ESSnet Trusted Smart Statistcs – Web Intelligence Network

Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 4
- Methodology and Quality

Deliverable 4.1: Minimal guidelines and recommendations for implementation

Revised version, 23.09.2021

Prepared by:

1. **WP leader: Alexander Kowarik (STAT, Austria, alexander.kowarik@statistik.gv.at)**
2. Piet Daas (CBS, The Netherlands)
3. Mauro Bruno (ISTAT, Italy)
4. Magdalena Six (STAT, Austria)
5. Olav ten Bosch (CBS, The Netherlands)
6. Giuseppina Ruocco (ISTAT, Italy)
7. Claire de Maricourt (DARES, France)
8. Valentin Chavdarov (BSI, Bulgaria)

Funded by
the European Union

This deliverable was funded by the European Union.

Funded by
the European Union

Table of Contents

Funded by the European Union

1. Introduction

This document was prepared by the work package 4 "Methodology and Quality" within the ESSnet project Web Intelligence Network (WIN). It is heavily based on previous work conducted during the projects ESSnet Big Data and ESSnet Big Data 2018-2020.

The goal of this document is to present a concise yet as exhaustive as possible overview of the knowledge gained within the ESS until the beginning of the WIN project and to adapt existing information (in wording) to the data source "web". Newcomers to the topic should be able to use this document as a concise on boarding document to be up-to-speed with colleagues that worked in previous projects.

The document covers important aspects of methodology, quality and architecture with a chapter for each of these topics. Since methodology and quality can be interpreted as two sides of the same coin, also the two chapters in this document cover partly overlapping topics.

Funded by
the European Union

2. Methodology of using web scraped data

Web scraping – the automatic collection of data on the Internet – is being used by official statistics organizations, e.g., National Statistical Institutes (NSIs), for a number of reasons. The most often mentioned reasons are to reduce the response burden, to speed up statistics, to derive new indicators, to explore and expand background variables and to characterize (sub) populations [ten Bosch et al., 2018]. A considerable number of studies have been performed and have proven that data collected from the web is a valuable source of information. Those studies have shown that web data can be used to collect specific information, such as prices [Cavallo, 2018] and job vacancies [Rengers et al., 2020], to study the dynamics of a phenomenon before designing a new costly statistical production chain [Barcalori et al., 2016a], to supplement administrative sources and metadata systems [Barcalori et al., 2016b], to replace survey questions [Kowarek et al., 2020] and to supplement and potentially replace already existing statistics [Daas and van der Doef, 2020]. There are probably more possible applications. Technical and legal aspects of web scraping are crucial but manageable; see Condron et al. [2019] and ten Bosch et al. [2018] for more details.

However, the main challenge in using web scraped data for official statistics is of a methodological nature. Where survey variables are designed by an NSI and administrative sources are generally well-defined and well-structured, data extracted from the web is neither under control of the NSI and is often not well-defined or well-structured. In this chapter, an overview is given of the state-of-art of using web scraped data by NSIs, the solutions developed and of the most important methodological challenges remaining. For web data those challenges are: i) combining web data with other data sources, ii) measuring the concept of interest and iii) dealing with population differences and dynamics [Daas et al., 2020a]. For web data, when text or images are being used, methods enabling the extraction of information from these types of data need to be applied. These have been developed by computer scientists and alike. An overview of those methods can be found in chapter 2 of Daas et al. [2020b].

2.1. Obtaining and using web data

Data from the web can be collected by scrapers, API's or sent directly by the owner to the NSI. Web scrapers are programs that search the web and are able to extract data from individual web pages or download complete web pages, web sites, documents and images. Advantage of downloading complete files is that they can be processed and studied after the scraping has been performed. An API is an Application Programming Interface [Singrodia et al., 2019] which is a software intermediary that allows two applications to communicate with each other. This has the advantage that, if an API is available, a program can be written that directly obtains the required information for a particular application in a predefined format. Web data obtained via the owner may seem advantages but has the risk that unknown adjustments or selections may have been made. The NSI may have less control in such cases.

Data collected from the web can be used in various ways. An interesting distinction is the one made between specific and generic scraping [Barcolari et al., 2016a; ten Bosch et al., 2018]. Specific scraping focusses on collecting 'small amounts' of information. Collecting prices from web sites for (a limited number of) products is a good example of such a task. Here, web pages are visited and when a product searched for is found on the web page, its description and price are collected (and likely nothing more). At the other end of the spectrum, generic scraping typically, collects complete web pages or may even collect a large number of web pages per

Funded by
the European Union

domain. As can be expected, generically scraped data is usually processed and analyzed after the raw documents have been obtained.

Another distinction focusses on the direct and indirect use of web data. In the case of direct use, the web page contains exactly the information needed by the NSI. Extracting prices from web pages is a good example of such use, as the price of a particular product is exactly what the NSI is searching for. In the case of indirect use, the web page data needs to be processed (in one way or another) to enable determining the concept of interest, usually via a model. A typical example of this is the use of web site texts to determine if a company is innovative or not [Daas and van der Doef, 2020]. The direct or indirect use is related to the availability of the concept that the researcher intends to measure (more in section 2.2.2).

2.2. Methodological challenges

2.2.1. Combining web data and statistical units

Producing statistics usually starts with defining a target population, such as all businesses in the Business Register, from which specific information will be collected. When using web data as the source of input (one of) the challenge(s) is in knowing which web site belongs to which statistical unit [van Delden et al., 2019]. This, in essence, is knowing which type of unit is associated with which URL (a location on the web). If you are lucky, such a relation is already included in the Business Register or has been determined by the Chamber of Commerce in your country. However, this information may not be complete or may be absent. If thatis the case, that relation needs to be established which is not an easy task [van Delden et al., 2019], especially for small statistical business units [Daas and van der Doef, 2020]. The link between a business unit and a URL (or a domain name) can be fairly dynamic, as companies sometimes stop their activity and web domains become inactive or switch ownership. One also has to realize that not all companies have a web site. The reader also needs to be aware that when different ‘units’ of observation are studied, such as products or job vacancies, similar issues may occur as one has to link those to the statistical ‘unit’ of interest. In all these cases, it’s better to use the term object here, as it applies to URLs, products and (job) vacancies.

To establish the link between a statistical object and an object extracted from the web, NSIs can either use an intermediate data source that includes a relation or try to find it out by themselves. Advantage of the first method is that itis usually much faster. However, in practice, sometimes both methods need to be applied to make sure that the information needed is as complete as possible. How this is done at various NSIs is described below by three examples: i) finding URLs, ii) finding product prices and iii) finding job vacancies.

URLs

In case of establishing the relation between a business unit and a URL (a domain name), this can -in principle- be derived via Domain Name Server (DNS) registration data. A DNS essentially links an IP-address (the unique number of an online server) to a domain-name. Since a domain-name needs to be created and registered prior to use, information on the organization or person that has registered a domain name also needs to be looked up. One can do this up on the web via a WHOis service, but this often only indicates the organization that is the owner of the web server where the URLs resides and not the business or person that owns the web site to which the URL refers. Alternatively, an NSI could visit the URLs, collect its web pages and try to derive the business from the data available. This is a considerable amount of work. There

Funded by
the European Union

are, however, commercial organizations that are able to provide this information (see van Delden et al. [2019] for an example). Another option is trying to find the corresponding URL, starting from business information via a so-called URL-search approach [van Delden et al., 2019]. A difficulty with all these approaches is trying to determine the exact relation between the type of statistical unit (e.g. enterprise, enterprise group, other, …) and the URL [Cordon et al., 2019]. This is a challenging task, as there are companies with multiple web pages and there are web pages associated with different types of statistical business units. On top of this, web pages can also be associated with other types of organizations and individuals.

Product prices

For products, identification is possible by looking at a unique article number or the descriptive texts. For scanner data this number is, for example, included in the data provided by supermarkets [van der Grient and de Haan, 2010]. For web data and for scanner data, one can study the product description. In case of web data one can, if available, also check the product picture; at the moment this is still a manual task. The biggest challenge here is the decision that needs to be made when a product is renewed. Here, usually, a domain expert checks if a product is just rebranded or has actually been replaced and that an alternative product needs to be selected. The substitution of products is important, as this may affect price development and hence the price index over time [Cavallo, 2017; van der Grient and de Haan, 2010].

Job advertisements

In case of job advertisements, different approaches have been used. Initially, the relation with the business unit to which the vacancy applied was studied. It was expected that the vacancy text contained that information, but in practice it was not the case for a considerable number of job offers [Stateva et al., 2020; de Mooij et al., 2020]. For example, vacancy texts may lack that information because an intermediate organization is involved in the application and selection procedure. Another disturbing fact is the finding that not every online job vacancy advertisement actually corresponds to an available job in a business [Stateva et al., 2020] (more on that in section 2.2.3). These observations make it challenging to use job vacancy data for official statistics via a unit oriented approach.

2.2.2. Concept measured.

Concepts are intricate real-world phenomena that are often not completely measured by a single variable or a single source. Moreover, data sources usually measure an aspect of a concept from a certain perspective. This holds for survey, admin and Big data. However, when data is collected by a questionnaire, a statistician has the advantage that he/she is in control on how the concept that he/she wants to determine is measured. Standard procedure here is to define questions, determine the sequence in which those questions are asked and finally test the questionnaire. Following these steps, a statistician is usually pretty confident that the person (or company) surveyed understands the questions and is able to answer them correctly. This procedure has the advantage that –it can be reasonably assumed- the values collected will very likely correspond to those intended to be obtained. In other words, when this procedure is followed, the statistician is usually quite confident that the concept intended to be measured is actually being measured; this is known as concept validity. However, one needs to realize that the data obtained is the result of an operationalization of the concept in the form of a questionnaire and not the “absolute or ground truth”.

Funded by
the European Union

When the data is given, for instance when using web data or administrative data, one cannot be that confident [Bakker, 2011]. In this case, the statistician either looks for a variable similar to the one he/she wants to measure or tries to develop an approach, usually via a model, to construct a variable close to the one intended to be measured. In either case, be it via direct or indirect measurement, it can be challenging -without contacting the owner of the website- to be absolutely sure that the envisioned concept is correctly extracted from the data available. For indirectly, i.e. derived, variables this question is all the more difficult to answer (more on that below). When comparing concepts between sources, so-called harmonization methods can be used to achieve and improve the comparability of the data in these sources [Bakker, 2011, section 2.2].

Web sites

When dealing with web sites, the simplest situation is extracting the value of a concept directly available from a web page. Product prices are a good example and are discussed in the next section. The situation is much more challenging when a concept is derived from the data (or text) available on a web page. An example that illustrates this challenge is determining the innovative character of a company, based on the text available on their web site (see Daas and van der Doef [2020]).

In this study a model was developed, based on a classified training set of companies (innovative or not) and the text extracted from their web sites. The research started with the response of the Community Innovation Survey (CIS) to obtain examples of innovative and non-innovative companies. Next, the corresponding web site of each company was determined and the text on the main page was extracted. This enabled the development of a (logistic regression) model based on the association between the words occurring in the text and a number of other features (such as the language). By testing the model on a non-included part of the training set, the test set, the model developed was found to correctly determine the innovative and non-innovative character of a company with an accuracy of 88% [Daas and van der Doef, 2020]. But, what are the hints that the concept intended to be measured (innovation) was actually being measured? Following results suggests this is the case:

1. The words with the highest weights included in the model were, often, logically related to the innovative character of a company. The ten words with the highest weights were: ‘com’, ‘system’, ‘inspiration’, ‘data’, ‘technology’,’ do’, ‘agenda’, ‘analysis’, ‘proud’ and ‘check’. For many words this really makes sense but what about ‘agenda’? It is difficult to exactly understand why this word is included here. For the words with the lowest (negative) weights, this is even more difficult as it is to be expected that there are many different types of non-innovative companies.
2. When the probabilities of being innovative for classified companies were plotted a clear U-shape emerged. This demonstrated an obvious distinction between two classes (innovative and non-innovative) and indicated a clear separation of both.
3. The number of large innovative companies (companies with 10 or more working persons) as determined by the CIS survey could almost be exactly replicated by the web text based approach. The CIS survey reported 19.916 ± 960 and the web text based approach found 19.276 ± 190 companies.
4. When studying the number of innovative companies at the municipality level, it was found that municipalities where a University or a University of Applied Science is located had relative large numbers of innovative companies compared to (similar sized) municipalities that did not.

Funded by
the European Union

5. Detailed data of the city of Amsterdam, at the zip code 4-level, revealed a clear association between zip-codes with large numbers of innovative companies and the occurrence of so-called startup incubators (organizations that stimulate innovation) in those areas.
6. Studies performed in Germany [Kinne and Lenz, 2019] and Flanders, the Dutch speaking upper part of Belgium, [Reusens, 2021], demonstrated that a similar text-based approach can be used to detect innovation in those countries. As study performed in cooperation with Statistics Netherlands in Sweden, however, they did not result in a positive finding [Daas and van der Doef, 2021; p. 21].

Clearly these points indicate that the concept of innovation or something very similar is being measured. However, the majority of the indications are association based and not all of them fully support this association. This suggests that the claim that the concept of innovation is being measured is actually debatable. Interestingly, in a study also performed at Statistics Netherlands, a web text-based approach was used to detect platform economy web sites. Here, the model was developed to pre-screen websites with the aim to identify (web sites of) companies potentially active in that particular area of the economy [Daas and de Wolf, 2021]. After developing and applying the model to all web sites associated with a company in the Dutch Business register, a list of companies with a probability of having a platform economy website was obtained. Next, a platform economy questionnaire was send to companies with a high probability (> 0.8) of having a platform economy web site. The response of the survey confirmed the fact that the concept of 'platform economy' was measured as it demonstrated that companies with a website with a probability of 0.93 and higher were all platform economy companies [Cakim, 2020].

From this, it is clear that it is challenging to determine -with absolute certainty- that a particular concept is measured when web data is used in an indirect way. Studying the causal relation between web data and the concept measured might be a good way to start shedding light on this problem [Pearl, 2009]. In addition, information from other data sources, such as governmental funding, patent and crowd sourcing data, could be used to determine the reliability of the web derived findings.

Product prices

When product prices are collected the concept measured is the price of a particular product. Here, it is essential to identify this product correctly and extract the corresponding price from the web page. When the product can be done, the prices can be used as long as i) the price on the web is comparable to the price of the same product sold offline [Cavallo, 2017], ii) it can be determined if the price shown is with or without VAT (and more) and iii) the numbers are not embedded in a picture. For the latter, Optical Character Recognition based approaches could be applied to extract them.

Job vacancies

For job vacancies, the challenge at the concept level is in determining if the advert scraped actually corresponds to a job vacancy as defined by Eurostat [2021a]. This is not always the case [Stateva, 2020]. Some of the adverts appear to be vacancies but are actually not; they are referred to as non-existent vacancies or 'ghost' vacancies [Swier et al., 2016]. In other cases, adverts where found to be ads of training courses and blog posts [Eurostat, 2021b]. After

Funded by
the European Union

language detection, these adverts can be identified via a combination of simple heuristics and machine learning with a precision of 99%.

2.2.3. Population covered

When web data is being used, the part of the target population included in the data collected plays a very import role. The ideal situation occurs when the target population is (nearly) completely included. Here, still corrections methods (may) need to be applied but the basis is very good. When the data collected does not include the complete target population, one -somehow- needs to correct for the selectivity of the population included in the data collected [Daas et al., 2020b; section 6.4]. When web data is used in combination with another source, such as survey data, that source can be used to correct for that difference [Daas et al., 2020b, section 2.5]. In all situations, under- and over-coverage (incl. duplications) and the dynamics of the population are very important issues.

Web sites

The data scraped from the web may only partly cover the population. Sometimes the data may even contain information from mixed populations; such as data of persons and companies. It is therefore essential to only select the data from the units included in the target population. The most convenient way is linking the data, via unique identifiers or a combination of variables (such as address information), to a register containing that information for the target population. However, not all data collected from the web contains enough information to enable reliable linking (see section 2.2.1 for more details). If that is not possible, one can decide to remove as much of the records as possible that obviously do not belong to the target population, to reduce linking errors. Next, one should remove duplicates and specifically check for under- and overcoverage. The job vacancy topic in this section provides more details on those steps.

Product prices

When prices are collected from the web usually a 'shopping basket' of products is used for which prices need to be collected. Based on the product description or unique code, those products are selected and the corresponding price is extracted. This, in principle, enables the calculation of the CPI from the data obtained. However, it is essential to weight the product prices and deal with the dynamics of prices on web sites. A paper that discusses the methodology developed in the 'Billion Prices project' of MIT, a project aimed to replicate the official CPI with web scraped data, illustrates the importance of this [Cavallo and Rigobon, 2016]. To enable the creation of a series that co-moves with the official CPI, the MIT researchers carefully selected product categories and retailers from which data was collected. Preferably, web sites of large multichannel retailers in a country that sell both online and offline were selected. Categories of goods, that are part of the official consumer price index baskets and *for which consumer expenditure weights were available*, were scraped. The latter was essential to create a similar CPI, since the weights could not be obtained from the web [Cavallo and Rigobon, 2016]. From this description, it's clear that apart from the prices of a representative set of products, obtaining the correct weights is essential. At an NSI, those weights are determined by the CPI-experts.

Dealing with the dynamics of price behavior is another essential issue for users of web price data [Powell et al., 2018]. By focusing on aggregated data for groups of similar products, this

Funded by
the European Union

effect can be reduced as much as possible. There are groups of products, however, of which the prices are more volatile, such as alcoholic drinks [Powell et al., 2018]. These are often caused by promotions. Time series models are used to deal with these changes as much as possible and survey data was used to calibrate the series [Powell et al., 2018]; e.g. to determine weights. Another approach to reduce the volatility of the CPI series when using web data (or other secondary sources) is using another type of index [Chessa, 2016].

Job vacancies

When job vacancies are collected from the web, a number of population related issues emerge. The first ones are concept related and are discussed in section 2.2.2. After these 'adds' have been removed, the collected set of adverts has to be deduplicated as they are usually collected from multiple sites that may display the same job adverts. A number of up to 20% has been reported for the number of duplicates [Eurostat, 2021b]. After identifying the adverts and comparing the texts, the duplicates are removed. It has also been suggested that methods are needed to deal with adverts containing more than one vacancy [Swier, 2017]. However, from the research performed in this area. it has become clear that not all jobs are advertised online and, hence, that the complete set of all jobs adverts does not represent the complete target population [Beręsewicz and Pater, 2021]. This selectivity of online adverts is strongly related to the linkability of that data with company data. Studies have been performed on inferring the number of job vacancies from online job adverts by attempting to correct for this selectivity. Here, the differences in the statistical unit and coverage have been taken into account, as well as possible existing auxiliary information and the possibility of using model based estimates or Bayesian inference methods. Usually an underestimation of the official number of vacancies is found [Beręsewicz and Pater, 2021; De Mooij et al., 2020; Condron et al., 2019]. There is an exception however. In Bulgaria, the number of online job vacancies was found to be much higher than the number obtained by the job vacancy survey. This was caused by many small and medium companies using accountants to keep their books in order. Since those accountants were usually also responsible for filling in the job vacancy survey, they often had no idea whether the company had vacant jobs or not and subsequently indicated zero (0) job vacancies. Hence, in Bulgaria, the number of online job vacancies is expected to be much closer to reality.

In general, it was concluded that the corrected online job vacancy data showed interesting trends (for some countries) but did not exactly reproduce the official statistics. Another approach to correct for the selectivity of this source is using capture-recapture methods in combination with a thorough cleaning method [Beręsewicz et al., 2021]. This (still) resulted in an underestimation of the number of job vacancies by 10-15%. Main causes mentioned were non-response and under-reporting of job vacancies online.

2.3. Conclusion

In this chapter an overview is provided of the methodological challenges when using web data for official statistics. The most important topics are discussed and references have been included to papers in which these topics are discussed in more detail.

Funded by
the European Union

3. Quality

This chapter on quality is meant to give an overview on relevant quality aspects and applicable guidelines for the work with web data. The main input is the deliverable K3 "Quality Guidelines for the Acquisition and Usage of Big Data" [Quaresma et al., 2020] from the ESSnet Big Data, which covered all data source classes present in that project.

The chapter is structured along two phases of the statistical production process, the input phase and the throughput phase. The throughput phase refers to two different processes and is split into two parts. The first part of the throughput phase is dedicated to deriving- so-called - statistical data from raw (or pre-processed) data. Whereas the second part deals with usage of these data to produce statistical output.

3.1. Input phase

In this phase, the most important steps are to

- identify new data sources (new web sites, APIs, etc.),
- contact/negotiate data owners,
- define workflows,
- prepare IT-infrastructure and to,
- acquire or record data and metadata.

3.1.1. General aspects about the new data source

The following questions about the statistical production should be considered:

- What could be the exact usages of online data? Which existing statistical outputs could benefit from these new types of data?
- What are the implications of using online data? What trade-offs are to be made? For example, we may obtain more granular indicators but with an unknown coverage bias.

The following questions about risks beside the statistical production should be considered:

- Could the outputs involved become vulnerable (e.g., strongly influenced by shifts in usage patterns of online tools)?
- Could there be any consequences to the reputation and the trustworthiness of the statistical office?
- Which legal aspects have to be considered?
- Are there socio-political aspects to be considered?
- What risk mitigation strategies can the statistical office develop?

Data access has to be taken into account: Long-term access has to be guaranteed.

3.1.2. Acquisition and testing of test data

The acquired test data has to be tested thoroughly.

Funded by
the European Union

During this stage it should be clarified:

- which (main) processes - technical and statistical - are necessary to use the new data source,
- whether the skills necessary to process the data are available in the statistical office,
- whether the available tools of the statistical office can adequately deal with the data.Particular attention should be given to the IT-issues of storage and computing capacities.

The forensic investigation of the test data involves:

- all the known steps involved in data cleaning,
- the production of aggregate statistics and the production of outputs,
- the linking of the test data with existing data.

3.1.3. General requirements for acquiring the data (more specific: for web scraping)

- Ensure that each data set will have a corresponding metadata set. Use a unified format for data and metadata storage.
- When collecting the data, ensure that there are reliable attributes that can be used to link to other data (e.g., geolocation, NACE, etc.).
- If possible, allow to access the raw data with the unified interface, i.e. the same name of fields for the specific dimension, e.g. company_id, NACE.
- If there are any methodological differences in the interpretation of the same dimension, e.g. job vacancy vs. job offer, please use the metadata.
- Ensure that all data is stored in a secure way and try to create different groups of users, e.g. external users vs. internal users to allow limited access to the data.
- Try to estimate the target population size, if possible, and use metadata to store this information.
- Use similar classifications, if possible, or at least create the transition key to encode/decode the list of possible values from one data source to another, i.e. level of education, recode lower secondary and upper secondary to secondary.
- Store the data in machine readable format, which can be processed by the computer. It means that the data must be collected in the column or raw two dimensional tables, e.g. ID; dim1; dim2; dim3; value1; value2.
- If possible, allow to access raw data in standard formats like JSON or CSV, to be easily loaded into most common data science environments.
- Replication and possibility of reproducing the data set for other purposes is one of the key issues with the framework presented in this document. Therefore, please use the most common unified formats to store and access this information.

3.1.4. ESS-web-scraping policy

For web scraping, follow the document “ESS web-scraping policy” prepared by ESSnet Big Data WPC [Condron et al., 2019].

This document provides the following Principles and Practices:

Principles

Funded by
the European Union

Web scraping will be performed in adherence with the principles of the European Statistics Code of Practice, and in compliance with intellectual property legislation (national copyright laws and the Database directive) (6).

The members of the ESS should use web scraped data solely for statistical purposes as laid down in regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics and the applicable national statistical legislation.

The principles guiding web scraping activities should be to maximise the benefits while minimising any burden, risks or potential impacts arising from these activities.

To this end, members of the ESS should:

- seek to minimise burden on the website owners;
- identify themselves to the scraped website;
- protect all personal data according to the GDPR;
- abide by all applicable EU and national legislation;
- respect the website scraping policies, being in agreement with the statistical principles laid down in regulation 223/2009 on European statistics;
- scrape for the purpose of creating new statistical information from the data;
- ensure transparency concerning the tools, and the methods and processes that are used for web scraping as well as the methods used for producing the relevant statistics;
- ensure that statistical information is produced in an unbiased manner, e.g when preparing training data sets for machine learning.

Practices (Implementation guidelines)

When web scraping, members of the ESS should:

- Respect the robots.txt exclusion protocol and only follow links to the extent necessary for maintaining the quality of statistics;
- Respect the wishes of website owners as set in terms and conditions insofar as practical to check those terms, and scraping is not essential to maintaining the quality of statistics as implied by statistical law;
- Identify themselves in the user-agent string and provide contact channels. This could include a link to a web-page explaining the scraper's purpose and what data it collects, responsible team contact details, and information on how to opt-out and request that extracted data be deleted;
- Follow internet standard conventions for scraping, such as standards established by the W3C consortium[1];
- Be transparent about its web scraping activities, possibly by providing information on the associated website;
- Inform website owners if a considerable amount of data is extracted on a regular basis. This would not be the case if a website is scraped with a low frequency and is not scraped in full depth;
- Seek to minimize burden on web servers, by:
    - adding idle time between requests,

[1] See : https://www.w3.org/TR/

Funded by
the European Union

    - scraping at a time of day during which the web server is not expected to be under heavy load,
    - optimise the scraping strategy for minimising the number of requests to a domain;
- Only scrape data within the scope of the statistical office's legal mandate, and do not re-use or distribute the raw data;
- Handle web scraped data securely;
- Avoid web scraping in using public APIs or other data provision options where available.

3.2. Throughput phase I – web scraping

3.2.1. Linking

Specific guidelines for web-scraping

- Find possible classifications that can be extracted from the website, e.g., date, territorial unit, sector as well as enterprise: business id, name, address etc.
- When you are referring to official statistics data sources, it is important to start with official classifications (e.g., NTS, business ID) and prepare algorithms to extract this information from websites if available.
- Use available statistical data to derive meaning and context from raw scraped data.

3.2.2. Coverage

General guideline:

Establish the population of interest.
The definition and study of coverage errors require the definition of the target population that should be explicitly identified in terms of type, time, and place.

Specific guidelines for web-scraping:

Representativeness - Try to estimate the population size and compare with traditional data.
For example, when you are scraping enterprise characteristics, try to count the number of websites that are accessible and can be used for web-scraping. Compare this number with the data from your business register.

Coverage - Relevant data available on website: Make a pilot web-scraping to assess what information is included on the websites.
Check if specific information, e.g., territorial unit or industrial sector can be extracted from the website. When information on the website is limited, it is also not very likely to monitor enterprise activity (e.g., innovations in enterprises) on the website. It is also important to monitor if the information is up to date and if changes over time can be identified (in longer time series).

Funded by
the European Union

3.2.3. Comparability over time

General guidelines

Closely monitor the structure of the data.
Check each data generation on structural changes in comparison to the previous one.

Continuous updating of the data acquisition and recording tools:
Web-scraping, text processing and machine learning tools have to be agile to follow the necessary changes of the data source. For example, if the website (e.g. a job vacancy portal) changes its structure, a person at the NSI responsible for web-scraping has to change the web-scraper to record the appropriate data. In other words, to scrape the data in a long time series, we need to monitor changes on the website and quickly modify web-scrapers.

Fit an appropriate statistical methodology for producing the output.
According to the Analyse Stage of data generating process by AAPOR (2015), apply a statistical method not sensitive to extreme data and define statistical tools for smoothing the break in the time series related to structural changes of the data source or coverage changes over time.

Specific guidelines for web-scraping:

Check if the modification/update date can be extracted from the website.
When web-scraping specific information from the website (e.g., job vacancies), try to extract the date of publishing this information. If the website is not up to date, it is unlikely to detect enterprise activity in longer time series. Since scraped data can disappear or change over time, the risk of missing data is high: monitoring and maintaining scrapers up to date should be considered as a priority.

3.2.4. Measurement errors

General guidelines

Establish the target information.
The definition and study of measurement errors require the definition of the target variable of interest.

Research on measurement errors.
If possible, measurement errors should be evaluated (on a small sub-sample) with an appropriate method, e.g., manual reviewing or comparison with other data.

Track changes need to be observed.
If values are changed or imputed because of detected errors or implausibilities , these changes should be tracked.

Specific guidelines for web-scraping

Funded by
the European Union

Verify if the data in the web fits the definition from official statistics.
It should be noted that sometimes the same variables may have different definition.

3.2.5. Model errors/ process errors

General guidelines

Estimating the quality of models is of great importance:

Apply appropriate model selection and evaluation criteria.
Techniques like cross validation, out-of-sample tests, etc. should be applied wherever possible to assess the model quality and possible errors.

Compare multiple machine/statistical learning methods.
Since it is not always straightforward to choose the right tool for the job, different methods should be tested and evaluated.

Evaluate the bias of the training data set.
In supervised learning, an unbiased training data is very important to not estimate based on a biased model.

3.3. Throughput phase II

3.3.1. Replacement of questions from surveys

Example: ICT usage in Enterprises

For a description of the example, see Subchapter 3.3.2

Guidelines for this example

Compare information coverage.
First, it is important to compare the coverage of the traditional survey with the possibilities of the big data source. Coverage is one of the most important aspects. Sometimes, for example in Online Job Vacancies data, the definition of job vacancy in the traditional survey may be different than the one used in the big data source (online job vacancies).

Compare definitions.
The second issue is to have a unified metadata set – it is necessary to compare all definitions of data gathered in traditional data sources vs. metadata in big data sources.

Measure and report accuracy of applied models.
Due to the complexity of new data sources, e.g., the data of websites may lead to the use of machine learning algorithms, it is also important to measure accuracy of the data set and the information provided.

Funded by
the European Union

3.3.2. Validation / comparison of results with results from traditional data source

Example: Survey on ICT usage and e-Commerce in Enterprises

A subset of the estimates currently produced by the sampling survey on "Survey on ICT usage and e-Commerce in Enterprises", yearly carried out by EU member states, includes as target estimates the characteristics of websites used by enterprises to present their business (for instance, if the website offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data is collected by means of traditional questionnaires.

These results can be compared with results based on new data sources, e.g., data collected by accessing the websites directly (i.e., via web scraping). The collected internet texts have then been processed to individuate relevant terms, finally the relationships between these terms and the characteristics of interest for the estimates are modelled.

Hence, the sequential application of web scraping, text mining and machine learning techniques represent the prediction approach to produce estimates that can be compared to the ones based on surveys.

In this kind of applications, the comparison allows a large number of quality evaluations: it is possible to compare the variability and the bias due to sampling variance, total non-response and measurement errors in the traditional survey vs the model bias and variance in the prediction approach. Further, one can produce aggregate estimates as well as to predict individual values.

Quality guidelines relevant for this application

- Assess the coverage of the population considered by the new data sources compared to the target population (mainly risk of undercoverage);
- Assess the prediction errors of the model-based approach.

3.3.3. Survey based estimation with auxiliary information / calibration

Example: Business survey with web-scraped information

In a business survey the question if the company has a web page is asked. Additionally, for all enterprises in the frame, online presence is tested with web-scraping methods. As this information might not be totally equivalent to the survey question definition, it cannot be used directly to estimate the total number / or ratio of enterprises with a web page. However, the web-scraped information will probably be strongly correlated to the response to the survey question and is known for the whole population. A straightforward way to improve the precision of the survey estimates might be to calibrate the survey weights in such a way that the survey estimates for the number of enterprises with an online presence (according to the web-scraping definition) match with the same number for the whole population.

Quality guidelines for this application:

Check definitions.

Funded by
the European Union

The variables from the big data source are checked regarding contents and definitions before used in a non-response analysis, weight adjustment or in general in a model assisted survey estimate.

Information must be trustworthy.
The quality of the information needs to be checked before it is used in such methods, since the survey theory regards the information to be known true population values in most scenarios.

Prefer auxiliary information on unit level.
If the auxiliary variable is available at the unit level, it is preferable to a situation with only information on the macro level, e.g. totals.

Estimators based on base weights are compared with adjusted estimators.
The base weight is a factor; usually the product of the design weight and a non-response factor assigned to each sampling unit before calibration. Estimators of the relevant key figures of the concerned statistics are analysed (e.g., the number of unemployed in LFS). Marginal totals of persons, households or businesses for important breakdowns are analysed.

Describe methodology and short-comings.
It should be described and publicly available how the method is applied and what effect can be seen compared to the base weights (see previous guideline). Possible short-comings should be clearly stated.

Funded by
the European Union

4. Software architecture

4.1. BREAL

One of the results achieved by the ESSnet Big Data II is the creation of a European reference architecture for Big Data. BREAL (Big Data REference Architecture and Layers) was conceived to assist NSIs in planning Big Data investments, and foster the implementation of standardised solutions for Big Data. BREAL is compliant with several reference architectural frameworks, such as the ESS Enterprise Architecture Reference Framework (EARF), as well as official statistical standards (GSBPM, GSIM, CSPA). BREAL is a set of artifacts structured according to the following architectural layers:

- The Business Layer, describing the main activities and behaviours (What) to deal with Big Data. This layer includes several artifacts, namely: (i) a set of principles; (ii) a set of business functions; (iii) the main steps of a production process based on Big Data, called Big Data Life Cycle; (iv) the main Actors and Stakeholders involved in Big Data acquisition and processing.
- The Application Layer, concerning the application components and services (How) to develop, in order to realize the business functions and the Big Data Life Cycle.
- The Information Layer, analysing data models related to the business functions and the Big Data Life Cycle.

4.1.1. BREAL Business Layer

The main goal of BREAL Business layer is to provide an overview of the business functions related to the use of Big Data for statistical purposes [Scannapieco et al, 2019]. ArchiMate, an open and independent language for Enterprise architecture design, defines a business function as follows: "*a behavior element that groups behavior based on a chosen set of criteria (typically required business resources and/or competences)*". More precisely, a business function describes behaviours related to the organization of resources, skills, or knowledge, while a business process refers to behaviours based on a sequence or flow of activities performed to achieve a specific product or service. BREAL business functions are grouped in two main subsets: "Development, production and deployment" and "Support". The first subset refers to the core abilities that allow to ingest, process and disseminate new data sources, while the second group includes the abilities supporting the core business belonging to the first group. The following figure shows BREAL business functions and their alignment with official statistical standards and reference architectures.

Funded by
the European Union

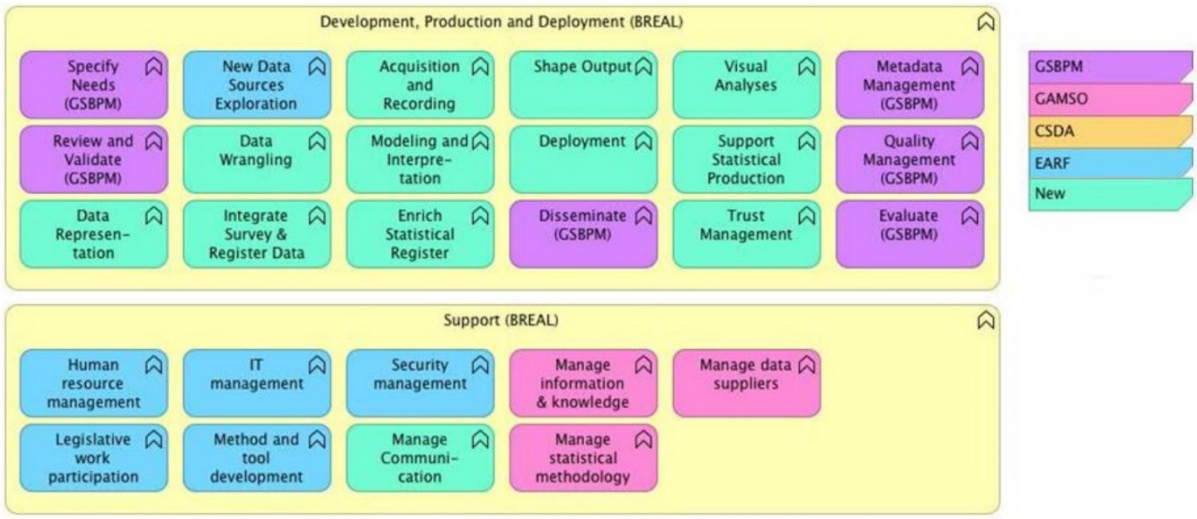

Figure 1:BREAL business functions[2]

The following table reports a summary of the definitions describing BREAL business functions, grouped according to the related subset and the reference framework.

Table 1:: BREAL business functions

Business Function	Description	Reference framework
Development, production and deployment		
Acquisition and recording	Ability to collect data from Big Data sources	
Data representation	Ability to derive structure from unstructured or partially structured data	
Data Wrangling	Ability to transform data from the original source format into a specific target format	
Deployment	Ability to take and integrate into production the (new) statistical product obtained from Big Data sources	
Disseminate	Release of the statistical products created from Big Data sources to users	GSBPM
Enrich statistical Register	Ability to enrich the statistical register(s) with the information retrieved from Big Data sources	

[2] Source: Scannapieco M., Bogdanovits F., Gallois F.; Fischer B., Kostadin G., Paulussen R., Quaresma S. et al. (2019): BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT.

Funded by the European Union

Business Function	Description	Reference framework
Evaluate	Ability to assess a specific instance of a Big data process considering relevant aspects such population coverage, accuracy and fitness of the model	GSBPM
Integrate Survey &Register Data	Ability to reuse and integrate data from surveys & registers to enrich Big Data statistical output	
Metadata management	Relevance of the creation, use and storage of statistical metadata throughout data processing	GSBPM
Modelling and Interpretation	The ability to design, implement and test new algorithms and models for Big Data processing	
New data sources exploration	Ability to find and explore Big Data sources to use for statistical purposes	EARF
Quality management	Ability to understand and manage the quality of the statistical products to improve the business process	GSBPM
Review and Validate	Ability to identify data errors and discrepancies and perform input data validation	GSBPM
Shape output	Ability to format and present data emerging from data patterns or during data exploration	
Specify needs	Ability to identify statistical needs that could benefit from Big Data sources	GSBPM
Support Statistical Production	Ability to support the current statistical production system(s)	
Trust management	Ability to trust the use of Big Data sources and guarantee a similar or higher quality of derived data compared to traditional surveys	
Visual Analysis	Ability to format and present data in such a way as to optimally analyse the results	
Support		
Human Resource Management	Ability to create and maintain human resources with specific skills required to deal with Big Data sources	EARF
IT management	Ability to manage tasks and decisions regarding IT assets for Big Data sources	EARF

Funded by
the European Union

Business Function	Description	Reference framework
Legislative work participation	Ability to cooperate to the legislative work providing the legislative basis of official statistical production. The privacy issue is particularly relevant for Big Data sources	EARF
Manage Communication	Ability to communicate and involve all stakeholders	
Manage data suppliers	Ability to manage different activities related to the relationships with data suppliers	GAMSO
Manage information &Knowledge	Ability to use different information and Knowledge assets, such as policies, guidelines and standards for improving data management and governance	GAMSO
Manage statistical methodology	Ability to manage different activities related to the statistical methodology used in the production process	GAMSO
Method and tool development for new statistics	Ability to develop methods and tools to support the exploration and innovation of new statistical products based on Big Data sources	EARF
Security management	Ability to ensure data confidentiality, integrity and availability through adequate security measures and policies	EARF

Most of the business functions described above relate to the Big Data Life Cycle that includes three main business process areas:

- Development and Information Discovery, including Big Data exploration and integration with other data, as well as the discovery of information that can be obtained from the new sources.
- Production, concerning the creation of statistical products based on the use of Big Data sources.

  Continuous Improvement, related to Big Data monitoring and assessment of its usage, especially in terms of population coverage and methods.

The following table reports the business functions directly related to each business process area.

Table 2: BREAL business functions and Big data life cycle

Big Data Life Cycle	Business functions	Reference framework
	Acquisition and recording	

Development and Information Discovery	Data Wrangling	
	Data representation	
	Review and Validate	GSBPM
	Modelling and Interpretation	
	Integrate Survey &Register Data	
	Enrich statistical Register	
Production	Deployment	
	Support Statistical Production	
	Disseminate	GSBPM
Continuous improvement	Evaluate	GSBPM
	Quality Management	GSBPM

Big data Actors and Stakeholders

A particular feature of Big Data systems is the distribution of tasks and resources between several organizations. Although each stakeholder is responsible for specific activities, the following roles allow to identify the different actors involved in Big Data projects, namely:

- IT & Statistical Pipeline Actors
- Capacity Providers
- Global roles (Statistical Institutions and System Orchestrator)
- Audit, Control, and Compliancy Actors
- Further Actors, related to (but not involved in) the process.

The first subset groups several actors involved in the first process area of Big Data Life Cycle: Development and Information Discovery. More in detail, Computing Architects specify the architectural requirements of the data system, covering both infrastructure and data platform frameworks. Information Architects are in charge of all activities related to 'Data Wrangling' and 'Data Representation'. These activities include data validation, data cleaning, data conversion, data aggregation, as well as data modelling. Once the data is prepared, Data Scientists explore its potential for statistical purposes, by managing the 'Analysis and Visualization' function. Finally, the main task of Domain experts is to design statistical models and indicators, to investigate the topic of interest, assess the consistency of the information retrieved and validate the achieved results.

The main actors, either internal or external to the organization responsible for the project, playing the role of 'Capacity Providers' are: Data Provider and Framework Providers. The Data Provider gathers primary data from the related source, addressing relevant issues, such as data persistence and definition of policies for data access and use by third parties. The Framework Provider is responsible for the management of infrastructure frameworks, data platform frameworks, and processing frameworks, thus delivering resources or services used by several actors for a specific application.

Among the global roles, the System Orchestrator acts on behalf of the Statistical Institutions, to integrate internal and external functions and define the whole set of requirements to be fulfilled by the system. These requirements concern also the business context and include the specification of resources, policy, governance, as well as the activities for monitoring or

Funded by
the European Union

auditing the implemented solutions. The System Orchestrator acts as the business owner of the system, supported by a high-level body within the NSI, such as scientific counsellors or an advisory board, and is in charge of the design and management of the Big Data process.

The actors included in the ”Audit, Control, and Compliancy” subset are related to three transversal functions, namely: Overall Data Management, Security and privacy, Scientific and data relevance. In order to guarantee data consistency between different data sources, the Chief Data Officer will identify and prioritize user needs, to select and explore Big data sources to exploit for statistical purposes. Concerning the second business function, the Data Protection Officer (DPO) is responsible for the fulfilment of security and privacy requirements and other measures established by the legal framework, or by the System Orchestrator. In relation to the third business function, Scientific & Academic Communities, sharing knowledge and expertise, foster process enhancements and the increase of data consistency and relevance.

The last subset of actors that may benefit from Big data potentials and analysis, although not directly involved in the process, groups several stakeholders, such as:

- Other Statistical Institutions, sharing or providing reference frameworks, processes, methodologies and tools that have an impact on NSI’s activities.
- Governments, responsible for issuing laws to protect citizens’ privacy and define rules and constraints to regulate Big Data use, also for statistical purposes. In addition, Governments may express the need of new statistical analysis for decision making.
- Medias, publishing statistical results and explaining methods and data sources used in the statistical process. They also may underline new information needs, and dealing with Big data they contribute to the design of new processes.
- Citizens, which allow to collect Big data by using smart devices in everyday life. They may contribute to the statistical process in an active way, by sharing data they are gathering for research, studies or other purposes.

4.1.2. BREAL Application Layer

The Application Layer of BREAL provides an overview of the software applications implementing most of the business functions described above [Scannapieco et al, 2021]. A subset of business functions has been excluded from the following analysis in terms of application services, due to their dependencies on NSI’s internal policies and production chain.

The Application services implementing Development, Production and Deployment business functions are listed and briefly described in the following table.

Table 3 The implementation of Support business functions relies upon the application services briefly described in the following table.

Business functions	Application services description
Acquisition and recording: ability to retrieve and store Big data sources	*External structured data retrieval*: to retrieve structured data from heterogeneous sources through APIs, from published tables or databases

Funded by the European Union

	External unstructured data retrieval: to retrieve unstructured data, such as text or images captured by satellites, or retrieved from websites, news, social networks
	Data storing: to store data acquired through the data retrieval services
Data Wrangling: ability to transform and manipulate raw data format and volume to facilitate data handling	*Data preparation, filtering, and deduplication*: to transform data in a machine-usable format, filter relevant parameters from unstructured data, remove duplicate data
	Data standardization: to transform acquired data (e. g.: data scaling, data encoding), thus allowing the comparison with other data sources
	Data aggregation: to create large clusters of observations and reduce the amount of computational resources, if a greater raw data granularity is not requested
Data representation: ability to contextualize and structure processed raw data with additional information, useful for the following function "Modelling and Interpretation"	*Data derivation*: to derive additional variables from acquired data, such as the Municipality from latitude and longitude
	Structure modelling: to derive and enhance data structure to model unstructured data, by creating categories or other structural features
	Data encoding: to add meta-information concerning formats or classifications to data
Modelling and Interpretation: ability to apply statistical methods to the pre-structured data from the previous function in order to add meaning to these data	*Data linking and enrichment*: to enhance data meaning by linking and enriching Big Data with information from other sources
	Methodology application: to derive relevant statistical information by applying domain-specific methods, mainly machine learning techniques
	Statistical aggregation: to aggregate the outputs of the modelling and linking steps to benchmark the results with other data sources
Shape output: ability to transform the output of the previous services for integrating the achieved results in the statistical pipeline	*Data exchange*: to transform and shape Big Data output for integrating the derived information in the statistical process
Evaluate: ability to store, update and report relevant information for assessing the process implemented in the Big Data pipeline	*Update evaluation database*: to provide a central storage of information concerning a particular instance of the process, executed by a specific service for evaluation purposes
	Create evaluation report: to access the evaluation database and provide evaluation reports concerning effectiveness and efficiency measurements
	Create feedback reports: to gain operational insights based on the analysis of evaluation reports

Funded by the European Union

Trust: ability to truck and report relevant information concerning process transparency and privacy preservation. This function is Complemented by the Quality Management function	*Update provenance database*: to collect and centrally store relevant information related to process transparency
	Update privacy database: to collect and centrally store relevant information related to privacy measures adopted throughout the Big Data process
	Create trust reports: to access the information stored in the provenance and privacy databases and provide several reports for process auditability
Validation: ability to assess data consistency according to a predefined set of rules, thus guaranteeing that data meets the planned purposes	*Structural validation*: to check the compliance of a dataset with the IT structural requirements, specified for data acquisition and treatment
	Content validation: to check data consistency, limiting the analysis to the dataset to validate, or considering auxiliary data sources, also from other providers
	Validation rules management: to manage rules complexity, feasibility, redundancy, completeness thus avoiding inconsistencies and overlapping
	Metrics on data validation: to assess the quality of data validation procedures, adjust data validation rules and evaluate the results of the validation process
Visual analysis: ability to examine data through graphical representation to assess its reliability and consistency	*Cleaning*: to check data reliability by detecting data inconsistencies and errors, missing or duplicated values, as well as outliers
	Exploratory data analysis: to explore data, in order to: 1) highlight useful data content, trends and structure; 2) confirm or reject initial assumptions about data
	Confirmatory data analysis: to confirm or reject initial hypotheses about data, identify enhancements and additional information needs
	Graphical data representation: to provide visual data analysis, in order to: 1) provide a brief description of a phenomenon; 2) represent the results of statistical analysis
Metadata management: ability to create, identify and manage metadata concerning Big data sources according to metadata standards	*Structural Metadata Management Service*: to describe the structure of the information objects involved in the process
	Process Metadata Management Service: to describe and track data processing to guarantee process reproducibility and data quality
	Administrative Metadata Management Service: to track and manage data access and the information to be preserved
	Reference Metadata Management Service: to link the concepts and objects described in the Metadata Management system with statistical units, variables, classifications and rules
	Report Metadata Management Service: to produce detailed reports for the four services described above
Quality management: ability to measure and	*Input Quality Management Service*: to assess and monitor data acquisition and recording stages

Funded by
the European Union

guarantee data Relevance, Accuracy, Timeliness and Punctuality, Accessibility and Clarity, Comparability and Coherence	*First Throughput Quality Management Service*: to assess and monitor raw data processing executed, to achieve intermediate "statistical" data
	Second Throughput Quality Management Service: to assess and monitor the production of a statistical output, based on the intermediate "statistical" data
	Output Quality Management Service: to assess and monitor the dissemination and evaluation stages
	Report Quality Management Service: to produce detailed reports for the four services described above
Support Business functions	**Application services description**

The implementation of Support business functions relies upon the application services briefly described in the following table.

Table 4 BREAL support business functions and application services

Support Business functions	Application services description
Data organization: sub-function within the IT Management function. Ability to organize and manage Big data distribution within the Big Data pipeline	*Data storage*: to support data organization and distribution and improve technical performances of Big data systems, especially in terms of capacity and transfer bandwidth. The first aspect concerns dealing with relevant data volumes, while the second one is related to the amount of information that can be retrieved or transferred in a certain time-lapse
	Data Access services: to manage the data access requests from authenticated users and pipelines. In order to allow data access, this task involves also capturing and updating descriptive and administrative metadata
	Data Platform Services: to combine logical data organization and distribution with the related application programming interfaces (APIs) or methods, including data registries, metadata services or semantic descriptions through ontologies or taxonomies
Infrastructure: networking and computing. Sub-function within the IT Management function. Ability to manage underlying processing and computations performed on stored data	*Processing services***:** to set up technical infrastructures and methods, according to the volume and velocity of data to process, and the requirements of Big Data pipelines
	Networking: to enhance the infrastructure connectivity, both on the internal and the external side, to facilitate data access and transmission, as well as the deployment of responsive processing in Big Data pipelines
Security management: ability to perform the core activities	*Authentication and authorizations services*: to provide and manage authentications and authorizations to the several users and components accessing to data and service resources available in the Big data pipeline

Funded by
the European Union

Support Business functions	Application services description
according to security and privacy requirements	*Monitoring and Audit Services*: to collect, monitor and store the events affecting the system, such as access or changes to data or services. The type of events to be collected is compliant with security policies requirements
Manage communication: ability to establish effective communication with the several stakeholders involved in a statistical project	*Stakeholder Identification and Analysis*: to collect and store information that allows to identify, contact and classify the stakeholders involved in a specific project, as well as the agreements between the parties
	Communication and planning: to track and manage the communication targeted for each stakeholder, or for different groups of stakeholders, according to the communication needs and strategy
Provision agreement management: ability to manage the agreements with information providers to monitor the fulfilment of data provision requirements	*Storing Agreement database:* to collect and store the main characteristics of data provision agreements, such as: deadline for data delivery, data confidentiality and sensitiveness, transmission protocol, data structure, data pre-processing performed at the provider's premises
	Update instance database: to monitor the compliance of each data transmission instance with the provision agreement and issue a warning in case of any discrepancies
	Provision Agreement reports: to produce detailed reports for users/managers, or related services, mainly: Trust Management services, Metadata and Quality management services, Validation services, Acquisition and recording services to track and check data quality and provenance
Legislative work participation: ability to cooperate to the legislative work regulating official statistical production	*Legislation database*: to provide an inventory of laws, directives and provisions enabling the use of relevant and reliable data sources for statistical purposes
	Updating Legislation database: to identify legal enhancements resulting in opportunities to explore emerging Big Data sources and feed back to the New Data Sources Exploration services
	Legislation report: to coordinate the New Data Sources Exploration services with the current regulation related to the new data sources that could be used for the statistical purposes. In addition, legislation gap reports foster the coordination with the Provision Agreement Management services, highlighting lack of legislation that may affect timeliness and quality of official statistics
Method and tool development for new	*Methods inventory*: to gather evidence from the methods applied to perform specific tasks

Funded by
the European Union

Support Business functions	Application services description
statistics: ability to identify methods and tools for processing new data sources. The implementation of a catalogue of methods and tools is essential to promote service reuse and process standardization	*Tools repository*: to provide the elements that allow to deploy and test the implemented tools, such as: general description, requirements, implementation and testing instructions
Human resource management: ability to build specific skills for Big Data projects	*Recruitment services*: to select and enrol new employees
	Training services: to gather information about training initiatives
	Assessment services: to evaluate employees' performances
	Motivational services: to gain evidence for rewarding and motivating and develop a strategy for career planning
	Employee accompaniment service: to collect the capabilities demonstrated or acquired by the staff involved in each project
Information and Knowledge management: ability to capitalize different knowledge assets to improve the information governance and organisation	*Knowledge and methods:* to provide an inventory of information assets produced during a Big data process, such as methodological documentation about data sources, methods and algorithms
	Information feedback: to collect and share the feedbacks from statistical pipeline loops to improve the algorithms applied throughout a process
	Intellectual property management: to document the several aspects of intellectual property and specify conditions of use for each information asset resulting from Big Data statistical processes

4.1.3. BREAL Information Layer

The analysis of the information layer allows to track the transformations of the main information objects involved in the statistical process and identify the target data entities, as well as their relationships. Based on the 'hourglass model' proposed for Trusted Smart Statistics, in BREAL the information layer is composed by:

- The Raw Data Sublayer, including the data sources acquired and stored by the BREAL 'Acquisition and Recording' business function
- The Convergence Sublayer, containing data resulting from 'Data Wrangling' and 'Data Representation' business functions and represented in terms of units of interest
- The Statistical Sublayer, concerning data produced by the following business functions: 'Modelling and Interpretation', 'Integrate Survey and Register Data', 'Enrich Statistical Registers', and 'Shape Output'. This sublayer groups the target concepts resulting from data analysis.

Funded by the European Union

Each sublayer is composed by specific data entities, namely: BD data entities, GSIM entities and provenance metadata entities. The following table provides a brief description of data entities belonging to the different information sublayers.

Table 5 BREAL information layer

BD data entities description	GSIM entities	Specific provenance metadata entities
Raw data sublayer		
A **Big Data Source** characterized by Volume (defined by two entities: Big Data Size and Non-Big Data Size) Variety (defined by two entities: Structured and Unstructured) and Velocity of data acquisition (defined by two entities: In-Motion and At Rest)	A Big Data Source is a specialization of **GSIM Data Resource**	**Acquired Entities** specifying data selected from an existing source **Provider Agents** specifying data provider
Convergence data sublayer		
Big Data Unit Type: entity composed by BD Variable which can be classified in : Identification, Core and Derived	A Big Data Unit Type entity is a specialization of **GSIM Unit Data Type**	**Throughput Activity** concerning specific operations executed on raw data and specialized into two entities: **Lower layer** (including metadata related to "Data Wrangling" activities) and **Upper Layer** (including metadata related to "Data Representation" and "Data Modelling and Interpretation" functions)
Statistical data sublayer		
Big Data Information Set: entity containing data derived from Big data which can be used to produce internal or external output	Big Data Information Set is a specialization of **GSIM Information Set**	**Lineage of Output entities**: starting from the provenance metadata, this entity considers the whole process executed on raw data from the output viewpoint

Funded by
the European Union

The following figure shows the entities belonging to each sub-layer, as well as their relationship. More in detail, the specific BD data entities are represented in blue colour, GSIM entities in pink and specific provenance metadata entities in yellow.

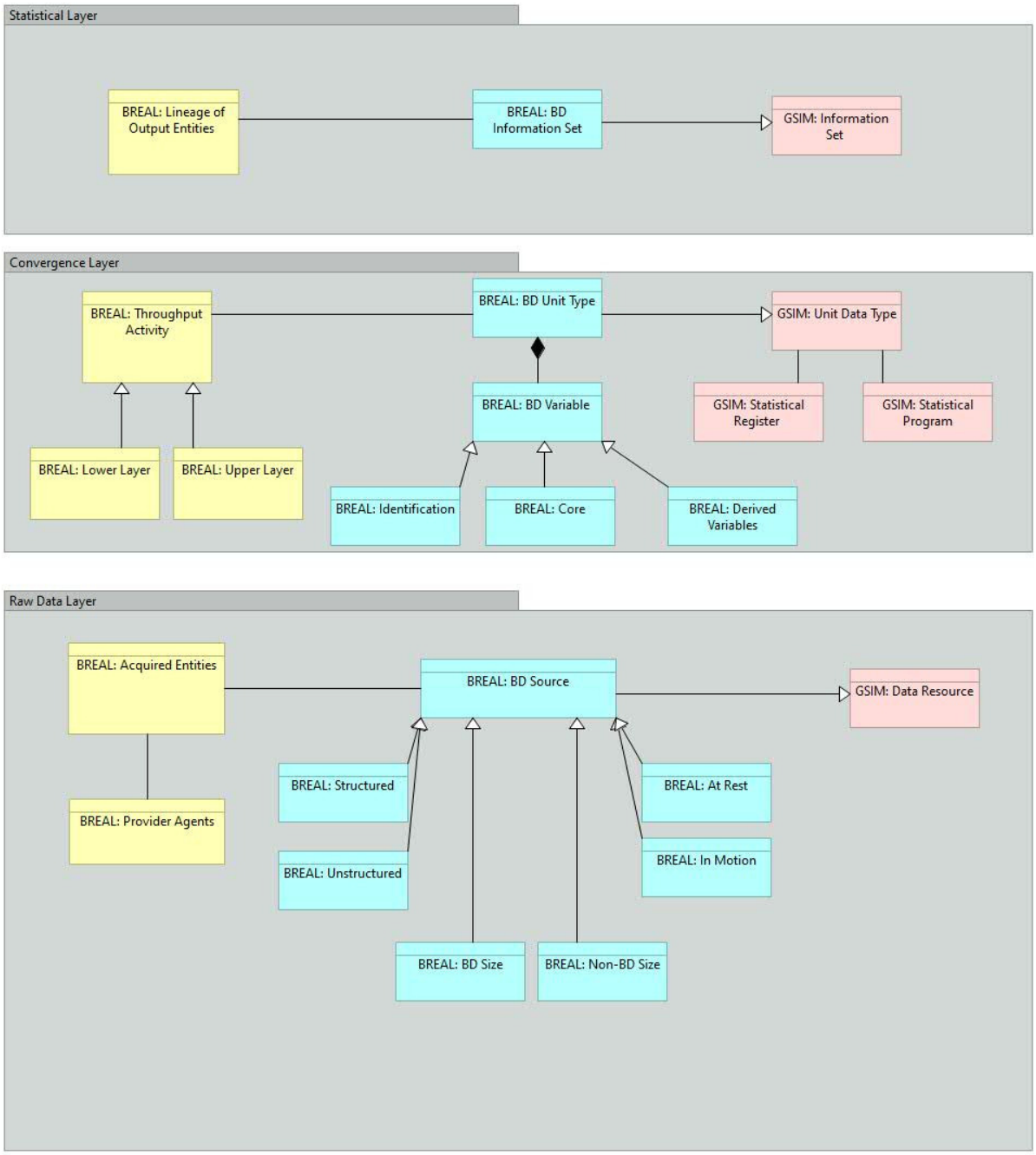

Figure 2 BREAL Information layer[3]

[3] Source: Scannapieco M., Bogdanovits F., Gallois F.; Fischer , Kostadin G., Paulussen R., Quaresma S. et al. (2021): BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2021-03-31. Edited by EUROSTAT

Funded by the European Union

4.1.4. BREAL Operational Model

The main goal of an operational model is to define how the implemented solutions will be deployed in relation to specific dimensions. Starting from the ESS Enterprise Architecture Reference Framework[4], BREAL operational model is based on a combination of the following types of sharing:

- Sharing application services
- Sharing infrastructure platforms
- Sharing data sources.

According to ESS EARF, application services can be grouped in:

- Autonomous services: implemented and executed in NSI's environment, without coordination with other countries.
- Interoperable services, having a similar service interface across the NSIs, and different back-end implementations.
- Replicated services: deployment of the same application service for different NSIs.
- Shared services: all NSIs may access to an available and shared application service.

Concerning Big data infrastructure platforms, depending on their location and owner, the following scenarios may occur:

- Local platform completely developed and managed by the NSI.
- Local platform of the data provider where NSIs may perform the statistical analysis to achieve a statistical output.
- Shared platform, operated and managed by a third party and accessed by NSIs for the statistical production.

The relevance of data sharing depends on the level of data coverage and considering this dimension, data can be grouped as follows:

- Local data, including data that can be processed only by a specific NSI, due to privacy issues, or to their relevance only at national level.
- European data, concerning data covering all European countries.
- Worldwide data having a global coverage.

The proposed operational model, described in the figure below, is the result of the following combination of platforms, application services and type of data:

- Local platforms operated by NSIs providing autonomous, interoperable and replicated services to process local data. European and worldwide data could be processed as well, but due to the amount of resources needed, it is recommended to share these data sources (grey coloured in the figure below) through the platforms operated by data providers or through shared platforms.
- Local platform of data providers, where shared application services may allow to process European or worldwide data sources. Local data sharing, or the execution of autonomous or interoperable services (grey coloured in the figure below) would be more difficult in this case, due to privacy regulation, software maintenance or trust issues.

[4] https://ec.europa.eu/eurostat/cros/content/ess-enterprise-architecture-reference-framework_en

Funded by
the European Union

- A shared platform providing mainly share application services is the best solutions to process European or worldwide data. In addition, if available on the platform, local data as well as autonomous or interoperable application could be accessed, in compliance with the privacy regulation.

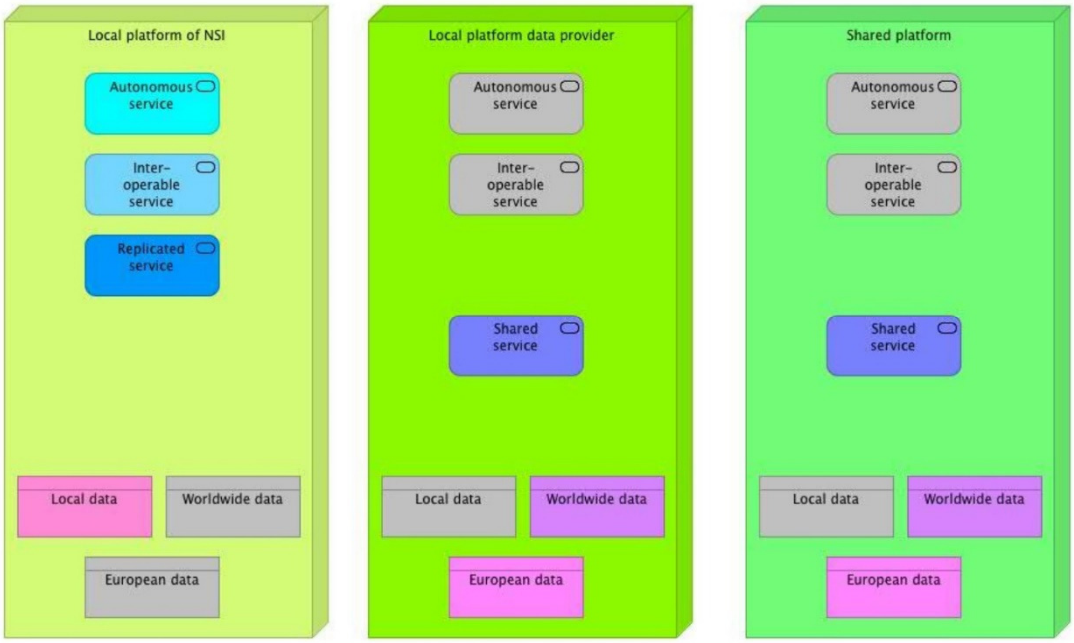

Figure 3 BREAL Operational model[5]

The analysis of BREAL layers has highlighted the advantages of adopting a reference architecture, such as:

- The reuse of existing software solutions and application components.
- The compliance with official statistical standards, and the alignment with other relevant frameworks contributing to the statistical process enhancement.
- The development of skills and capabilities, both at national and European level.

[5] Source: Scannapieco M., Bogdanovits F., Gallois F.; Fischer , Kostadin G., Paulussen R., Quaresma S. et al. (2021): BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2021-03-31. Edited by EUROSTAT

Funded by the European Union

4.2. Use Cases

The following paragraphs describe the architectural layers modelling, according to BREAL, the software solutions developed during the ESSnet Big data II[6] for Online Job Vacancies (OJV) and Online-Based Enterprise Characteristics (OBEC). The activities carried out by the implementation work packages have proved the relevance of BREAL as a reference architecture, to identify and bridge the gap between traditional statistical processes and Big data pipelines.

4.2.1. Online Job Vacancies (OJV): application and information architectural layers

During the ESSNet Big Data II, work package B has concerned the estimates produced for Online Job Vacancies (OJV), in order to integrate in the statistical process, the methodologies resulting from the ESSNet Big Data I and develop prototypes. Concerning the application layer [Scannapieco et al, 2020a], the following table reports a summary of the specific application services and components of the OJV pipeline, specialising BREAL application services related to BREAL business functions.

Table 6 BREAL business functions, application layer and OJV application services and components

BREAL Business processes and functions	BREAL Application services	OJV External application services specialising BREAL Application services	OJV Application Components and Services
Acquisition and recording	External structured data retrieval	API access	API Querying
		Specific scraping	Scraping and Crawling
	External unstructured data retrieval	Generic scraping	
		Website retrieving	
	Data storing		Create URLS inventory
			Create scraped data storage
Data wrangling	Data preparation, filtering, and deduplication	Ex-post data filtering	Web data processing
		Language detection	
		Deduplication & mapping	
	Data standardization	Format standardization	
		Text processing	
	Data aggregation	Data representation	TD matrix
			Word and sentence embedding
			Pretrained NLP model
		Pseudo-stock transformation	Compute aggregates
Data representation	Data derivation	Create indicators	

[6] A description of the Implementation track workpackages is available from: https://ec.europa.eu/eurostat/cros/content/implementation-track-0_en

BREAL Business processes and functions	BREAL Application services	OJV External application services specialising BREAL Application services	OJV Application Components and Services
	Structure modelling	Classification & standardization	Build training and test set
			Data set labelling
			Train, test and validate
			Apply ML
Modelling and interpretation	Data linking and enriching	Linking service	Linking data
		Deduplication service	
Shape output	Data exchange	Match and register data	Enrich register units
		Output building service	Small simulation
			Benchmark comparison
			Results validation (Quality)

The first business function, "*Acquisition and recording*" is realized by some application services that allow to retrieve structured and unstructured data through a "Scraping and crawling component". This module can perform both, a specific or generic scraping and website retrieving. In addition, an "API Querying" component provides an API access to retrieve information from websites having a known structure. The last step of data acquisition and recording is completed through two application components, to create URL inventory and store data retrieved through web scraping.

The second business function, "*Data wrangling*" is implemented by several components, namely: "Web data preparation", "TD matrix", "Word and sentence embedding" and "Pretrained NLP models". The main steps of raw data preparation, such as: ex-post filtering, language detection, deduplication and merging are executed by the "Web data preparation" components, also used to perform data standardization for the next steps. The following components, "TD matrix", "Word and sentence embedding" and "Pretrained NLP models" allow to perform a first data aggregation to check the validity of a job offer, based on the publication date for instance. The module "Compute aggregates, used also for the "*Data representation*" business function, computes derived data used in the subsequent steps. In addition to this component, "Data representation" is realized through the following modules, performing the main iterative steps for the application of machine learning (ML): "Build training and test set", "Data set labelling", "Train, test and validate" and "Apply ML".

After data representation, the "Modelling and interpretation" business function is implemented by the "Linking data" component, conceived to process data related to the same job vacancy, performing data deduplication and linkage. The last business function, "Shape output", is realized by the "Enrich register units" module, using the data retrieved from job offers to enhance the available information for register units. Other components: "Small simulations", "Benchmark comparisons" and "Result validation", perform iterative tasks to produce and validate the final output.

Funded by the European Union

The whole process is supported by two crosscutting components: "Workflow manager" and "Metadata management service", respectively to manage the process steps execution, and share process metadata. The information layer related to the OJV pipeline is summarized in the following table.

Table 7 OJV information layer

OJV data entities description	**BREAL BD data entities/GSIM entities**
Raw data sublayer	
A **Job portal** is the only data resource, raw data are acquired through web scraping or crawling and API querying	A Job portal is an instance of **GSIM Data Resource**
Convergence data sublayer	
OJV Unit Type: entity composed by OJV Variables which can be classified in: Identification (ID number, information about portal URL$_s$), Core (job title and location) and Derived (grouping the remaining variables obtained from job vacancies)	OJV Unit Type entity is a specialization of **BD Unit Data Type**
Statistical data sublayer	
OJV Information Set: set of entities containing data derived from OJV such as: Skills, Education degree, Change in demand by occupation, skills region, Flow of Oja$_s$, Number of Oja$_s$, Uncertainty metrics, Statistics on skills, Regional dimensions of Oja$_s$. These data entities are used to produce internal or external output	OJV Information Set is a specialization of **BD Information Set**

4.2.2. Online-Based Enterprise Characteristics (OBEC): application and information architectural layers

In order to enhance the national business registers, within the ESSnet Big data II, work package C [Scannapieco et al, 2020b] has explored web scraping, text mining and inference methods to gather and process enterprise information, such as Web presence, location, type of activity. During the implementation activities, two main use cases: 1) URLs Inventory of enterprises, and 2) Variables in the ICT usage in enterprise survey have investigated the combination of several sources to produce experimental statistics at national level. The description of the application layer for both use cases is reported in the following table.

Table 8 : BREAL business functions, application layer and OBEC application services and components

BREAL Business processes and functions	**BREAL and OBEC Application services**	**OBEC Application Components and Services**
Acquisition and recording	External structured data retrieval (URLs inventory use case)	Scraping

BREAL Business processes and functions	BREAL and OBEC Application services	OBEC Application Components and Services
	External unstructured data retrieval	
	Data storing	Create URLs inventory (URLs inventory use case)
		Create Document Base of Scraped data
Data wrangling	Data preparation, filtering, and deduplication	Web data preparation
	Data standardization	
Data representation	Data derivation	Web data preparation (URLs inventory use case) TD Matrix (Variables in the ICT usage)
	Data encoding	
Modelling and interpretation	Methodology application	Build Training and Test set
		Training, test and validate
		Apply ML
	Data linking and enriching (URLs inventory use case)	Perform linkage
Shape output	Data exchange	Enrich register units
		Perform linkage (URLs inventory use case)

The "*Acquisition and recording*" business function is realized by the "Scraping application component" to retrieve both structured and unstructured data, depending on the number of websites to scrape. Another module implementing this business function in the first use case is "Create URLs inventory", which allows to build the list of eligible enterprises' URLs. In addition, in both use cases, the application component "Create Document Base of Scraped Data" provides a service to store scraped web data in relational or NoSQL databases.

Concerning "*Data wrangling*" business function, the "Web Data Preparation" allows to prepare scraped data for the next data processing steps. More in detail, this module transforms scraped data in a predetermined format. In the case of URLs Inventory, this component provides functionalities to: i) define a structure from unstructured data; ii) perform data standardization; iii) implement the "*Data representation"* business function by deriving new variables and adding metadata about formats or classifications. In the second use case, for ICT Variables, the TD Matrix component and its services for structuring scraped data without structure, allows to obtain additional variables and add metadata, thus implementing the "*Data representation"* business function.

In relation to the "*Modelling and Interpretation"* business function, three main components provide services to apply machine learning techniques. More specifically, the first module (Build Training and Test Set) allows to prepare sets of data for a chosen machine learning method. The following application component (Train, Test and Validate) is intended to train and validate the machine learning model, while the third module (Apply ML) allows to apply

Funded by
the European Union

the trained model. For URLs Inventory, an additional application component (Perform Linkage) performs the linkage of data from different sources, such as Statistical Business Register, to enrich available data. This component serves both, "*Modelling and Interpretation"* and *"Shape output"* business functions. Furthermore, a specific component (Enrich Register Units) is dedicated to the preparation of statistical outputs for producing statistical results.

The information layer related to the first use case, URLs Inventory of enterprises, is briefly described in the following table.

Table 9 Information layer for the use case concerning URLs inventory

URLs Inventory of enterprises Data entities description	**BREAL BD data entities/GSIM entities**
Raw data sublayer	
Raw data are acquired from several sources: **Enterprise Website** usually providing unstructured information, and a smaller number of **Search Engines, APIs, Yellow Pages** with structured data formats	All data sources: Enterprise Website, as well as Search Engines, APIs, Yellow Pages are instance of **GSIM Data Resource**
Convergence data sublayer	
OnlineBasedEnterprise Unit Type: entity describing the information about OBEC units of interest, based on the available data sources **Survey data**: entity describing core and structured information about OBEC units of interest, collected through traditional sources and used to structure and explain OBEC raw data **Business register**: entity describing OBEC data source storing relevant information, such as enterprises ID, location, URL	OnlineBasedEnterprise is an instance of **BD Unit Data Type** Survey data and Business register are instance of **GSIM Datasets**
Statistical data sublayer	
Indicator on Internet presence: entity describing the statistical output resulting from the use case and providing some statistical indicators about enterprises presence on internet and their characteristics. This information is the core input for the second use case, Variables in the ICT usage in enterprise survey	Indicator on Internet presence is an instance of **BD Information Set**

A summary of the information layer designed for the second use case, Variables in the ICT usage is reported in the following table. While the the convergence data sublayer is the same in the use cases analysed, the data entities included in the raw and statististical data sublayers differ, due to the specific input data sources and the different output resulting from data processing.

Funded by the European Union

Table 10 Information layer for the use case concerning Variables in the ICT usage in enterprise survey

Variables in the ICT usage in enterprise survey **Data entities description**	**BREAL BD data entities/GSIM entities**
Raw data sublayer	
The main raw data source is the **Enterprise Website,** usually providing unstructured information	Enterprise Website is an instance of **GSIM Data Resource**
Convergence data sublayer	
OnlineBasedEnterprise Unit Type: entity describing the information about OBEC units of interest, based on the available data sources **Survey data**: entity describing core and structured information about OBEC units of interest, collected through traditional sources and used to structure and explain OBEC raw data **Business register**: entity describing OBEC data source storing relevant information, such as enterprises ID, location, URL	OnlineBasedEnterprise is an instance of **BD Unit Data Type** Survey data and Business register are instance of **GSIM Datasets**
Statistical data sublayer	
Indicators on estimated variables: entity describing the statistical output resulting from the use case and providing some statistical indicators concerning enterprises ecommerce, job advertisements, social media and other features, such as presence on internet by enterprises size, NACE levels, NUTS levels, number of employees	Indicators on estimated variables is an instance of **BD Information Set**

Funded by
the European Union

4.3. Design Principles

The BREAL business principles provide guidance and general rules to promote the use of Big Data sources for statistical purposes. Most principles, as well as the related statements and rationales, have been developed combining reference architectures and standards that inspired BREAL business functions. Further, they are grouped in several subsets, listed in the table reported below:

Table 11 BREAL design principles

Category	Principle	Reference framework
Data capturing	Use an authoritative source	CSDA principle - paragraph 5
	Information is captured and recorded at the point of creation/receipt	CSDA principle - paragraph 4
	Data is only "used by" the Statistical Office	Big Data Task Force - principle 2
	Standardise and harmonise data as quickly as possible	
	Capture metadata at source	Common Metadata Framework principle
Data processing	Processing method (algorithm) transparent to all involved parties	Big Data Task Force - principle 1
	Push computation out (BDTF)	Big Data Task Force
	Consider all capability elements	CSPA principle, paragraph 69
Relationship management	Engage and partner with the input parties	Big Data Task Force - principle 3
Quality	Quality control is built in	EARF principle
Security	Security is built in	
Reuse (and service-based approach)	Reuse before adapt, before buy, before build	
	Reuse of data	

The following paragraphs explore three relevant aspects that enhance the reuse of available software solutions: reusability, interoperability, and portability.

4.3.1. Reusability

The reusability of data, functionalities, services and knowledge is one of the key requirements of the platform. The WIHP will support three modes of reusability:

- **Data Reusability:** It should be possible to share data between different users, so that processed and semi- processed data are made available to others than the group producing the data.

- **Workflow reusability:** It should be possible to share workflows between use cases, such that one data gathering process can copied and used as a starting point for another use case.
- **Services reusability:** It should be possible for different use cases to make use of functionalities on the platform exposed as services. This could include standard data processing functions, such as extracting standard metadata from the HTML, provide language detection or deduplication of web pages.

The WIHP architecture supports the three modes of reusability with the following application level functions.

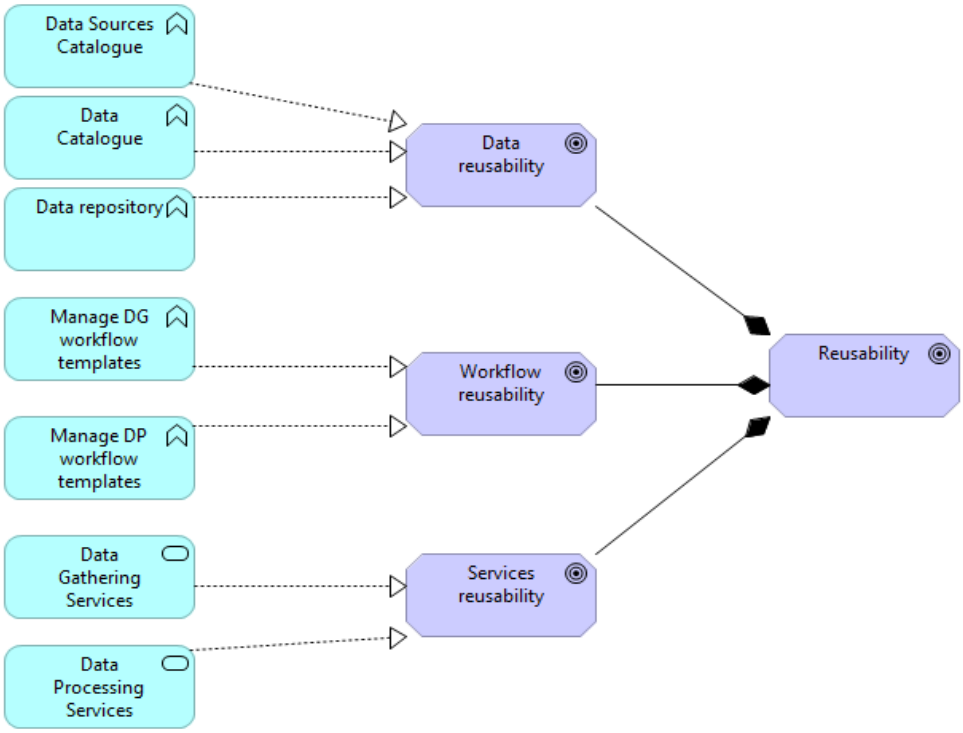

Figure 4 WIHP Reusability view (Eurostat TSS Taskforce)

Table 12 WIHP Reusability modes

Reusability Goal	Description	Application Components
Data Reusability	Data Sharing is the most straight forward means of sharing as it involves sharing the data and metadata in a repository for others to use.	**Data Sources Catalogue:** The platform should enable users to explore existing data sources used in other use cases to assess the possibility of reuse. **Data Catalogue:** The platform should enable users to explore existing data sets for reusability.

Funded by the European Union

		Data Repository: To provide a repository of output data from the different use cases. This can also include semi-processed data.
Workflow Reusability	Workflow sharing is a means of sharing the full setup of executables of a workflow which can be used as a starting point (template) for adaption to other use cases.	**Manage Data Gathering (DG) workflow templates:** Function that allows the exchange of templates for gathering so they can be copied and exchanged between use cases. **Manage Data Processing (DP) workflow templates:** Function that allows the exchange of templates for processing so they can be copied and exchanged between use cases.
Services Reusability	The Services sharing involves the development and generalisation of specific services for e.g. data gathering or data processing. This means of sharing requires dedicated development and maintenance of the service.	**Data Gathering Services:** Services to support the design and execution of data gathering of web data that will ease the setup of a new data gathering workflow and automate tasks. **Data Processing Services:** Services to process the extracted data.

Service Reusability

Services provided by the WIHP can be grouped as follows:

- Data Gathering Service
- Data Processing Services
- Output Data Sharing Services.

A set of WIHP service candidates is displayed in the following figure. This list may change as services should be identified as more use cases start using the platform and may end up differing.

Funded by
the European Union

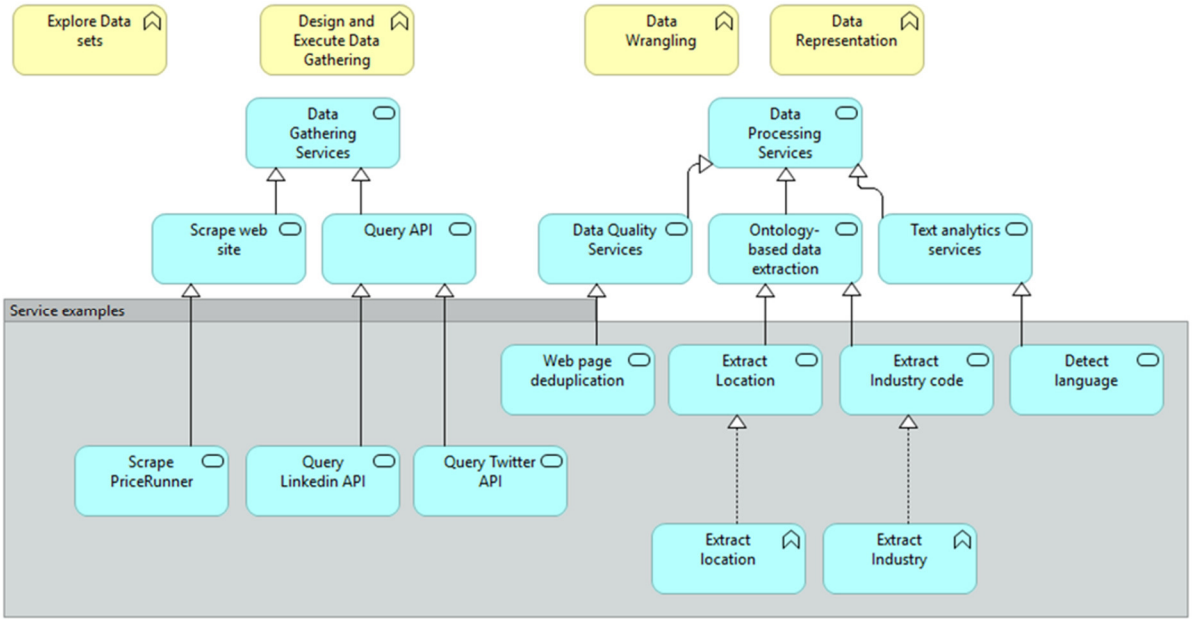

Figure 5 WIHP Data Gathering and Data Processing Services (Eurostat TSS Taskforce)

The evolution of the actual data gathering services and data processing service should take place gradually and driven by the use case making use of the platform. Identifying and developing services without a certainty of a use case using it, would involve a significant risk of developing services that are not used.

The technical implementation of services may evolve over time and could initially be implemented by sharing code on GitHub, which can then be executed in the environment of the use case. If justified, this can later evolve into a more mature service wrapping, which can be called via e.g. a REST API.

Table 13 WIHP Services: Business and Application components

Business level component	Application Components
Design and Execute Data Gathering	**Data Gathering Services:** Services to support the design and execution of data gathering of web data that will ease the setup of a new data gathering workflow and automate tasks. **Scrape web site:** Service to help scrape a websites' specific pages and data. **Query API:** Service providing readily available access to API's that are considered valuable across multiple use cases.
Data Wrangling and Data Representation	**Data Processing Services:** Services to process the extracted data. **Detect Language:** To detect the language in which a web page is written. **Webpage Deduplication:** To eliminate duplicate web pages either via URL deduplication or analysis of multiple factors including the description, publication date, etc.

Funded by the European Union

	Data Quality Services: Services to check and improve the quality of the data. This includes deduplication of data, identifying faulty data based on searches, etc. **Ontology-based data extraction services:** Services extracting data based on ontologies, such as mapping a company to an industry code. These are typically using a combination of reference data and machine-learning techniques. **Text analytics Services:** These are services supporting text analytics techniques by creating data structures well suited for text analytics. This includes creating bag-of-words, sentence extraction, language detection and translation, word stemming, etc.

4.3.2. Interoperability

According to Wikipedia, *"Interoperability is a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, at present or in the future, in either implementation or access, without any restrictions"*[7].

In the context of the WIHP, to achieve interoperability the following principles should guide the design and implementation activities:

- Use of commonly agreed standards for API, for data models, for scripting workflows.
- Open source repositories for scripts (*github* public repositories).
- Accessible environment for testing/benchmarking workflows.
- Possibility to adapt of workflows (parametrization) through UI.
- Support of hybrid workflows (multi platform).
- Sharing quality metrics.

4.3.3. Portability

"Portability in high-level computer programming is the usability of the same software in different environments. The prerequirement for portability is the generalized abstraction between the application logic and system interfaces. When software with the same functionality is produced for several computing platforms, portability is the key issue for development cost reduction."[8]

In the context of the WIHP, to achieve portability the following principles should guide the design and implementation activities:

- **Few dependencies**: contextually independent systems can be deployed almost anywhere because they are relatively independent/self-contained (*support micro service architecture)*.

[7] https://en.wikipedia.org/wiki/Interoperability.
[8] https://en.wikipedia.org/wiki/Software_portability.

- **Well-defined interfaces**: the means of communicating with contextually independent systems are very well defined (*state of the art API management and integration architecture)*.
- **Easily fulfilled dependencies**: the few dependencies that the contextually independent systems have are also easy to fulfil (*platform agnostics and open source standard are preferred)*.

Funded by
the European Union

5. References

AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report, Public Opinion Quarterly, 79, pp. 839–880.

Bakker, B. (2011). Micro Integration. Statistical methods series 201108, Statistics Netherlands, the Netherlands. Located at: https://www.cbs.nl/en-gb/onze-diensten/methods/statistical-methods/throughput/throughput/micro-integration

Barcalori, G., Scannapieco, M., Summa, D., Scarno, M., (2016a). On the use of Internet as a Data Source for Official Statistics: A Comparative Analysis of Web Scraping Technologies. Paper for the NTTS 2015 conference. Brussels, Belgium.

Barcaroli, G., Scannapieco, M., Summa, D. (2016b). On the use of Internet as a Data Source for Official Statistics: A Strategy for Identifying Enterprises on the Web. Rivista Italiana di Economia Demografia e Statistica LXX(4), Ottobre-Dicembre.

Beręsewicz, M., Pater, R. (2021). Inferring job vacancies from online job advertisements. Statistical Working Papers, Eurostat. Located at: https://ec.europa.eu/eurostat/documents/3888793/12287170/KS-TC-20-008-EN-N.pdf

Beręsewicz, M., Cherniaiev, H., Pater, R. (2021). Estimating the number of entities with vacancies using administrative and online data. Located at: https://arxiv.org/pdf/2106.03263.pdf

Cakim, K. (2020). Measuring the Dutch Platform Economy: A collaboration with Statistics Netherlands. Bachelor thesis of the school of Economics, University of Utrecht, the Netherlands.

Cavallo, R. (2017). Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers. American Economic Review 107(1), pp. 283–303.

Cavallo, R. (2018). Scraped Data and Sticky Prices. The Review of Economics and Statistics 100(1), pp. 105-119.

Cavallo, R., Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives 30(2), pp. 151–178.

Chessa, A.G. (2016). A new methodology for processing scanner data in the Dutch CPI. Eurona 1, article 2, pp. 49-69. Located at: https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2016-art2.pdf

Condron, A., Kowarik, A., Gussenbauer, J., Summa, D., Zardetto, D., De Fausti,F., Consalvi, M., Stateva, G., Georgiev, K., Maślankowski, J., Niewiadomska, E., Małek, P., Bis, M., ten Bosch, O., Windmeijer, D., Lorenzi, L., Lee, P. (2019). Reference Methodological Framework for processing online based enterprise characteristics (OBEC) data for Official

Funded by the European Union

Statistic. . Workpackage C, Implementation – Enterprise Characteristics, ESSnet Big Data II, Deliverable C2.

Condron, A., Kowarik, A., Summa, D., Stateva, G., Maślankowski, J., ten Bosch, O., Wood, M., Lee, P. (2019). ESS web-scraping policy template. Workpackage C, Implementation – Enterprise Characteristics, ESSnet Big Data II, Deliverable C1.

Daas, P.J., de Wolf, N.J. (2021). Identifying different types of companies via their website text. Abstract for the SDSS 2021 symposium, online, USA. Located at: https://ww2.amstat.org/meetings/sdss/2021/onlineprogram/AbstractDetails.cfm?AbstractID=309790

Daas, P., Puts, M., Maslankowski, J., Salgado, D., Quaresma, S., Tuoto, T., Di Consiglio, L., Brancato, G., Righi, P., Six, M., Kowarik, A. (2020a). Report describing the methodological steps of using big data in official statistics with a section on the most important research questions for the future including guidelines. Workpackage K, Pilots Track – Methodology and Quality, ESSnet Big Data II, Deliverable K10.

Daas, P., Maslankowski, J., Salgado, D., Quaresma, S., Tuoto, T., Di Consiglio, L., Brancato, G., Righi, P., Six, M., Kowarik, A. (2020b). Revised version of the methodological report. Workpackage K, Pilots Track – Methodology and Quality, ESSnet Big Data II, Deliverable K9.

Daas, P.J.H., van der Doef, S. (2020) Detecting Innovative Companies via their Website. Statistical Journal of IAOS 36(4), pp. 1239-1251, doi/10.3233/SJI-200627.

Daas, P.J.H., van der Doef, S. (2021) Using Website Texts to detect Innovative Companies. CBDS discussion paper 01-21. Statistics Netherlands.

De Mooij, M., Blatt, D., Melser, C., Roos, M. (2020). On line job vacancy data as a source for official statistics: combining online job vacancy data with the job vacancy survey (in Dutch). CBDS Working paper no. 04-20, Statistics Netherlands, Netherlands.

Eurostat (2021). Experimental Statistics: Labour Market Concentration Index using OJAs data - Methodological Note. Draft version.

Kowarik, A., Gussenbauer, J., Mikesa, L., Weinauer, M., Peterbauer, J., Rannetbauer, W. (2020). Webscraped data for replacing and validating survey questions. UNECE paper. Located at: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2020/mtg1/SDE2020_T2_Austria_Gussenbauer_Paper.pdf

Kinne, Jan und David Lenz (2019), Predicting Innovative Firms Using Web Mining and Deep Learning, ZEW Discussion Paper No. Located at: https://www.zew.de/publikationen/predicting-innovative-firms-using-web-mining-and-deep-learning

Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys 3, pp. 96-146.

Funded by
the European Union

Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J., Winton, J. (2018). Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting. J. R. Statist. Soc. A 181(3), pp. 737–756.

Rengers, M., de Lazzer, J., Stateva, G., Saucy, F., Lucarelli, A., Wu, D., Elezovic, S., Grahonja, C., Maślankowski, J., Eidelman, A., Dumesnil de Maricourt, C., Necula, M., Alexandru, C., Columbano,A., Schmassmann, S., Amarone, M., Aprile, D., Chianella, D., Sorrentino, M., Špeh, T. (2020). Report on the Statistical Output, Required Quality and Definition of the Necessary Metadata ate European and national Level. Workpackage WP B, Implementation – Online Job Vacancies, ESSnet Big Data III Deliverable B4.

Reusens, M. (2021) A better statistic on innovative companies in Flanders using web scraping and machine learning. Presentation available at the UNECE wiki, located at: https://statswiki.unece.org/download/attachments/290359872/2020-10%20Statistics%20Flanders%20web%20scraping.pdf

Scannapieco M., Bogdanovits F., Gallois F.; Fischer B., Kostadin G., Paulussen R., Quaresma S. et al. (2019): (Deliverable F1) BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT

Scannapieco M., Bogdanovits F., Gallois F.; Fischer , Kostadin G., Paulussen R., Quaresma S. et al. (2021): (Deliverable F2) BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2021-03-31. Edited by EUROSTAT

Scannapieco M., Bogdanovits F., Gallois F.; Fischer , Kostadin G., Paulussen R., Quaresma S. et al. (2020a): (Annex WPB) BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2020-07-11. Edited by EUROSTAT

Scannapieco M., Bogdanovits F., Gallois F.; Fischer , Kostadin G., Paulussen R., Quaresma S. et al. (2020b): (Annex WPC) BREAL. Big Data Reference Architecture and Layers. Application layer and Information layer. Version 2020-07-11. Edited by EUROSTAT

Singrodia, V., Mitra, A., Paul, S. (2019). A Review on Web Scrapping and its Applications. International Conference on Computer Communication and Informatics, Jan. 23 – 25, 2019, Coimbatore, India.

Stateva, G., Saucy, F., Lucarelli, A., Wu, D., Maślankowski, J., Dumesnil de Maricourt, C., Grahonja, C., Rengers, M., de Lazzer, J., Necula, M., Alexandru, C., Aprile, D., Chianella, D., Sorrentino, M., Schmassmann, S., Columbano, A., Elezović, S., Špeh, T. (2020). Methodological framework for processing online job adverts data for Official Statistics V.2. Workpackage WP B, Implementation – Online Job Vacancies, ESSnet Big Data II Deliverable B3.

Swier, N., Jansson, I., Wu, D. , Nikic, B., Pierrakou, C., Körner, T., Rengers, M. (2016). Interim Technical Report, Workpackage 1, Web scraping / Job vacancies, ESSnet Big Data I, Deliverable 1.2.

Swier, N., Hajnovic, F., Jansson, I., Wu, D. , Nikic, B., Pierrakou, C., Rengers, M. (2017). Final Technical Report, Workpackage 1, Web scraping / Job vacancies, ESSnet Big Data I, Deliverable 1.3.

Funded by
the European Union

Ten Bosch, O., van Delden, A., van den Heuvel, G. (2018). Web scraping meets survey design: combining forces. Paper presented at the BigSurv18 Conference, Barcelona, Spain.

Quaresma, S., Maslankowski, J., Salgado, D., Tuoto, T., Di Consiglio, L., Brancato, G., Righi, P., Daas, P., Six, M., Kowarik, A. (2020). Revised version of the quality guidelines for the acquisition and usage of big data. Workpackage K, Pilots Track – Methodology and Quality, ESSnet Big Data II, Deliverable K3.

Van Delden, A., Windmeijer, D., ten Bosch, O. (2019). Finding Enterprise Websites. Paper for the European Establishment Statistics Workshop 2019, Bilbao, Spain.

Van der Grient, H., de Haan, J. (2010). The use of supermarket scanner data in the Dutch CPI. Statistical Methods paper, Statistics Netherlands, The Netherlands. Located at: https://www.cbs.nl/en-gb/onze-diensten/methods/surveys/aanvullende-onderzoeksbeschrijvingen/the-use-of-supermarket-scanner-data-in-the-dutch-cpi

Funded by
the European Union