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 5 – New Data Sources

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

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

1. Introduction

In the current situation, characterized by a strong dynamism, the need for varied and timely information at multidimensional level increases and official statistics is called to respond efficiently and effectively by intervening and transforming the production data model. Understanding the transformations taking place and therefore "measuring" society and the economy are the new challenges for the statistical institutes in the third millennium. It is an increasingly complex task, made even more challenging since the NSIs have to respond to these new requests leaving unchanged the awareness of being producers, researchers and guarantors of the quality of official statistics.

The approach of the Italian National Institute of Statistics (Istat) with respect to the new complexities, both of phenomena and of data, has been to adopt new strategies that provide the ability to integrate data from different sources: traditional surveys, data already held by the public administration and innovative sources such as Big Data. The aim is to reduce the statistical burden on respondents while enriching the offer, quality and timeliness of the information produced.

Traditional quality frameworks are not sufficient to tackle the complexity of Big Data. By their nature, their characteristics – environment in which they are produced, usability, validity, need for harmonization with respect to the definitions and classifications of official statistics – are different from the canonical sources, so that they make traditional systems of data collection and data processing no longer adequate. For these reasons Istat has started to implement infrastructures and software for the treatment of Big Data (Sandbox and Cloudera) with projects involving enterprises, universities and research centres.

Starting in 2013, Istat has launched several projects on the use of Big Data: *Persons and Places* (Mobile Phone Data), for example, aims to integrate the use of mobile (tracking) data in the statistical process of estimating flows of inter-municipal population, producing origin/destination matrix of daily mobility for work and study at the spatial granularity of municipalities; *Labour Market Estimation* (Google Trends) can be used to improve estimates made on the labour market in terms of forecasts and nowcasting. Other projects launched concern the use of 'S*canner Data'* for estimating consumer prices; of web-scraping techniques for the estimation of the use of Information and Communications Technology by companies and social media (Twitter, Facebook) for the estimation of confidence indicators. In these ongoing experiments within Istat the results obtained are generally encouraging and promising, in comparison to analogous statistics obtained with official data.

Furthermore in Istat, for the first time, experimental statistics **[1]** have been produced by using Internet data, one of the most important Big Data sources, to obtain a subset of the estimates currently produced by the sampling survey on yearly survey "ICT usage in Enterprises". Target estimates of this survey include the characteristics of websites used by enterprises to present their business (for instance, if the websites offer e-commerce facilities or job vacancies). ICT survey data have been used as a training set for fitting the models to predict values, and administrative data in the SBR have been used for handling representativeness problems[1].

Many other examples of the use of Internet data sources can be reported. In more detail, these experiments concern the use of web scraping and text mining techniques. They have been applied with the aim of substituting traditional instruments of data collection and estimation or of combining them in an integrated approach to produce the required estimates, harnessing both survey data and data from the Internet.

Using this experience, while planning the development of permanent censuses, in 2017 Istat started to experiment the possible use of Internet data as a new source for statistical business registers, with the aim to promote and start the integration of traditional and new sources, improving the already started processes. For business registers this opportunity becomes more effective due to the massive presence of business units on the web. Consider, for example, companies that use the Internet for advertising purposes and have full interest in being on the network with constantly updated information; or to the receptive structures that find on the net the way to spread their contacts.

In the last part of 2017 a preliminary analysis started on the feasibility of the acquisition and the possible use of web data for the SBR, that could continue and even become wider taking advantage of the opening of the new Laboratory for Innovation (LabInn) in 2018, that offers the opportunity to dedicate time to research and development of innovative projects and provides useful infrastructures to test their ideas in a dedicated space. The next sections are dedicated to briefly describe the preliminary analysis (§2.) and the strategy chosen for the project (§3.); in the following ones the techniques used (§4.) and the results obtained (§5.) in this first experimentation will be presented, while in section §6. the reference framework architecture for web data processing will be schematically shown. Finally the lessons learned and future work will be summarized to provide an overview of the main conclusions (§7).

2. The *register-based* approach to the BigData and the multidisciplinary context of Istat LabInn

In October 2017, an experiment was started for the enlargement of the informative content of the SBR in order to provide concrete support for statistical production. In particular, the aim was to integrate the structural data of the business units in the SBR with the new information obtained using web

[1] This result is of the utmost importance, as it will allow to make use of the Internet data instead of the traditional survey data in many circumstances, whenever a (small) subset of data will be made available as training set, not necessarily obtained by costly repeated official sample surveys: it will be sufficient to select (under a rigorous probabilistic sample design) a subset of enterprises with related websites, to manually access them and to observe the values of the target variable we are interested to. Then it will be possible to fit models applicable to the generality of websites that will be accessed with the usual web scraping techniques.

scraping and record linkage techniques and using the new information on enterprises to start a more detailed statistical analysis, finding new classifications and new taxonomies to support a better interpretation of new emerging economic phenomena. The objective therefore consisted of three scenarios, whose activities have been carried out in different ways, even in a parallel way, even if the focus in this first part of the project was exclusively on the first two:

- Target 1: Increase quality of information in the SBR for improvement of sample designs
- Target 2: Expanding informative content of the SBR (new items, documentation system)
- Target 3: New business taxonomies (‘emerging’ populations, to be mapped over time)

The team of Istat researchers (namely the authors of this paper) has been working on these targets since February 2018, while the last months of 2017 were dedicated to this preliminary analysis. The main idea was 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.

The first months of the experimentation were used to carry out both a feasibility study, with a strong thematic significance and a domain analysis. Firstly, the main elements of a company website were studied, to understand which could be the best way to address the next investigation. A small number of enterprises has been observed (around 350 units, taken from the SBR), having different size and economic activity performed, and their websites were deeply analysed ‘manually’ by the SBR staff focusing on their website structure, to find out if some informative blocks could be identified, containing the economic phenomena to detect, possibly corresponding to precise areas in the websites.

The result of this **mapping activity** was the set-up of a theoretical matrix ‘thematic areas–domains–phenomena’ and a corresponding empirical one ‘thematic areas–keywords’. Based on the phenomena of interest for our Directorate for Economic Statistics, the first matrix contains the main macroareas: *Identification*/*Activity*/*Localization*; *Products*/*Markets*; *Knowledge Economy*; *Internationalization*; *Governance*; *Customers*/*Suppliers*; *Economic information*. Each informative block is linked to the statistical domains who could be interested in acquiring this kind of information. For example, *Internationalization* is crucial for FATS and International Trade statisticians, while *Knowledge Economy* is significant for R&D and ICT. Furthermore they are also linked to the measurable phenomena of interest, such as the existence of a *pdf* document describing the governance for *Internationalization*, the R&D or innovation features or the existence of patents for *Knowledge Economy*.

Afterwards by surfing the sample websites it was possible to create the empirical matrix, by counting the frequency of keywords in each block and in each area of interest, also looking for the corresponding area of the website that better identifies that specific economic phenomenon, since it was found out that the hierarchical structure of the sites is standard enough for each macro area. Thus a list of significant keywords for each information block has been prepared, looking at keywords that in a common way could identify information and their cataloguing. The assumption to be verified was whether these sets of words are a-priory keys or, more likely, they follow the thematic organization of the site. The basic idea was that the same keyword could be meaningful in a context and at the same time have a different meaning in another context. In the following automatic step they should be searched only in the proper area of the website[2].

[2] The list of keywords by area of interest and their frequencies in the sample are not reported here, since for the purpose of its use the list was built in Italian.

Looking at the design and organisation of the sites it was decided at this stage that the next web scraping activity could stop at the ‘second level’ of the website structure, since here all the information useful for the subsequent processing is exhausted. Another important result of this preliminary mapping activity was that for each enterprise the presence on its own website of the focal pieces of information was registered so as to make us have an initial idea on how much these will be usable, to check how well the information is structured and what the level of encoding is. Hence for each company the presence / absence of the main information of interest was recorded, such as: the web address (URL), the fiscal code, the VAT code, contact information (company name, home address, contact details...), downloadable balance sheet (pdf), typology of site (reliability - degree of site security), and so on. Furthermore in this way it was possible to have also a first mapping of the websites (*Table 1*).

As an important outcome of this examination, given that the difference in the structure and content of the sites depends to a large extent on the size of the enterprises, on the complexity of their organization, and on the economic activities they carry on, in order to evaluate in the correct way the results of the next web scraping activity it was decided to extract a stratified sample from the SBR, that could take into account the disparities caused by these factors.

Table 1 - Enterprises of the first sample by mapping of their websites

Typology	**Description of the mapping**	**%**
Elementary - illustrative	Simple, with basic information, few pages and basic information	15,9
Intermediate - illustrative	Well-illustrated, with general information, reporting certifications, periodically updated and with exhaustive information	11,2
Advanced – illustrative	Broadly illustrated, it shows the whole business of the company, many sub-menus, all activities detailed, reporting certifications, continuously updated, dynamic	14,0
Enterprise Group	Collector node for other sites belonging to a specific group	22,5
Multimedia	Audio, video, photos, with little other information	0,9
Advertising products and brands	Mainly the qualities and information related to the products are indicated	1,9
Alias	Existing only on Yellow Pages or Blank Pages, business communities, sites directories…	9,3
NO website	It was not possible to find out a web site for these enterprises	24,3

After this preliminary analysis, thanks to the positive results obtained, Istat high level management decided it was time for an in-depth investigation and for this aim to make all the big data projects on enterprises converge in a single management project, that should be *register-based*, in the sense that definitively all statistical, administrative and web information will be ‘catalogued’ in the SBR for multiple purposes, thus assuring consistency and giving a concrete support to the statistical production whose base is the SBR. The register-based approach to the Big Data ensure to proceed in a structured and integrated manner, placing the register at the centre.

In this sense there is a “two-way” information flow, to and from the register. The two parts evidently benefit from each other’s presence: on the one hand, the register acquires the new and earlier information from the web and catalogues it in a coherent way for multiple purposes, enlarging its content, for a concrete support for statistical production; on the other hand, the data of the web, once inside the informational world of the register, are integrated with all the other variables from administrative and statistical sources, and in this way the unstructured data of the web acquire a “structure” and could benefit from all the available statistical classification: the link with the register provides information on the subpopulations of interest and therefore on which is the reference universe of the data; they receive a name, a size dimension, a collocation in the territory, etc. using all the classifications in the register. The wide coverage of the SBR becomes a support platform for Big Data.

For this aim a working group has been put in place, added to the Istat participation to the first ESSnet BigData[3]. The analysis of available resources results in the formation of the work team. It was a difficult task for the necessity of managing connections among separated departments. However it was clear to everyone from the beginning that we had to put together different types of resources, having peculiar professional skills. All the experts have been working in Istat for many years. Some of them are directly involved in web scraping and text mining activities especially from a methodological point of view, others are responsible for carrying out and managing the statistical production processes of the SBR and competent for the thematic aspects of this research. Some IT persons have been involved too, also to deal with the future production of the tools and techniques that are being experimented, in view of their implementation in the regular data production. To underline the presence of the various professional skills that have become necessary, the work team is composed by one top level manager for handling relationships, three methodological experts, three statisticians managing the SBR production process and two IT experts.

A multidisciplinary context means there is high interaction among experts and also integration of different skills and expertise. Experts have been always working contextually, side by side. The ‘thematic’ expert governs the process, raises the problem, indicates its requirements, and validates the results. He/she does not know in advance the contents coming from the web. The ‘methodological’ expert analyses the requirements of the problem and finds the methodological solutions. The IT expert provides the tools and techniques and turns specifications into procedures. And everything happens in a circular and recursive process.

The project started on the 5th February 2018, when the kick-off meeting of the work team took place in Istat. Since it was immediately clear that Big Data need big resources, the first step was to participate in the selections to become part of the first projects that could have taken advantage of the facilities made available in the Istat LabInn.

Inaugurated on the 21st March 2018, the LabInn is one of the tools that Istat has been providing to introduce process and product innovations. The laboratory was created to respond to new statistical information needs and to strengthen the Istat commitment to research, also encouraging the integration among Istat researchers by enhancing the specific skills of those who are included in production departments. It also represents an opportunity to enhance human capital and professional growth, with positive impacts in terms of motivation and growth in job satisfaction.

Within a dedicated physical and technological space, the most innovative ideas are developed that refer to the areas identified as strategic and of particular interest to the Institute, like: the use of new data sources and new data collection acquisition methods; improvement of statistical processes and use of administrative files; innovative outputs; use of new ICT technologies and methodologies. The inauguration of the Laboratory was also an opportunity to illustrate projects that have passed an initial evaluation phase and have been developed by Istat researchers during 2018.

[3] Istat will participate also to the next edition of the ESSnet, in many WPs among which the WP C Implementation – Enterprise Characteristics.

The project about Big Data usage to support the SBR has been one of the first projects to participate in the LabInn.

3. Main steps of the first experimentation and set up of the SBR sample

The applied strategy for the three months[4] experimentation in the LabInn consisted in the following three points, that can be regarded as separated steps only by a theoretical point of view and not by execution:

1) acquisition of the enterprise's web address:
   - from the administrative sources used to update the SBR (available only for 5% of SBR enterprises);
   - downloading information from some proper thematic directory sites;
   - performing batch queries on the search engines by means of the available SBR enterprises identification characteristics (URL retrieval with machine learning techniques);
2) identification of the enterprise – in whatever way the web address has been found, it is considered crucial to proceed with its identification, in order to confirm the exact correspondence between websites and enterprises in the SBR:
   - 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;
   - detection of identification variables from the website (Fiscal Code, VAT Number, Business Name, Address, ...) through the use of *Information Retrieval* techniques through *pattern matching* on strings;
   - comparison with the same information available in the SBR register through *matching techniques* and *string similarity metrics* (Jaro-Winkler, Levenshtein, etc.)
3) extraction and analysis of information:
   - Web Scraping techniques for web data acquisition (with different approaches depending on whether the enterprise web address is available or not in advance);
   - Text Mining techniques (using also Natural Language Processing techniques) for extracting the information to integrate the Business Register (including metadata and *.pdf* file).
   - Machine Learning techniques, for the use of algorithms that simulate a learning process for the construction of predictive models.

The starting point of the project has been the choice of the enterprises in the SBR to involve in the experimentation, considering the subset of register where the web addresses of the enterprises were available from administrative sources.

The register *PagineGialle* (the Italian Yellow Pages) by Consodata S.p.A. is the administrative source that provides each year all the information about URLs and telephone/fax numbers. In this project the most recent data have been used, that is the web addresses recently updated (no earlier than 2016). In this way only 5% of the enterprises in the SBR has a non-empty field for the URL, about 240,000 active units.

As a reference population only a subset of the SBR 2015 has been considered, namely about 1,600,000 active enterprises in the reference year, having the legal form of a corporation or a

[4] From the 21st March 2018 until the end of June 2018.

partnership. To improve precision, a stratified sample of about 100,000 units has been extracted from the reference universe, with proportionate allocation[5].

The experimentation considers a sample of enterprises stratified by size, in terms of employment and turnover, and by economic activity, in order to take into account the differences in the structure and contents of the web sites possibly caused by these factors.

After few tests to find a good partitioning of the enterprises in the least number of homogeneous strata, for the construction of the variable "size of the enterprise", both the number of employees and turnover were taken into account at the same time, thus giving rise to four classes, as showed in *Table 2* or in *Figure 1*. In order to take into account the different economic activities performed by the enterprises, the NACE Rev.2 classification has been used and in particular the codes have been grouped in the high-level SNA/ISIC aggregation A*10/11 [6], actually reduced to nine classes since Agriculture, forestry and fishing activities have been excluded from the experimentation. Finally the 36 sample strata are then created, making use of the three variables, as a combination of them (9x4, economic activities and sizes). The 10% sample thus obtained, representative of the SBR units, consists of 115,751 enterprises.

Table 2 – Size of the enterprises in terms of employees and turnover

Size	Employees		Turnover
MICRO	< 10	and	< 2 mln
SMALL	10 |– 50	and	2 |– 10 mln
MEDIUM	50 |– 250	and	10 |– 50 mln
LARGE	≥ 250	OR	≥ 50 mln

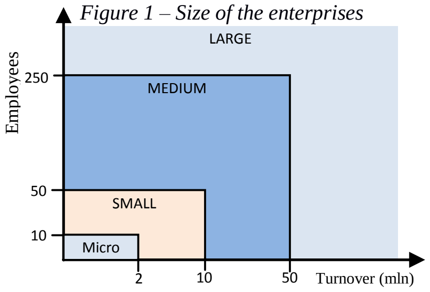


Figure 1 – Size of the enterprises

4. Web Mining Process

The automatic detection of characteristics in website has recently emerged as a very important task in several contexts. A huge amount of information is freely available through websites, and it could, for instance, be used to accomplish statistical surveys. However, the information of interest for the specific task under consideration has to be mined among that huge amount, and this turns out to be a difficult operation in practice. Indeed, the sheer size of the problem makes implausible to non-automatic intervention. The main contributions of this work is the presentation of a practically viable technique to perform automatic detection of enterprise's characteristics by using the information contained in their websites. For this purpose, a complex strategy has been defined and developed. This strategy includes 3 main phases: web scraping, text mining and machine learning.

[5] In this way the sampling fraction in each of the strata is proportional to that of the total population.

[6] 1. Agriculture, forestry and fishing; 2. Manufacturing, mining and quarrying and other industry; 3. Construction; 4. Wholesale and retail trade, transportation and storage, accommodation and food service activities; 5. Information and communication; 6. Financial and insurance activities; 7. Real estate activities; 8. Professional, scientific, technical, administration and support service activities; 9. Public administration, defense, education, human health and social work activities; 10. Other services.

4.1 The Web Scraping phase

In the web scraping phase the procedure extracts the information from each corporate website and saves it to a NoSQL DBMS. In this phase, there are two steps:

- First step:
  The procedure identifies each enterprise on the web and creates an URLs list. In case of available URL from administrative sources the procedure checks whether the URL is valid or not and finds a valid one if it is not valid. In case of non-available URL the procedure uses either the URL Retrieval techniques performing batch queries on the search engines by means of the enterprise identification characteristics, or it directly download information from some proper thematic directory sites.
- Second step:
  The procedure extracts all the information from websites by using web scraping techniques. Besides the text appearing in the pages, it reads also some additional information: such as images, HTML tags, meta-keywords etc.

For further explanation:

In the first step, in case of available URLs from administrative sources, the procedure verifies if the URLs are valid and exist at start. It proceeds with the syntactic validation of the strings, the check of the recurring errors and the authority of the web addresses. Then, in case of non-existing URL it performs a web search using search engines in order to find the most similar URL and computes its probability of correctness by using a machine learning approach.

In case the enterprise web address is not available in advance, the procedure performs batch queries on the search engines by means of the available enterprises identification characteristics. In particular, the name of the enterprise is used as a search string, then a query on a search engine is performed, collecting the first 10 links returned as the result of the query. For each link the probability of correctness is evaluated by using a machine learning approach, hence the link whose probability exceeds a given threshold is accepted as valid.

In the second step, the scraping procedure also reads, besides the text, additional information such as: the attributes of HTML elements, the name of the image files, *.pdf* files, the meta-keywords of the pages etc. Two types of web scraping have been considered: i) generic web scraping; ii) specific web scraping.

Generic web scraping assumes that the structure and the content of a website are not known in advance so the site is scraped and processed in order to collect information of interest. In such cases any specific information will be retrieved in the next phases.

Specific web scraping, instead, deals with the case where both structure and content of the websites to be scraped are well known, so scraping programs have to simulate the behaviour of a user visiting the website and collecting all needed information.

4.2 The text mining phase

The number of web pages available on the Internet has constantly been increasing in the last 25 years, and nowadays a huge amount of data is freely available through this channel. Therefore, the automatic extraction of many statistical information from this source is extremely appealing. Otherwise the amount of data is surely a big problem itself, but it is worsened by its completely not standardized structure and by noisy and not completely homogeneous data.

Therefore to identify the relevant parts of the extracted information a quite articulated procedure has to be developed that requires the use of several steps of text mining.

Text Mining is the branch of Data Mining concerning the process of deriving high-quality information from texts. References can be found for instance in **[8]**. This area underwent considerable improvements in recent years, with a number of concurrent factors contributing to its progress, first of all the continuous expansion of the Internet and the demand for effective automatic search and manipulation strategies. Modern text mining techniques require the integration of natural language processing techniques (see, e.g. **[5]**) with several advanced machine learning techniques. (see, e.g. **[9, 10]**).

Natural language processing is an approach which allows to find meaning of the free text. This is done by using several techniques **[12, 14]** such as:

- Tokenization: cutting string into still useful linguistic units using string splitting (whitespaces) or regular expressions.
- Lemmatization: given a word, its inflectional ending is removed in order to return the word to its basic lemma. This allows to group together the different inflected forms of a word (e.g. plurals of nouns, tenses of verbs, etc.) so they can be analysed as a single item.
- Part-Of-Speech recognition (POS tagging): every word is identified as a particular part of speech (such as: noun, verb, etc.).
- Word embedding: set of language modelling and feature learning techniques in NLP mapping words or phrases from the vocabulary to real numbers vectors.

In general, a very high level model for text analysis includes several text processing tasks, such of these are:

- Language identification:
  this task automatically detects the language(s) present in a document based on the content of the document. It is a key task in the text mining process.

- Information retrieval:
  this task gathers results that are potentially of quality and relevant for the specific needs of the user, which can be:
  - informational: if the user wants to know something about a certain topic;
  - navigational: if the user wants to go to specific page;
  - transactional: if the user wants to perform a certain operation with the mediation of the Web; e.g. access a service, download items, make a purchase
  - mixed: if user behaviour comprises a mixture of the previous points, e.g. look for a good starting point for research on a certain topic ("Bed and Breakfast Milan").

- Information Extraction:
  this task automatically extracts structured information from unstructured and/or semi-structured machine-readable documents. It requires the implementation of methodologies able to support and automate the identification and the organized collection of data and information from a “source”, written in natural language and the development of a semantic interpretation capacity related to the language used and the expressive context and terminology of the domain to which it refers.
  It includes three levels of intervention named entity recognition, coreference resolution and relationship extraction.

- Textual entailment recognition:
  this task decides if the meaning of one text is entailed (can be inferred) from another text, given two text fragments.

- Summarization:
  this task aims to provide a self-contained and internally cohesive text which serves as a selective account of the original.

In some cases, according to the specific needs, the sequential application of some of these tasks allows to obtain directly the requested information. For instance, in order to estimate the presence in website of links or references to the enterprise's social media profiles Information Retrieval techniques have been applied.

In other cases, instead, the problem of automatic detection of enterprise's characteristics needs to be solved as a supervised classification problem. Each record is obtained from the automatic analysis of one enterprise website by Information Extraction and Summarization techniques and the class is the presence or the absence of some phenomenon.

For instance, determining whether an enterprise website offers online job application facilities is an interesting case of this problem **[2, 4]**.

4.3 The Machine learning phase

Engineering a statistical production process which includes Big Data sources could include Machine Learning (ML) techniques to guarantee their automation. Moreover, in order to increase the level of automatization, which reduces the risk related to uncontrolled data sources, it is useful to define a calibration phase of ML algorithms that help set the typical algorithmic parameters required for processing **[3]**. In general, ML algorithms embody the principles of Data Mining aspiring to enables computer systems to learn behaviour from data in order to take autonomous decisions. The choice of which ML algorithm to apply, largely depends on the user's domain knowledge, the desired results and on the performance of computing platform. There are many different kinds of machine learning algorithms for discover patterns in Big Data that lead to actionable insights. At a high level of abstraction, these different algorithms can be classified into three groups based on the way they "learn" about data to make predictions: supervised learning, unsupervised learning, semi-supervised learning.

Supervised learning is based on the availability of a set of labelled records, called training set, that constitutes the source of information to learn a classifier **[3]**.

Unsupervised machine learning is more closely aligned with the idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Finally, many commonly used machine learning algorithms actually fall into the category of semi-supervised learning where only some of the data is labelled.

When the prediction of characteristics in websites is needed, the analytical process involves supervised classification techniques **[2, 4]**. Therefore, to apply the described approach for the automatic detection of characteristics in website, a set of websites is needed for which it is available, or at least attainable, the class labels with respect to the considered categorization.

In this work, a strategy has been applied for classifying a list of websites in order to automatically categorize them on the basis of presence or absence of some phenomenon. To clarify about the former enterprise website online job offer for example, the procedure has been the following: first all the available text is extracted from this list of websites by means of an automatic scraping procedure; then this text is used to prepare the data records, as described in the previous §4.2.

Subsequently, several steps are performed in order to identify and select only the relevant parts of the above information and to exclude the noisy portion as much as possible. This is done by using several natural language processing techniques, until a number of relevant words are obtained, together with n-grams and additional features that can be used to summarize each website. After this, dimensionality reductions techniques have been applied to obtain a set of standardized data records describing the websites. Finally, such records have been classified by using state-of-the-art classification algorithms **[10]**. Among others, Deep Forest **[14]**, Random Forest **[6]** and Support Vector Machine **[7]** have been used in this experimentation.

5. First preliminary results: additional information from web

The identified enterprise websites have been analysed by means of the Big Data analytic techniques, described in the previous section, in order to produce a set of new enterprise-level information linked to the SBR, whose content allows:

- to complete the missing information in certain variables of the Business Register;
- to check some information for variables enclosed in the Business Register;
- to add new information to cover additional variables for the Business Register.

The new information can be classified according to the type of business enterprise characteristics extracted from web. In particular, the following types have been considered:

- structural characters: they are related to the structural features of the enterprises, such as anagraphical characteristics and personal data, business data, dimension, etc. In this experimentations the following ones have been considered: Tax fiscal code, VAT number, business name, company capital, telephone number, email address, certified email address, business enterprise's street address;
- qualitative characters: they concern information not directly measurable but are representative for the enterprise, such as, the short description of business activity, the presence of links or references to the enterprise’s social media profiles, the presence of online job application facilities, the identification of *.pdf* documents concerning financial statements or product catalogues.

The design and development of the prototype that implements the strategy described in the previous section has been realized on the extracted sample of enterprises. These enterprises have been organized in groups according to the approach used for identifying them on the web. In particular, the acquisition of the company’s web address has been realized by source type:

i. URLs available from administrative sources (URLs available)
These URLs are derived from administrative sources or marketing companies and checked by a validation procedure.

ii. URLs obtained from Certified email available in administrative sources (URLs from Certified email)

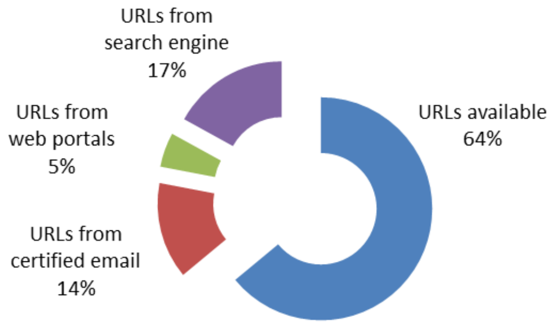
These web addresses are the result of a process of extracting URLs from the Certified email address contained in administrative sources. The procedure extracts the domain of the enterprise website from Certified email and verifies if a correspondent valid URL exists.

iii. URLs obtained from some proper thematic directory sites (URLs from web portals)
These URLs are the result of a specific web scraping process to extract the identification variables of enterprises from some thematic business portals

iv. URLs obtained through the use of search engines in batch (URLs from search engine)
These URLs are the result of a URL Retrieval procedure that performs batch queries on the search engines by means of the available enterprises identification characteristics.

However, for all types of source the procedure verifies the correct identification of URLs, in order to confirm the exact correspondence between websites and enterprises in the Business Register. *Figure 2* shows how the enterprises of the sample achieved the URLs in this experimentation.

Figure 2 - Enterprises of the sample by type of source for the URLs



To all the extracted test, the Information Retrieval techniques have been applied, for the recovery of the relevant textual information contained in it. In particular, through pattern matching on strings. Here in the following, some examples of the results obtained when the retrieved information is compared with the same characters available in the SBR register, through matching techniques and string similarity metrics.

5.1 Additional information: structural characters

The “***Tax code/VAT number***” character has been acquired for about 84% of the enterprises in the sample. Quite all the remaining 16% do not show this information on the website. Since this character allows to uniquely identify the company, it is important to underline that in case of a difference between the tax code/VAT number extracted from the website and the one contained in the register, it should be necessary to investigate better if a demographic event happens that could justify this occurrence.

The “***Company name***” character has been acquired for all enterprises of the sample. It has been compared with that contained in the SBR by means of matching techniques and similarity metrics between strings, and it results to be the same in 70% of the enterprises. Also this character is useful to uniquely identify the company. For the future, the next step will be to check if the ‘discordant’ values found in the web could be found as a new name for the same enterprise in the following release of the SBR, so that the web data could be used to acquire early news on changes in the name.

As regards the “***Enterprise street address***”, a total of 103,000 street addresses have been acquired from the enterprise websites. They have been identified and allocated between administrative headquarters and other local units (*Table 3*). The purpose is to provide the SBR with a new source of information useful for identifying the correct location of administrative offices and other local units of the enterprises, and to correct the under coverage of administrative sources.

The “***E-mail address***” has been acquired from enterprise websites and distinguished into e-mail address and certified e-mail address, for a total of 198,000 email addresses (*Table 4*). This character enriches the SBR with updated information that is useful for surveys’ managers and supervisors, for keeping communication with the enterprises, to enlarge information of business contacts.

Table 3 – Addresses by type of local unit

Type of local unit	%
Addresses related to administrative headquarters	70
Addresses related to other local units	30

Table 4 – E-mail addresses by type of address

Type of address	%
E-mail address	92
Certified e-mail address	8

As regards “***Telephone number***”, a total of 230,000 phone numbers have been acquired from company websites. They have been recognized and distinguished among telephone numbers, mobile phone numbers and fax numbers (*Table 5*). This information is very useful since it is not well covered in the administrative sources used to update the Italian SBR. It is used in the check phases of all the business surveys and it is also useful for SBR staff when a contact with the enterprises is needed.

Table 5 – Phone numbers by type

Type of phone number	%
Phone	58
Mobile phone	12
Fax	30

Concerning “***Company capital***”, the value present on the website has been acquired and recovered in 21% of the enterprises of the sample. This information refers just to those companies that publish the share capital on the website. This type of information is also useful for creating new dimensional class that measures the companies by risk capital.

5.2 Additional information: qualitative characters

The string identified as a “***summary description of the enterprise activity***” has been extracted from the web for all companies. This information has been obtained combining the meta-information (meta-tag) of the website with the results of querying the search engine. The output is useful for the SBR staff to classify the economic activity performed by the enterprises because it represents a descriptive showcase of the website. *Table 6* shows a ‘summary description of activity’ for three enterprises chosen in the sample.

Table 6 – Examples of summary description of activity for three enterprises in the sample

Id Code	Company name	Activity description
XX1	Enterprise 1	Company specialized in calendaring and bending of pipes and profiles
XX2	Enterprise 2	Installation and maintenance of security alarm systems, telecommunications and home automation
XX3	Enterprise 3	Production of precious fabrics to satisfy every customer's need

Concerning the "***Presence on Social Media***", more than 45% of the enterprises of the sample use these new tools for marketing activities. Identifying companies on social media provide a new channel for getting further enterprises information. This information can be considered as an indicator of the use of IT for marketing. *Figure 3* shows the distribution of social media usage.

The "***Online job application facilities***" character identifies enterprises that publish job application facilities on their website. It allows the enterprise to find specialized profiles by using web technologies and provides another useful skill based indicator for classifying the economic activity of companies. *Table 7* shows the distribution of enterprises with online job facilities.

Figure 3. Enterprises by presence on social media

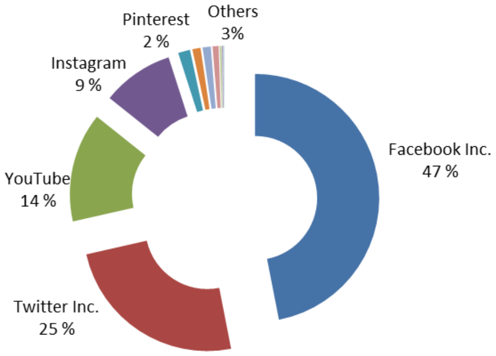


Table 7 – Enterprises by presence/absence of online job facilities

Online job facilities	**%**
Enterprises WITH online job application facilities	15
Enterprises WITHOUT online job application facilities	85

The problem of automatic detection of this character has been solved by using a machine learning approach. A dataset of 12,000 enterprise websites has been considered, with known class label, as training set. In this problem, the class label is the presence/absence of online job facilities. The training set was composed of enterprises respondent to the Italian ICT survey and for them the actual value was known. Finally, the fitted model has been applied on the training set to predict the class for all enterprises in the sample. In particular, the classification algorithms that have been used are: Support Vector Machines, Random Forest and Deep Forest.

The "***presence of .pdf documents***" concerning *financial statements* and *product catalogs* has been also investigated. Some enterprises publish this information on their website. Only these two typologies of documents have been considered, but many other *.pdf* files are available to continue the analysis. The files has been downloaded, then identified by analyzing their contents by a document classification procedure, and finally archived by category.

Table 8 – Presence / absence of .pdf file by typology of document

Type	**%**
Presence of *.pdf* files concerning financial statements and product catalogues on websites	40
of which:	
- *financial statements*	*30*
- *product catalogues*	*70*
Absence of *.pdf* files concerning financial statements and product catalogues on websites	60

Financial statement allows to extract enterprise's characteristics and other relevant information about economic activity, useful also to profile the enterprise, while the product catalogues provides useful information for better classifying the economic activity performed. Furthermore, the analysis of both kind of documents is useful for the Target 3 (New business taxonomies - 'emerging' populations, to be mapped over time), in particular when studying the "Enterprise 4.0" phenomenon.

Table 8 shows the presence/absence of this two kinds of *.pdf* files in the sample websites.

5.3 First preliminary results: the Web Mining process

Table 9 – Enterprises in the sample by identified characters and by source of information for the URL

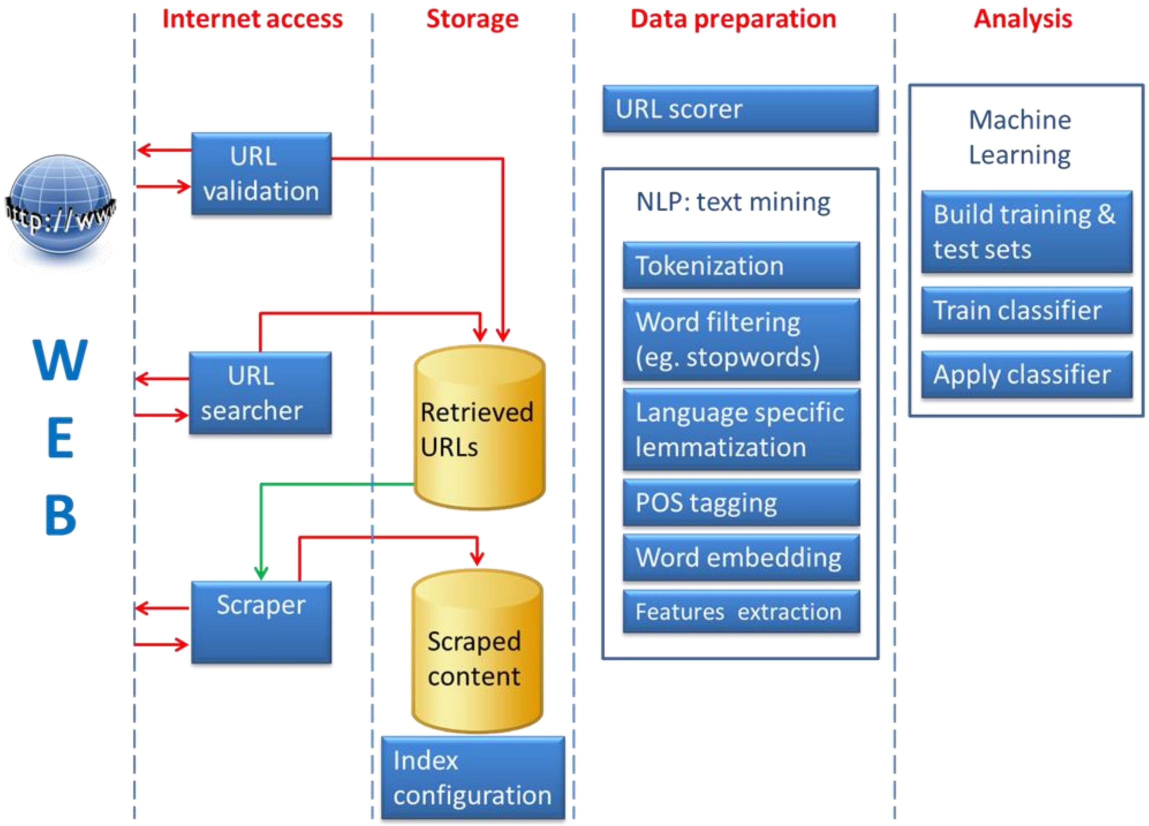
Characters	Total	URLs available from admin source	URLs from certified mails	URLs from web portals	URLs from search engine
Company name	100,000	64 %	14 %	5 %	17 %
Tax code/VAT number	80,440	60 %	9 %	5 %	26 %
Enterprise street address	103,000	70 %	8 %	7 %	15 %
E-mail Address	198,000	60 %	10 %	3 %	27 %
Telephone number	230,000	63 %	8 %	7 %	22 %
Company capital	21,580	40 %	4 %	35 %	21 %
Presence on Social Media	131,830	65 %	9 %	6 %	20 %
Presence of online job application facilities	15,000	52 %	16 %	2 %	30 %

6. Reference Framework Architecture for Web Data Processing

One of the objective of this experience is to investigate whether the combination of web scraping, text mining and machine learning techniques can be used to collect information about enterprises to be used for improving the SBR. In order to do this, a custom version of the generic reference logical architecture already used in the ESSnet Big Data project **[16]** was adopted. As shown in *Figure 4* this reference architecture is made of several building blocks organized into four main layers, namely: "Internet access", "Storage", "Data preparation" and "Analysis". The Web scraping part of the process (data acquisition and subsequent storage) is done in the first two layers ("Internet access" and "Storage"), then the scraped content is processed in the third layer ("Data preparation") by means of text mining techniques in order to transform the unstructured or semi-structured nature of the data in a structured form suitable to be used in the analysis layer by means of machine learning techniques that will produce the final statistical outputs.

The **URL validation** block performs a check about the correctness of a list of URLs (from administrative sources) provided as input. In particular it verifies whether the URLs are syntactically valid and still exist, moreover it checks the recurring errors and the authority of the web addresses. In case of non-existing URL a software performs a web search by using Bing search engine in order to find out the most similar URL and computes its probability of correctness by using a machine learning approach.

Figure 4 – Reference Framework Architecture for Web Data Processing



The objective of the **URL searcher** block is to retrieve a list of websites related to a given enterprise. Usually this list is obtained by querying a search engine on the web using the name of the enterprise as a search term. **[15]** The underlying assumption is that, if an enterprise has an official website, this should be found within the results provided by a search engine. For this block a custom software named UrlSearcher **[17]** was used.

The **Retrieved URLs** block is basically a container of URLs obtained in the URL searcher and URL validation blocks, it can be implemented in different ways, ranging from a *txt* file to a relational DBMS.

The **Scraper** block is responsible for acquiring the content available on each URL in the list of URLs provided as input. It can have additional features such as URL filtering (if a list of URL to filter out is provided) and is usually configurable by setting different parameters such as the level of scraping (e.g. just the homepage or the homepage plus a first level of links from that, etc.). For this block a custom software named RootJuice **[18]** was used. It is worthwhile to mention that RootJuice organizes the content scraped from web pages in a way that reflects the structure of the HTML document, this means that it will be possible to store in the storage platform not just plain unstructured text but a list of HTML tags and attributes (eg. title, metatagDescription, etc.) each of them containing the related text present in the page. In addiction other software were used in order to download enterprise related information from some thematic directory sites.

The **Scraped content** block is a container of the content scraped by the Scraper block. Usually it is necessary to implement this block by using Big Data technological solutions due to the fact that the amount of information could be huge and mainly consisting of unstructured data. In this work the software that implements the Scraper block is Apache Solr **[22]**, an open source enterprise search platform that is also a NoSQL DB. The scraped content is loaded into Solr by a custom software named SolrTSVImporter **[19]**.

The **Index configuration** block represents a strategy of indexing the scraped data stored into the Scraped content block. In a Big Data context, the huge amount of data that can be stored may make data indexing a mandatory step, in order to easily retrieve information in subsequent phases. This block is normally included in the storage platform.

The **URL scorer** block is used to assign a score to an URL on the basis of some predefined parameters such as the presence of some features of interest inside the URL's content. Given a list of URLs related to an enterprise, this block can be used alone or in conjunction with other blocks in order to prepare the decision strategy to identify the most probable official URL for a given enterprise. For this block a couple of custom software named UrlScorer **[20]** and UrlMatchTableGenerator **[21]** were used.

The **NLP: text mining** block contains several sub-blocks. The Natural Language Processing (NLP) is the ability of a software to understand human language, in this work the focus is on textual data, this implies that text mining techniques must be used in order to proficiently process the scraped content and allow the feature engineering that is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

The **Tokenization** block processes the textual content of the scraped resources by transforming it in a text that becomes input for further processing such as parsing and text mining or for analysis blocks.

The **Word filtering** block is used to filter out some words/tokens (if a list of words to be filtered out is provided) from the scraped textual content or to enrich it with a list of go words.

The **Language specific lemmatization** block lemmatizes the tokens found in the scraped textual content in order to reduce the number of textual elements to be analysed. Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.

The **POS tagging** block is responsible for the Part-Of-Speech recognition, this means that each word contained in the scraped text is classified as a specific part of speech (e.g. noun, verb, adjective, etc.).

The **Word embedding** block includes a set of language modelling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers.

The **Feature extraction** block is responsible for localizing and retrieving from a scraped resource a set of predefined features of interest (e.g.: addresses, telephone numbers, names, VAT codes, etc.) by searching for a specific regular expressions, matching sentences etc. Usually it is implemented in a SW program.

The **Machine learning** block (and its sub-blocks) produces the final statistical outputs by using one (or more) learner(s). In this work a supervised approach has been adopted. In particular the following learners have been used: SVM, Random forest, Deep forest.

7. Conclusions: lessons learnt and future work

Internet as data source represents new opportunities and challenges for Official Statistics that should incorporate all innovative potential data sources as much as possible into their conceptual design. An increasing number of Statistical Institutes, including Istat, are indeed experimenting the use of new sources of data, also known as Big Data, in order to produce the same or new statistical information in a multisource environment, more efficiently and with higher levels of quality.

In this context some experiments are being conducted at Istat which will be fully operational and enter the production process in the medium term. There are important tasks yet to be faced with the use of Big Data and in general for Statistical Institutes. The main challenge for using Big Data is to move from experimentation to production. This step involves various aspects ranging from respect for privacy to the need to acquire new infrastructures (methodological, technological, organizational ones), as well as new skills. Istat is working in this direction and the exchange of experiences on an international level is fundamental.

In recent decades, the combined use of data from different sources for statistical purposes has become a consolidated practice. Along with the use of administrative data for setting-up statistical registers, the experimentation of the use of Big Data has been started for some time in order to update, add and validate information in the SBR. The most used source for finding Big Data is the web. The vast amount of information available presents new opportunities, but at the same time new challenges for data integration experts, given the structural difference between administrative and internet data.

Definitely in the administrative sources the identification of the unit is sure. In addition, the integration of their data is simple, as they use a common identification code (tax code, VAT number). Through a consolidated process of integration, an ID code for legal units and possibly an ID code for enterprises are assigned.

Without any doubts instead, Big Data present serious risks. Some of them are evident, like the difficulty to manage rapidly growing volumes of data resulting in a high consumption of computing and storage resources; moreover there are technical limits to solve, like the long run time necessary for the crawler to get the entire content and the restrictions caused by the security barriers inside the website preventing automatic access. Some others are statistical problems, such as: the difficulty of certifying the quality of information and the data reliability; how to attribute information with certainty to an 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. A point of strength is that the web is an independent source of data, while all the other sources, administrative and statistical, can be considered in some way linked to each other and influencing each other. The picture provided by the Big Data may be a representation closer to reality: this is how the enterprise sees itself and how it wants to present itself to the outside.

Some issues like the updating and maintaining strategies are questions that still need a proper answer, like the matter on when and how it would be necessary to repeat the extraction of the same data from the web. Although very expensive, maybe a cyclic monitoring to capture the differences could be useful, anyway it must still be understood when differences can be considered significant as true changes, and moreover how to get the "date" in which the change takes place, crucial for the SBR metadata. These are issues that have not even been touched upon in this experimentation.

Indeed there are many difficulties related to the use of these new sources. The Big Data change the way we collect, analyse and integrate data. The added value lies exactly in the information that is hidden in the data, and in the proactive use of the data, namely *data-driven*, that is, reading the data and using them as a starting point to create a strategy. Therefore to integrate the Big Data analytics tools in the context of current production processes is actually a difficult task. This means to combine *data-driven* processes – based on *input* data that do not come from sources that are specific for statistical purposes, as they are not homogeneous, not structured or semi-structured, and not stable over time – with processes based on an *output-oriented* approach, since in the context of official statistics, production processes are constructed with a view to obtaining statistical outputs.

The challenge in this project is to integrate these two logics according to a *register-based* approach. While the step taken so far is to extract insight, useful information for the study of a phenomenon through the combined use of data mining and machine learning techniques using mainly methodological/engineering skills, now in this new context it is the thematic expert to be crucial, to guide and closely follow the integration of web data, large and unstructured, with those well-known and structured of the register.

It is clear that in order to be able to combine the two types of process in the best way, a multidisciplinary investment is necessary, specifically the specialized, thematic, statistical, algorithmic and technological competences should be enabled to work together, sharing their high level skills. For this aim a good training is essential, it is important that the various specialists are able to understand the Big Data analytics tools, obtaining the skills on techniques and tools compatible with the technologies and production processes, since complexity requires that they be tackled by collaborating on a shared competence base.

In this experimentation the different professional skills have created synergy, there were step-by-step advances obtained thanks to these interactions and to the critical analysis of results carried out together, since the validation of the extracted data is the responsibility of the thematic competence. As a result a prototype procedure has been produced that validated the standard: the experience in the LabInn has been the starting point for achieving more and more timely information, obtained from unofficial sources, and therefore producing new experimental statistics expanding the SBR, in addition to the traditional business statistics.

8. References

[1] https://www.istat.it/en/experimental-statistics/experiments-on-big-data

Web Mining Process

[2] Barcaroli G., Bianchi G., Bruni R., Nurra A., Salamone S., Scarnò M. Machine learning and statistical inference: the case of Istat survey on ICT. Proceeding of 48th scientific meeting of the Italian Statistical Society SIS 2016, Salerno, Italy. Editors: Pratesi M. and Pena C. ISBN: 9788861970618.

[3] G. Bianchi, R. Bruni, Effective Classification using Binarization and Statistical Analysis. IEEE Trans. on Knowledge and Data Engineering Vol. 27, 2015.

[4] G. Bianchi, R. Bruni, F. Scalfati, Identifying e-Commerce in Enterprises by means of Text Mining and Classification algorithms, Mathematical Problems in Engineering, in press, https://www.hindawi.com/journals/mpe/aip/7231920/, 2018.

[5] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python. O'Reilly Media, 2009.

[6] L. Breiman. Random Forests. Machine Learning. 45 (1): 532 (2001).

[7] C.-C. Chang and C.-J. Lin, Training *v*-support vector classifiers: Theory and algorithms, Neural Computation, 13(9), 2119-2147 (2001).

[8] R. Feldman, J. Sanger, The Text Mining Handbook. Cambridge University Press,2006.

[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, 2002.

[10] W. Klosgen, J.M. Zytkow (eds), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 2002.

[11] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830, 2011.

[12] R. Rehurek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Malta, 2010.

[13] H. Schmid, Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland, 1995.

[14] Z.-H. Zhou and J. Feng. Deep forest: Towards an alternative to deep neural networks. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pages 3553–3559, Melbourne, Australia, 2018

Reference Framework Architecture for Web Data Processing

[15] Barcaroli G., Scannapieco M., Summa D. (2016). On the use of Internet as a data source for official statistics: a strategy for identifying enterprises on the web. Available at: https://www.istat.it/it/files//2018/06/a4_RIEDS-2016.pdf

[16] ESSnet Big Data project, WP2 deliverable 2.4. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/ee/Wp2_Del2_4.pdf

[17] UrlSearcher - https://github.com/SummaIstat/UrlSearcher

[18] RootJuice - https://github.com/SummaIstat/RootJuice

[19] SolrTSVImporter - https://github.com/SummaIstat/SolrTSVImporter

[20] UrlScorer - https://github.com/SummaIstat/UrlScorer

[21] UrlMatchTableGenerator - https://github.com/SummaIstat/UrlMatchTableGenerator

[22] Apache Solr - http://lucene.apache.org/solr/