26th Meeting of the Wiesbaden Group on Business Registers
- Neuchâtel, 24 – 27 September 2018

Bianchi G., Consalvi M., Gentili B., Pancella F., Scalfati, Summa D.
Istat

Session No 2 – Innovation in Statistical Business Registers

New sources for the SBR: first evaluations on the feasibility of using big data in the SBR production process

Abstract

(300 – 500 words)

Keywords: Big data; sources for SBR; Web Scraping techniques; Text Mining and Natural Language Processing techniques

Istat has been working on a project about big data usage to support the Statistical Business Register since February 2018. It is one of the first projects to participate in the Laboratory for Innovation (LabInn), a tool that Istat has been providing since 2018 to introduce process and product innovations. The main idea is to use big data as an additional source for the SBR, through web scraping and text mining technologies, with the aim of integrating the 'structured' business data with the 'unstructured' data coming from web pages. Expected results will be the dissemination of new experimental statistics expanding the SBR, in addition to the traditional business statistics.

The experimentation considers a sample of about 100,000 enterprises of the SBR, stratified by size in terms of employment/turnover, by economic activity and legal form. The implementation starts with two different approaches depending on whether the enterprise web address (URL) is available or not in advance. In case the URLs from administrative sources are available, URL syntax validation is performed, including check of the recurring errors and domain extraction from URL, testing new procedures created specifically for this purpose. In case they are missing, some URL Retrieval techniques are used, performing batch queries on the search engines by means of the available SBR enterprises identification characteristics, or directly downloading information from some proper thematic directory sites. In both approaches it is relevant to proceed with the identification of the true web address, in order to confirm the exact correspondence between websites and enterprises in the Business Register.

The techniques used will be briefly shown in the paper: Web Scraping techniques for web data acquisition (massive scraping of URLs from list; scraping by using the search engines in batch mode; specific scraping from thematic directory sites) and Text Mining techniques (using also Natural Language Processing techniques) for extracting the information to integrate the Business Register.

The main risks and some critical issues will be treated in the paper, regarding both the Istat internal organization aspects, the applicability and usability of the first prototype. There are also serious risks

such as the difficulty of certifying the quality of information (reliability of data), how to attribute information with certainty to a SBR enterprise and how to classify with certainty the universe to which a set of data extracted from the web belongs. Despite risks there are also obvious benefits such as enriching statistical production with new information, increasing the timeliness of statistical products in a short time and increasing the relevance of business statistics at a lower cost than enlarging the existing data. Some issues like the updating and maintaining strategies are questions that still need a proper answer.