Estimating Enterprise Characteristics from Web Data: Achievements and Future Developments

**Keywords:** Web data, Big Data, Enterprises Websites

# Introduction

Internet is one of the most interesting Big Data sources for Official Statistics. Indeed, while for other sources, like mobile phone data or smart meters, there is the need to engage partnerships with their providers, Internet data are publicly accessible. Internet as a Data Source (IaD) data can be used in substitution or in combination with data collected by means of traditional survey-based instruments. In case of substitution, the aim is to reduce respondent burden, in case of integration the increase in accuracy of the estimates is the main goal.

Among the possible uses of IaD, data from enterprise websites are particularly relevant for Official Statistics. During the last few years, the vast majority of enterprises acquired an Internet domain in order to set up an official website, thus making available (almost) for free several information that previously was available only via traditional collection systems. Hence, it is recognized as an opportunity for National Statistical Institutes to collect and to mine the publicly available information on these websites to describe a wide range of phenomena in near real-time.

Given this context, the ESSnet Big Data Pilots I was launched by Eurostat early 2016 and is concluded in July 2018. Within such a project, the purpose of workpackage “Web Scraping of Enterprise Web Sites” was to investigate whether web scraping, text mining and inference techniques could be used to collect, process and improve general information about enterprises. The project will have a follow-up, namely the ESSnet Big Data Pilots II that includes again a specific workpackage, “Enterprise Characteristics”, aiming at conducting to an implementation stage the piloting activities carried out within the first ESSnet project.

In this paper, we first summarize the results achieved within the first project strand, then, we will highlight the main developments foreseen for the future project activities.

# ESSnet Big Data Pilots I: Web Scraping Enterprise Characteristics

Within the workpackage “Web Scraping of Enterprise Web Sites”, the members (IT with a coordination role, BG, NL, PL, SE, UK) identified six (principal) use cases aimed to estimate characteristics from enterprises websites. For such use cases, a piloting activities was conducted, eventually resulting in the publication of experimental statistics on the project wiki[[1]](#footnote-1). More specifically, the following output indicators where computed:

* Rate of retrieved URLs from an enterprise list
* Rate of enterprises engaged in web sales on their website
* Rate of enterprises with job advertisements on their website
* Rate of enterprises that are present on social media
* Percentage of enterprises using Twitter for a specific purpose, estimated from web data

## The Processing Pipeline

The use cases were implemented according to a processing pipeline shown in Figure 1.

After a step of *Internet acc*ess, consisting of the URL retrieval and the websites scraping, a *Storage step* organizes (unstructured) scraped data for a *Data prepar*ation step that process texts and, in particular chooses a *Text representation* method, among: (i) Bag of words, (ii) Language encodings (word embeddings) and (iii) Engineered features (selection of keywords intended to represent the text). Finally, the *Analysis step* can be either *Machine learning* based or *Deterministic*.

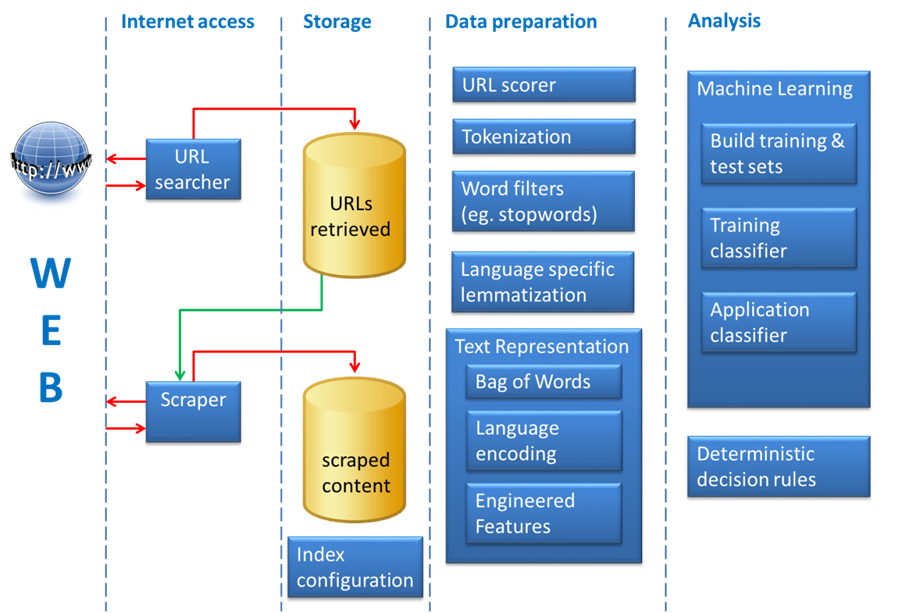


Figure 1: The logical architecture underlying the processing pipeline

## Methods and Results

From a *methodological perspective*, both deterministic and machine learning methods were used in the pilots. On one side, we learned that even with different methods good results can be achieved. On the other side, however, we saw that in some cases there can be a convergence of methods (e.g. the URL retrieval pilot where Italy, Bulgaria and the Netherlands applied the same methodology). Predicted values can be used for a twofold purpose: (i) at unit level, to enrich the information contained in the register of the population of interest; (ii) at population level, to produce estimates. Both unit level and population level quality have been evaluated in the piloting phase. With respect to quality evaluation at unit level, when employing supervised machine learning methods the quality can be measured by comparing against labelled data in a hold-out ‘test’ set. If this test set is representative of the whole population and not used for training the model, the performance measures (like accuracy and F1-score) calculated for the test set can be considered a good estimate for the overall performance. The unit level quality evaluation has been performed by all pilots. The issue of measuring the quality of population estimates making use of predicted values has also been addressed, for instance by Italian pilots, but more work is needed for a drawing conclusions on this issue.

From an *IT perspective*, several tools needed to be ad-hoc developed and software solutions are available on the project website[[2]](#footnote-2). Such solutions include: (i) Web scraping frameworks, (ii) implementation of machine learning and deterministic decision models, and (iii) full-fledged software solutions implementing specific use cases.

The overall quality of the results at unit level was good, in some cases (e.g. Rate of enterprises that are present on social media) even excellent. The quality of the results at population level has been only partially evaluated but comparisons with survey estimates are promising.

As concluding remarks for the first phase of the project, we highlight that:

1. The project proved that a whole pipeline from data collection to analysis can be put in place and produce quite good results. The pipeline implemented in the first ESSnet was an ***experimental*** one and paved the way for the ***implementation*** in the second ESSnet***.***
2. Moving towards implementations the proposed solutions poses two main categories of issues, namely:
   1. Methodology: quality evaluation in the new framework of estimates from Web data still deserves some investigation.
   2. Technology: web scraping and text processing tools need to be evolved for an enterprise-level usage in terms of volume, access, processing speed and technical organisation. Massive scraping might require improved bandwidth and parallel scraping.
   3. Organizational: re-organization of production processes and capability building actions should be designed and implemented.

# ESSnet Big Data Pilots II: Enterprise Characteristics

The ESSnet Big Data Pilots II is a continuation of the successfully finalized ESSnet Big Data Pilots I. Among the achieved results, two major categories will constitute the starting point of the new ESSnet:

* URLs retrieval methodology, a process and software implementations for detecting websites of enterprises based on search engines and machine learning techniques;
* Methodologies, processes and software implementations for detecting characteristics of enterprises such as E-commerce activities, Social media presence, Job advertisements, NACE code, etc.

These results can already be implemented and deployed in some ESS countries, with some adaptation to the local circumstances. In the new ESSnet project, the methodology and tools of the previous ESSnet project will be generalized and extended to be used by any ESS country, taking into account the variety of requirements that such countries may have.

In the scope of the new project, the workpackage WPC “Enterprise Characteristics” will be in charge of such efforts with participating countries: AT, DE, FI, IT, NL, PL, UK and BG as a coordinator. The main tasks of WPC are:

* *Development of an ESS web-scraping policy*, aimed at designing and developing transparent web-scraping policies in order to allay public concerns about the data collected and the usage of them. Extracting knowledge from online data draws attention and public concerns how NSIs are utilising online data, including relatively uncontroversial cases such as NSIs utilising textual data on company websites. GDPR also adds further requirements to the web scraping activities.
* *Development of Methodological Framework/Guidelines*, with the goal of producing generalized and extended methods, procedures and implementation requirements for web scraping on enterprise characteristics. The methodology will be based on results from implementation of use-cases: Enterprise URLs Inventory, E-Commerce in Enterprises, Social Media Presence on Enterprises webpages, NACE identification and Job vacancies ads on enterprises’ websites.
* *Production of Experimental Statistics*, including reference metadata, aimed at producing experimental statistical products using web scraped enterprise characteristics data as their main source.
* *Development of Starter Kit for NSIs*, consisting of procedures for testing and maintenance of web scraping.
* *Development of Quality template for statistical outputs*, finalized to the development of a quality management template for web scraped enterprise characteristics to ensure that the quality level is satisfactory for disseminating experimental statistics. The starting point of the quality template will be the UNECE Framework for the Quality of Big Data and all the quality aspects already identified by WP8 of the previous ESSnet.

# Conclusions

The undertaken route for the production of Official Statistics using Internet as a Data Source includes the ongoing work described in this paper as an important milestone.

As described, the work on using enterprise websites to support business statistics, addresses the whole production pipeline from data collection to data dissemination and impacts on technical, legal and organizational levels.

The promising achieved results, as well as the concrete actions planned for facing implementation issues, are expected to move this pipeline towards a full-fledged statistical production in the short-medium term for several countries of the European Statistical System.

1. <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Experimental_statistics1> [↑](#footnote-ref-1)
2. Available at: <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Links1> [↑](#footnote-ref-2)