Navigating quality challenges in landscaping web data: New aspects and source stability

Magdalena Six, Alexander Kowarik[1]

Abstract

This paper delves into the challenges of landscaping web data, specifically focussing on the development of quality aspects for new data sources. It highlights the limitations of traditional quality dimensions when working with web-scraped data and emphasises the need for additional considerations along the data processing pipeline. It explores the process of website selection, emphasising the importance of a standardised assessment tool to ensure comparability between different countries. Moreover, it discusses the impact of source stability on data quality, illustrating how unstable access to data sources can block accurate analysis and limit the reliability of statistical indicators. Real-world examples showcase the complexities of interpreting observed web data, further emphasising the significance of reliable and stable data sources.

Keywords: Web scraping, quality, relevance of sources, landscaping.

1. Landscaping of websites for web scraping with focus on selection models

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

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.

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 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.

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 where landscaping seems to include all processes before the data ingestion starts, we give our own definition as follows:

- 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.

1 Magdalena Six (Magdalena.six@statistikgv.at), Alexander Kowarik, (Alexander.kowarik@statistikgv.at), Statistics Austria. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of Statistics Austria. This work has been part of the European project "2020-PL-SmartStat Web Intelligence Network". Part of the text has been released published elsewhere, especially as deliverables for the mentioned project

The three sub-processes cataloguing, measurement and selection build and depend on each other. Starting from the first cataloguing, this can be an iterative process between measurement and selection and even cataloguing.

Depending on the topic of interest, the difficulty of the landscaping exercise as a whole and of each sub-process can vary enormously.

There are two dimensions that are central for understanding the complexity of the landscaping for a certain/topic:

- if all websites should be captured or representatives should be selected;
- if additional information is available or not.

The most common kind of additional information is information from the statistical agency itself, *e.g.* names of businesses in a certain branch from the business register or relevant companies for online-shopping of clothes.

If a representative subset of websites should be selected, the concept of representativity, which is not defined mathematically, but mere as an idea, has to be operationalised in some way, *e.g.* by a random sample or a cut-off sample.

1.1 Cataloguing

This process is easier if the starting point is the list of enterprises being part of a certain population of interest. If you have the companies names you can use them specifically in your online search, which will lead to more specific results.

It can also be the aim to actually identify the population of enterprises/units active in a certain field, 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.

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 / respective topic of interest. Especially, when you want to capture all websites of a large target population, the sheer number of URLs makes it impossible to visit each website "manually" to decide if the URL belongs to the target population. You therefore need an automated mechanism, which estimates for each URL in your catalogued URLs if it belongs to the target population.

1.2 Selection of websites

In the case of starting from a population of enterprises/units, you have a list of enterprises 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. So, the selection step is selecting the valid URLs that correspond to statistical units, mostly via linking procedures. Additionally, you have to check technical criteria if a website can be scraped, *e.g.* by looking at the robots.txt. The validity check often contains checking unique identifiers that (by law) must be present on a company's website such as the value added tax or company registration number.

If no additional information is available for the population, selection mainly corresponds to classifying websites if they belong to your target population, *e.g.* selling or developing a certain product. This can be done for example with:

- word-based methods: counting the occurrences of certain keywords;
- semi-supervised learning: a set of positives examples needs to be available;
- supervised learning: a training data set with positive and negative examples must be available.

1.2.1 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 obvious.

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.

The need to decide for/rank several websites with respect to specific criteria such as the popularity of the website, the trustworthiness of the website owner, the structure of the data on the website etc. shows that the task fits well to the framework of Multi-Criteria Decision Making (MCDM) . 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).

There are in general three groups of criteria that can be used in the selection process: 1) information from the website itself, 2) information about the website and 3) information from previous scraping of a specific website. The inclusion of all three categories of information might be costly but leads to the most trustworthy selection of websites suitable for scraping.

2. Quality indicators to measure the relevance and stability of selected OJA sources

In this chapter, we focus on a specific aspect of the whole production process – the relevance and the stability of the selected sources based on the examples of job portals. We do this from the perspective of a user of the pre-processed data from the central Eurostat platform – the web intelligence hub, who has limited insight and control in Eurostat's landscaping decisions and advanced selection models. We believe that this is a very meaningful exercise as it is close to the current actual situation when using the centrally scraped OJA (Online Job Advertisement) data– it shows if the producers' efforts in landscaping OJA sources fulfil indeed the users' expectations about the relevance and the stability of the sources.

2.1 Indicators for relevance of selected sources

If a producer of Official Statistics is not in charge of the landscaping process itself, it is crucial to check if the included sources are indeed the most important ones and correspond to the sources that would have been considered by NSI domain experts as well. For this we propose to consider the following indicators:

- if your NSI scrapes OJA data itself, compare the included sources from your own scraping processes with the included sources on the Web Intelligence Platform (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.

2.2 Indicators 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 with the indicator 'number of vacant positions advertised online'. 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 stability of the time series is impacted by changing sources included in the OJA data.

Creating a time series by simply adding up all unique OJAs over an instable number of sources probably won't capture effects from the labour market, but effects reflecting the inclusion of certain sources at a certain point in time. *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. For this we propose to look at the following indicators:

- determine if it is always the same sources in the course of the time span considered;
- 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?

2.3 Indicators 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 in the job market. The following indicators can give you hints if such a phenomenon occurs in the data:

- 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.

2.4 Indicators for the stability of sources over different versions of data

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

To measure the stability between different versions of data, we propose the following steps:

- 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.

References

Kowarik, A., P. Daas, M. Bruno, M. Six, O. ten Bosch, G. Ruocco, C. de Maricourt, V. Chavdarov. 2021. “Minimal guidelines and recommendations for Implementation”. *Deliverable 4.1 ESSnet Trusted Smart Statistcs – Web Intelligence Network, Grant Agreement Number: 101035829 – 2020-PL-SmartStat*. Luxembourg: Eurostat. https://cros.ec.europa.eu/system/files/2023-12/deliverable_4_1_minimal_guidelines_and_recommendations_for_implementation_essnet_tss_win.pdf.