compare definitions.
The second issue is to have a unified metadata set – it is necessary to compare all definitions of data gathered in traditional data sources vs. metadata in big data sources.

Measure and report accuracy of applied models.
Due to the complexity of new data sources, e.g., the data of websites may lead to the use of machine learning algorithms, it is also important to measure accuracy of the data set and the information provided.

8.3. Validation / comparison of results with results from traditional data source

Example: Survey on ICT usage and e-Commerce in Enterprises

A subset of the estimates currently produced by the sampling survey on “Survey on ICT usage and e-Commerce in Enterprises”, yearly carried out by EU member states, includes as target estimates the characteristics of websites used by enterprises to present their business (for instance, if the website

Funded by
the European Union

offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data is collected by means of traditional questionnaires.

These results can be compared with results based on new data sources, e.g., data collected by accessing the websites directly (i.e., via web scraping). The collected internet texts have then been processed to individuate relevant terms, finally the relationships between these terms and the characteristics of interest for the estimates are modelled.

Hence, the sequential application of web scraping, text mining and machine learning techniques represent the prediction approach to produce estimates that can be compared to the ones based on surveys.

In this kind of applications, the comparison allows a large number of quality evaluations: it is possible to compare the variability and the bias due to sampling variance, total non-response and measurement errors in the traditional survey vs the model bias and variance in the prediction approach. Further, one can produce aggregate estimates as well as to predict individual values.

Quality guidelines relevant for this application

- Assess the coverage of the population considered by the new data sources compared to the target population (mainly risk of undercoverage);
- Assess the prediction errors of the model-based approach.

8.4. Survey based estimation with auxiliary information / calibration

Example: Business survey with web-scraped information

In a business survey the question if the company has a web page is asked. Additionally, for all enterprises in the frame, online presence is tested with web-scraping methods. As this information might not be totally equivalent to the survey question definition, it cannot be used directly to estimate the total number / or ratio of enterprises with a web page. However, the web-scraped information will probably be strongly correlated to the response to the survey question and is known for the whole population. A straightforward way to improve the precision of the survey estimates might be to calibrate the survey weights in such a way that the survey estimates for the number of enterprises with an online presence (according to the web-scraping definition) match with the same number for the whole population.

Quality guidelines for this application:

Check definitions.
The variables from the big data source are checked regarding contents and definitions before used in a non-response analysis, weight adjustment or in general in a model assisted survey estimate.

Information must be trustworthy.
The quality of the information needs to be checked before it is used in such methods, since the survey theory regards the information to be known true population values in most scenarios.

Prefer auxiliary information on unit level.
If the auxiliary variable is available at the unit level, it is preferable to a situation with only information on the macro level, e.g. totals.

Funded by
the European Union

Estimators based on base weights are compared with adjusted estimators.
The base weight is a factor; usually the product of the design weight and a non-response factor assigned to each sampling unit before calibration. Estimators of the relevant key figures of the concerned statistics are analysed (e.g., the number of unemployed in LFS). Marginal totals of persons, households or businesses for important breakdowns are analysed.

Describe methodology and short-comings.
It should be described and publicly available how the method is applied and what effect can be seen compared to the base weights (see previous guideline). Possible short-comings should be clearly stated.

Funded by
the European Union

9. Guidelines for a centralised web data infrastructure

Whereas the data acquisition of the use cases of WP3 happened at the infrastructure of the individual NSIs, it was the goal of WP2 to use a common and centralised infrastructure – the Web Intelligence Platform (WIP) – at least for the scraping process. For OJA, the process of landscaping as well as the processing of the scraped data, including information extraction, happened on the WIP. For OBEC, the WIP was primarily used as a scraping infrastructure. Some of the processing steps are publicly available as generic Python-scripts[12] and can be run on the WIP as well. However, the selection of the URLs to scrape, as well as any processing steps that include information from the countries' SBR, require to be run locally in each NSI.

The different roles of WIN members in the usage of the WIP – to scrape, to process or to analyse the scraped or the processed data, but also to advise the further development of the WIP, led to a wide range of experiences. In this chapter we want to collect lessons learned from a user's perspective in the form of guidelines.

9.1. Guidelines about technical requirements of a centralized web data infrastructure

Smooth operation of the scraping processes
The core of a centralised infrastructure is the scraping process. The major advantage for NSIs due to the centralised scraping infrastructure is the elimination of maintenance work for the scrapers. It is therefore most important to design and build the scrapers in such a way that they are robust against the most common changes in websites / portals.

Portability
Design the platform as a service which can be "brought to the data". Thereby the full platform or individual components of the platform can run in different cloud environments or locally on premise at the NSIs. This can avoid legal problems for NSIs who are not allowed to upload sensible information such as elements of their business registers to a shared platform.

Open-Source-First
Use open source software and cloud agnostic tools for all processes of the platform and its components. Open source components should have no knowledge of the environment they are running on, so that they can run on the centralised platform as well as on an NSI's scraping environment

Modularity
The components of the platform should work as standalone modules if needed. Therefore, a custom platform as combination of a (subset) of all platform components in combination with own components could be hosted. Example: a generic ecommerce detection module.

Web-native, user-friendly access modality
The platform should be created in a way that it is to be accessed via a browser, since extra steps such as VPN, SSH tunnels etc. are often not allowed on the NSI infrastructure.

[12] https://github.com/jmaslankowski/StarterKit

Funded by
the European Union

Identity and Access Management (IAM)
The IAM should be customizable in case of self-hosted platforms and use the "EU login" in case of a centrally hosted platform. Include a clear ownership and role system.

Metadata for a transparent, traceable scraping process
For each scraped page / each API accessed a clear log entry stating the time point, the duration and the result should be available.

Scheduling, Prioritising and Resource management
A task, computing resources and cost management tool is needed which lets users in a multi-user environment allocate resources and see priorities and used resources.

Connection to outside data sources
It should be possible to directly use data from other data sources such as registers

Make data easily accessible and transferable
The platform's data storage should be implemented in a way that so that it is easy for authorized users to access data and transfer data to other systems.

Maintenance of scrapers
Responsibilities for the maintenance of scrapers have to be clearly defined and infrastructure maintenance should be performed by the platform operator.

Metadata about the scraping process
Save metadata of the scraping process such as the IP address of the server that responds, the website's "robots.txt" file and the region (and/or IP address) of the bot. Also, the stability of the scraping process should be saved as metadata so as to understand breaks in time series better.

9.2. Guidelines about landscaping for a centralized web data infrastructure

Selection of scraped sources and documentation of the landscaping process
The decision about scraped sources should be in the hand of the future users of the data. For this, it is particularly important to involve the subject matter experts at the NSIs.
If the selection of the scraped sources happens centrally, provide users with a detailed and transparent description about the selection process and the selection parameters.
If there is a need to select similar sources across countries (NSIs), the NSIs should agree on a common selection model (see Chapter xx) or at least on important selection criteria.

Focus on sources where formal partnerships are possible
A centralised web data infrastructure should prioritize formal partnerships, licenses, or contracts established with owners or operators of web data platforms over massive scraping. Such partnerships allow for an authorized access to data, ensuring that the data ingestion is legal, ethical, and sustainable.

Decision about the scraped information
The future users of the scraped data should have in advance the possibility to determine the variables and the specifics of the scraped information. Again, subject matter experts should be involved.

Funded by
the European Union

Coordination of landscaping
Eurostat should be in a coordinating role about the selection of the data sources if the scraped data is used by multiple partners Guidelines about the centralized ingestion process

Transparency
Transparency is needed towards the public, the NSIs as well as the website/data owners, about which sources are scraped, how often the scraping process takes place and the reference period. Especially the future users of the scraped data need detailed metadata about the scraping processes.

Easy access
Trainings and easily accessible instructions should be available to access and make use of the centralized web data infrastructure

9.3. Guidelines about the centrally scraped raw data

Access to scraped raw data
The future data users need full access to the scraped raw data. This allows them to apply NSI-specific process steps on them and compare the results with potentially available scripts shared through the platform.

Versioning of scraped data
In case of repeated scraping, the scraped data has to be available in a versioned form.
Additionally, adding a time stamp to every observation from the web is helpful for later processing of the data and makes the data independent from metadata such as scraping frequency.

9.4. Guidelines about the processing of scraped data on the platform

Transparent, versioned code and models of processing of raw data on the platform
If process steps such as information extraction (e.g. extraction of a classification variables from scraped text) are available on a centralised platform, the code should be open to all data users. If the code is under development, it is important to provide versioned scripts. In this way, data users can understand how the extracted information is generated, identify potential sources of errors, and potentially further develop the code.
When automated information extraction models are used (e.g. classification models according to international classifications such as NACE or ISCO), the scripts and the models should be provided with the NSIs.

Shared code for processing of raw data
The scripts and tools to process the scraped data should not only be available to all users, users should also have the possibility to create their own processing scripts, to reuse existing ones or to run (parts) of the workflows themselves. Tools should be in place so that users can interact.

Transparent quality assessment of automatic information extraction model
Quality checks should be in place about the automatic information extraction from scraped data. It is necessary to do manual annotation exercises to control if automated classification works as well as expected. The resulting quality indicators such as accuracy should be made available to data users and made public if the resulting data is used for the production of Official Statistics, see also Deliverable “D4.8 Quality Assessment for the statistical use of webscraped data”.
These quality checks need to be repeated regularly so as to be able to see if the used models have

been improved and to check if the used models are still adequate or if they suffer e.g. from concept drift.

Standardized quality assessment
When different NSIs follow a common standard for processing scraped data, the same approach should apply to how they evaluate the quality of those processes. This approach ensures consistency, comparability, and reliability in both data processing and quality assessment across different institutes.

Consultation of subject matter experts
Subject matter experts at the NSIs should be involved in the assessment of the quality of the information extraction from the scraped data.

Training data
Generally, it is a difficult, expensive and work-intense task to generate curated manually labelled training data, sometimes referred to as gold standard. The size of such a data sets needed to be able to train a model for the complicated classifications widely used in official statistics, is the main driver of cost

Funded by
the European Union

10. References

[CON19] Condron et al. (2019). ESS web-scraping policy template. Workpackage C, Implementation – Enterprise Characteristics, ESSnet Big Data II, Deliverable C1.

[CO21] Colombo, E. et al (2021). Landscaping OJA Web data sources. Deliverable D1.1 – OJA landscaping methodological guide, Version 3

[DA22] Piet Daas et al. (2022). Web intelligence for measuring emerging economic trends: the drone industry, Statistical Working Papers

[KO20] Kowarik, A. et al (2020) Deliverable K3: Revised Version of the Quality Guidelines for the Usage and Acquisition of Big Data, Work Package K Methodology and Quality, https://raw.githubusercontent.com/essnetbigdata/deliverablesESSnetBigdataII/main/WPK/K3.pdf

[KO16] Körner, T. and Rengers, M. et al (2016), Inventory and qualitative assessment of job portals, Deliverable 1.1. Work Package 1 Web scraping / Job vacancies of the ESSnet on Big Data. https://cros.ec.europa.eu/group/31/files/1063/download

[PU21] PUlearning (2021), Website of the pulearn python library, https://pypi.org/project/pulearn/

[SIL11] Silla, Carlos N., and Alex A. Freitas. (2011), A survey of hierarchical classification across different application domains, Data Mining and Knowledge Discovery 22:31–72.

[SUN01] Sun, Aixin, and Ee-Peng Lim (2001), Hierarchical text classification and evaluation, in Proceedings 2001 IEEE International Conference on Data Mining, 521–528. IEEE

[ST22] Stateva, G. et al (2022). Deliverable 3.1: WP3 1st Interim technical report, Essnet Trusted Smart Statistics – Web Intelligence Network, Grant Agreement Number 101035829 — 2020-PL-SmartStat, Work Package 3 New Use Cases

[ST23] Stateva, G. et al (2022). Deliverable 3.2: WP3 2nd Interim technical report, Essnet Trusted Smart Statistics – Web Intelligence Network, Grant Agreement Number 101035829 — 2020-PL-SmartStat, Work Package 3 New Use Cases

[TR22] Trentini, F. (2022). Landscaping OJA Web data sources. Deliverable D4.1 – OJA sources ranking model report

Funded by the European Union