Quality Aspects of Web Data based on the Experiences of the ESSnet Trusted Smart Statistics – Web Intelligence Network

Magdalena Six
Statistik Austria, Vienna, Austria– Magdalena.Six@statistik.gv.at

Alexander Kowarik
Statistik Austria, Vienna, Austria – Alexander.Kowarik@statistik.gv.at

The ESSnet "Trusted Smart Statistics – Web Intelligence Network (WIN)"

The ESSnet "Trusted Smart Statistics – Web Intelligence Network (WIN)" is a project within the European Statistical System (ESS), which engages 17 organizations from 14 European countries. It aims to develop a web intelligence system at the ESS level, providing a greater chance to generate the right conditions for the integration of web data into official statistics. The Web Intelligence Hub (WIH) is the pillar of Trusted Smart Statistics that provides the fundamental building blocks for harvesting information from the Web to produce statistics. The ambition of the WIH is to become a high-quality source of web data, methodologies and algorithms ready to be used to produce European and national official statistics.[1]

One work package (WP2) of this ESSnet includes already well-established use cases such as online job advertisements (OJA) and online-based enterprise characteristics (OBEC), with the ambition to be moved into the statistical production stage soon. Another work package (WP3) focusses on new types of web data sources, such as web data about the real estate market, construction activities, online prices or hotel prices. For these use cases the aim of the ESSnet is to produce experimental statistics.
Building on the experiences made in the different use cases, a work package of its own (WP4) aims to consolidate knowledge gained in the WIN in the area of methodology, architecture and quality when collecting, processing and analysing web data.

In this paper, we will first present the structure and some examples of the already published quality guidelines for web based data. In the second part, we introduce ongoing work of WP4 and present concepts for the landscaping and selection of websites which have the potential to be scraped.

Quality Guidelines

Based on the more generic work of previous ESSnets (Big Data I & II), the members of WP4 already collected and published "Minimal guidelines and recommendations for implementation" w.r.t. quality, methodology and architecture.

[1] More information on the ESSNet can be found here https://ec.europa.eu/eurostat/cros/WIN_en

Structure of the quality guidelines

The included quality guidelines are structured along two phases of the statistical production process, the input phase and the throughput phase, see also Figure 1. The throughput phase refers to two different processes and is split into two parts. The first part of the throughput phase is dedicated to deriving- so-called - statistical data from web data.
In the Throughput Phase I we pay particular attention to the following quality aspects:

- linking (e.g. linking a scraped website to a business register)
- coverage
- measurement errors
- model/processing errors (e.g. errors when extracting classification variables)
- comparability over time (e.g. gaps in the times series due to failed scraping processes).

The Throughput Phase II deals with the usage of the derived statistical data to produce statistical output.

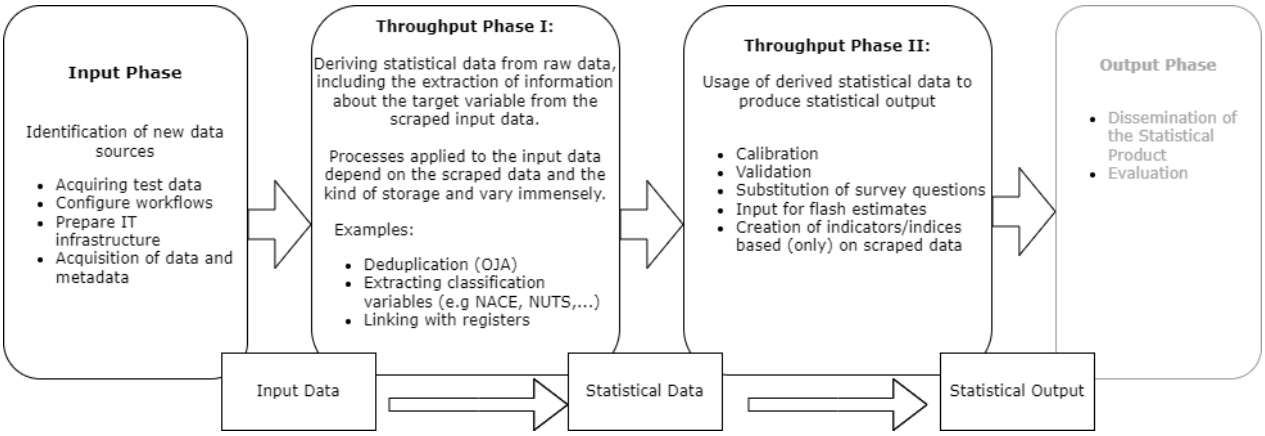

Figure 1 Phases of statistical production process

Examples of the quality guidelines

We only list here a few examples and refer to the document “Minimal guidelines and recommendations for implementation” for further information [Ko21].

Examples for guidelines with respect to coverage and 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.

- Make a pilot web scraping to assess what information is included on the websites. Check if specific information, e.g. territorial unit or industrial sector, can be extracted from the website. When information on the website is limited, it is also not very likely to monitor enterprise activity (e.g., innovations in enterprises) on the website.

Examples for guidelines with respect to comparability 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 date of publishing this information.
- When the website is not up to date it is unlikely to detect enterprise activity in longer time series.

Ongoing quality work of WP4

Landscaping and Selection of Websites

The existing quality guidelines covered the "input phase" of the production process – and gave valuable recommendations on how to ingest data from websites. They also incorporated the so called ESS-webscraping policy, including principles and practices for members of the ESS when engaging in webscraping.

The experiences from the ongoing work in the ESSNet WIN show, that from a quality perspective, too little focus was laid on the part of the production process before the actual ingestion of web data from websites starts: Landscaping data sources and the included selection of data sources plays a major role in the resulting usefulness and trustworthiness of the thereafter ingested data.

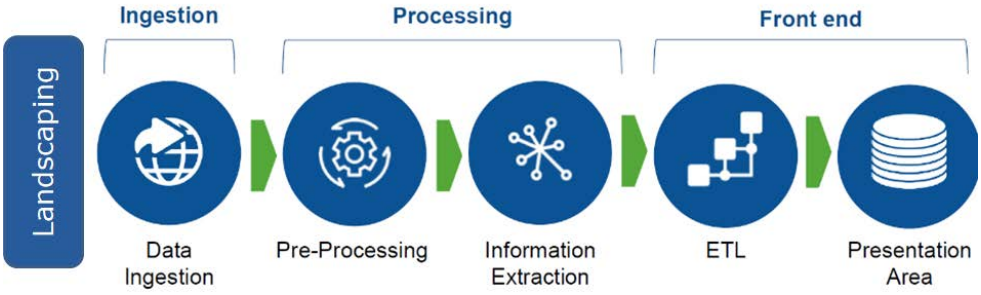

Figure 2 Data Pipeline for online job advertisement data, taken from the WIH for OJA data

Within a company or organisation, the term "landscaping" is used for cataloguing and measurement of all the data within the company or organisation. Similarly, in the world of web-based data**, landscaping** refers to the **cataloguing** and **partly also measurement of all web-based data sources** relevant for the topic of interest.

Depending on the topic of interest, the effort of the landscaping exercise can vary enormously: E.g., in case of satellite data, all the needed data might be available on one single website. In case of online job advertisements (OJA), real estate prices or other price statistics, the great extent of existing websites – with varying market share, varying trustworthiness, varying legal and technical characteristics etc. – makes it necessary to select specific websites. In other cases, such as the topic of enterprise characteristics, the relevant information should be scraped from all websites available.

It also depends on the topic of interest, if the number of scraped websites should be maximised or if there is also a downside to having a high number of sources and the selection of websites becomes central.

Two examples illustrate this observation:

- In the case of scraping enterprise characteristics (WP2, OBEC), the total size of the population studied - the number of enterprises with a website active in a country – is not known. It is the goal of the landscaping exercise to detect as many enterprise websites as possible. Each additional website of an enterprise which is detected and scraped provides additional value. Further, the scraping process is only repeated in relative long time intervals, since the scraping process is resource-intensive and updates of enterprise characteristics are not expected to change very often.
- In the case of online job advertisements (WP2, OJA), the total size of the population studied –all enterprises searching via online advertisements for new employees or all online job advertisements – is not known either. Additionally, the same job advertisements can be found on different job portals, leading to the non-trivial processing step of de-duplication of scraped advertisements. Further, job portals vary in their share of original and new job advertisements, some job portals may only contain redundant or outdated information. The scraping process is repeated in relative short time intervals (e.g. once a week) to harvest new job advertisements. Thus, the stability of the source and the repeatability of the scraping process is important, since failed scraping rounds lead to irregularities in the time series.
  All these observations make it clear that it can not be optimal to scrape every possible job portal: additional job portals lead to more redundant information which has to be filtered out afterwards and the probability of failed scraping rounds increases with the number of sources.

Selection of websites

When it is necessary to select 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 sources becomes obvious.

A generic answer to the question “Which websites should be selected?” would probably be: “The most important ones” or “The ones with the highest quality”. But quantifying importance or quality can be rather tricky.

We differentiate between three categories of information for the assessment if a website should be selected for scraping.

- **Information from the website itself (Inf1)**
  This includes all sort of technical and content-wise information than can be found on the website itself. If 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 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 websites played a role for several use cases in WP2 and WP3. Selection models were developed for the use case Online Job Advertisements in WP2 as well as a common selection model for all use cases of WP3.

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[2].

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 concern 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.

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),

[2] 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.

- the displayed form (structural field, text or mixed) for variables such as “Type of Occupation”, Type of Contract”, “Working Time” etc,
- further variables

These categorical variables have to be transformed into numerical values and combined to allow for an overall quantitative assessment of the website. This process has to 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 taken into account 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 groups 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 for each source. Then, the calculated distribution is compared to the distribution known from the Labour Force Survey (see [Co21]).
The website with the most similar distribution of the respective variable are ranked lowest. This holds also for stability and popularity: the more stable and the more popular, the smaller the rank.
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 1.

Table 1 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 possible to name a threshold 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.

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 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 rejected**).
  Examples for this 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 determined content-wisely (e.g. the price in ads about real estate has to be given). Further, mandatory variables can also be of technical nature such as the existence of a URL of the offer and an advertisement ID.

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 for 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 not that 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.

Literature

[Tr22] Trentini, F. (2022). Landscaping OJA Web data sources. Deliverable D4.1 – OJA sources ranking model report

[Co21] Colombo, E. et al (2021). Landscaping OJA Web data sources. Deliverable D1.1 – OJA landscaping methodological guide, Version 3

[Ko21] Kowarik, A. et al (2021). Deliverable 4.1: Minimal guidelines and recommendations for Implementation, ESSnet Trusted Smart Statistcs – Web Intelligence Network
Grant Agreement Number: 101035829 — 2020-PL-SmartStat, Work Package 4
- Methodology and Quality, https://cros-legacy.ec.europa.eu/content/deliverable-41-minimal-guidelines-and-recommendations-implementation_en

[St22] Stateva, G. et al (2022). Deliverable 3.1: WP3 1st Interim technical report, Essnet Trusted Smart Statitics – Web Intelligence Network, Grant Agreement Number 101035829 — 2020-PL-SmartStat, Work Package 3 New Use Cases, https://cros-legacy.ec.europa.eu/content/wp3-deliverable-31-wp3-1st-interim-technical-report-20220330_en