Statistical Journal of the IAOS 40 (2024) 549–562
DOI 10.3233/SJI-240032
IOS Press

A web intelligence information system to support the production of EuroGroups Register (EGR) statistics

Antonio Laureti Palma[a], Alexandros Bitoulas[b,*], Alexandre Depire[c], Fernando Reis[c], Pau Gaya Riera[c] and Ioannis Sopranidis[c]
[a]*Italian National Institute of Statistics (Istituto Nazionale di Statistica – ISTAT), Rome, Italy*
[b]*Sogeti Luxembourg, Luxembourg*
[c]*Eurostat, Directorate-General Statistical Authority of the European Union, Luxembourg*

Abstract. There is a growing demand for statistics to better understand the globalisation that is accelerating due to the removal of barriers in international trade and to technological progress. Key players in globalisation are the multinational enterprise (MNE) groups that have increased in number and complexity and need to be properly represented by macroeconomic and business statistics. To deal with this need, the European Union Member States, the European Free Trade Association (EFTA) countries and Eurostat have collaborated to create the EuroGroups Register (EGR). This paper explores the use of web intelligence to improve the accuracy and completeness of the EGR, which makes use of tools for extracting and exploring information from the World Wide Web. Additionally, it presents a methodology to assess the quality of the information retrieved from the web, based on an *ex-post* comparison with the official information contained in the EGR. The results provide indications about the possibility of using web intelligence tools to support MNE group monitoring and to complete the EGR data in cases where this information is missing. Finally, it dives deeply into the approach to harness Wikipedia as a data source and into the techniques and methodology used to extract data from this source.

Keywords: Multinational enterprise (MNE) groups, statistical business registers (SBR), web intelligence, big data, wikipedia

1. Introduction

Economic globalisation is a process that refers to the increasing interconnection of national economies around the world. It has grown significantly in importance during the last decades because of political decisions, innovation, and technological evolutions. Multinational enterprise (MNE) groups play a significant role in this context and as a result have gained enormous importance in world trade and production. The proportion of MNE groups in business statistics represents a high percentage of the value added of the business economy in most countries [1,2,3,4]. Therefore, business statistics without sufficient information on MNE groups are weak, with a significant risk to the quality of information produced. Appropriate information on MNE groups requires official business statistics to record information about them beyond their national borders.

The production of high-quality business statistics depends largely on the quality of information available in the statistical business registers (SBR) that represent the backbone of statistical production for business statistics for all National Statistical Institutes (NSIs). Dealing with worldwide MNE groups is a challenge for all NSIs since they must capture as much information as possible from their non-domestic part. The national statistical business registers enable NSIs to deal with information extended to cross-border transactions, but not the complete group structure beyond national boundaries.

*Corresponding author: E-mails: alexandros.bitoulas@ext.ec.europa.eu; alexandros.bitoulas@sogeti.com.

To overcome the national limits, several initiatives have started over the last two decades worldwide. The EU Member States, the European Free Trade Association (EFTA) countries and Eurostat have collaborated to create the EuroGroups Register (EGR) [5]. The EGR is the statistical business register of MNE groups operating in the EU and EFTA countries.

The EGR is a structural component of European business statistics and the information it contains can be used for statistical purposes. Its users are the EU and EFTA National Statistical Institutes, their Central Banks, and the European Central Bank. The data are collected annually from national SBRs, and they are combined in the EGR using the cross-border relations between enterprises and each group's structure at national level. The EGR extends its information on MNE groups outside the EU with data from commercial data providers. Different statistical sets are being generated at 11, 13 and finally 15 months after the initial data collection when the final statistical set is produced. Each set is made available to EGR users, and the related statistics are disseminated after the final set is made available.

In the EGR, about 10% of MNE groups account for 90% of employment. This clearly indicates a high polarisation of the MNE groups recorded in EGR, with just a few very large ones dominating.

Due to the importance of the largest MNE groups, which have a significant impact on economic and business statistics, they need to be monitored regularly and to be updated frequently to provide users with highly accurate and up-to-date statistical information. Several NSIs have created specific functions or organisational structures to monitor the information received from large MNE groups and they use a profiling process to keep the information of the most significant MNE groups up to date. The aim of profiling is to understand the enterprise's business model and to translate this into a useful structure for statistical data collection. At the European Commission level, this activity is organised in line with the European business profiling recommendations manual [6].

As part of the EGR innovation process, various activities are scheduled for improving the accuracy and completeness of the data of MNE groups in the EGR. To this end, new public sources can provide support for monitoring any changes or modifications that may occur in these MNE groups, coupled with web intelligence methods.

Web intelligence is a relatively new area that makes use of tools for extracting and exploring information from the World Wide Web [7].

The key hypothesis that this paper examines is whether there is added value from public data sources and the use of web intelligence methods to complement missing values of key variables in the EGR. The paper describes a method to assess and rank public data sources with a view to selecting the most appropriate source per variable to complete any missing values in the EGR and to demonstrate how such sources can be used in official statistics. The core process is to define a quality ranking for each variable from the usable public sources. The ranking is based on an *ex-post* comparison between the public information and the information contained in the EGR. This ranking and the analysis of the various public information collected is key; the top-ranked variables could then be used both as an input source for the EGR, in the case of missing information, and as support information for EGR producers and profilers. The article finally sheds light on the approach to harness Wikipedia as a source of public data for Official Statistics.

This data framework can support EGR producers and profilers by feeding the data received into an information system module, where the information from various sources could be visualised and compared with the EGR data. Wherever they complete or improve the data quality, they can then be used to update the EGR.

The use of publicly available data could be particularly useful to produce open registers with unit-level data accessible to the public based on web intelligence.

2. Web intelligence approach

The ability to retrieve, store and release big data from and on the web allows both private companies and institutional bodies to explore this information. Public information, reorganised and focusing on specific business domains, is considered as smart data, and seen as a significant support for producing official statistics [8]. For this reason, smart data has been gaining increasing attention from researchers and statisticians over the last years. A series of papers and methodological analyses have been published, with relevant applications on economic and social aspects. In general, these applications use: financial markets high-frequency data [9], electronic payments data [10], mobile phone data [11], satellite image data [12], scanner price data [13], online price data [14], and online job advertisements [15].

In the European Statistical System (ESS), the challenges related to web intelligence is studied under the Trusted Smart Statistics [8] framework based on the

Web Intelligence Hub (WIH), which was recently established by Eurostat and the National Statistical Institutes (NSIs). The WIH offers a foundational framework for collecting content from the web and extracting data with the purpose to generate statistics. The WIH already collects and produces data on online job advertisements (OJA), its most advanced use case, co-developed with Cedefop [16]. In the case of MNE groups, an initial feasibility study 'Smart Data for MNEs' was published in 2021 [17]. This study highlighted possible sources of information from the web and the possibility of web scraping these sources for a selected number of MNE groups, mainly operating in EU and EFTA countries.

We initially focused on a restricted number of 200 MNE groups operating in the EU and EFTA countries, and included some whose headquarters were based outside the EU. The web intelligence process was carried out in two phases: in the first 'discovery' phase, public and open sources for MNE groups were investigated; in the second 'implementation' phase, an information database was built integrating all available data that were retrieved through one of the two main ways of web content acquisition when collecting web data: an application programming interface (API) service, if available, or with web scraping if the former was not available. APIs are contracts of services between two applications, which define how two applications can communicate with each other. Compared with web scraping, APIs are generally considered faster and more reliable for querying data from an application (or data source, in this case). Web scraping on the other hand retrieves a complete web page and extracts content from specific web features.

3. Discovery phase

The discovery phase identified the public sources to be used, looking at the availability of information on the control structure of the groups, their global group heads, the country of the global decision centre,[1] the main activity codes, the consolidated persons employed, and turnover and assets.

The discovery phase was carried out earlier by a feasibility study [17]. In this study, a pool of web sources was assessed (we name this process landscaping) in terms of the relevance of the available information, and as regards any technical limitations imposed by the source on retrieval of the content of interest. Seven web sources were finally selected:

- Wikipedia (https://www.wikipedia.org)
- Wikidata (https://www.wikidata.org)
- DBpedia (https://www.dbpedia.org)
- GLEIF (https://www.gleif.org)
- Open Corporates (https://opencorporates.com)
- PermID (https://permid.org)
- EDGAR (https://www.sec.gov/edgar.shtml)

The study assessed many sources according to their pros and cons. For the identification process, GLEIF (Global Legal Entity Identifier Foundation), Wikipedia, Wikidata and DBpedia were the most relevant ones. GLEIF was selected for the availability of both information on the legal entity identifier (LEI) register and the 'who owns whom' information. Wikipedia and Wikidata, projects owned by the Wikimedia Foundation, a non-profit organisation created to fund several wiki projects, were also selected as relevant sources. Wikipedia is a multilingual free online encyclopedia written and maintained by a community of volunteers through open collaboration; it is considered as a relatively stable and up-to-date source with a large variety of information. Wikidata is a collaboratively edited multilingual knowledge base and a common source of open data. It is a useful source of structured information regarding the key functions in MNE groups, headquarter geographical coordinates (which can be particularly useful for geo-maps), and unique identifiers (including LEI identifiers in some cases). Both Wikipedia and Wikidata provide several ways for extracting information, in particular they offer an API for free extraction, under the terms of the Creative Commons Attribution-Share-Alike licence.

Finally, DBpedia is a project started by the Free University of Berlin and Leipzig University in collaboration with OpenLink Software (https://www.openlinksw.com/); it aims to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web also through semantic querying. The source is particularly useful for retrieving information about the legal name, the number of persons employed and the URL of businesses.

Sources such as GLEIF, Wikipedia, Wikidata and DBpedia provided reference dates and some kind of time series for specific variables, which resulted in a more useful analysis of quantitative data such as number of employees or turnover. For qualitative data, the best structured information came from GLEIF and PermID.

[1] Unit where the strategic decisions referring to an enterprise group are taken.

4. Implementation phase

The MNE information database was created based on EGR key variables at group level: Global Decision Centre (GDC) and its country code, number of persons employed, turnover, assets, and number of associated legal units.

The database implementation process was designed for each key variable and was based on the three standard steps: extraction of each piece of data from the web source; transformation of the data to harmonise it with the EGR variables; and loading of the web-harmonised data into the database.

In the web content retrieval step, the website's availability and the stability of the structure of the data are crucial when using APIs or scraping techniques, as a change on the website or the API service can heavily affect the collection process. Each public source allowed for different coverage of the selected MNE groups. Even when information about a group is present in the public source, the stability of its availability is not guaranteed.

The transformation step first required a metadata reconciliation with the EGR variables, which was done first by variable mapping and then by variable recoding. This step included the linking of the extracted records of the extracted information with the proper EGR group identifier. In 5% of cases, a change in group name or country code was identified: in 2.5% of cases, this occurred without any need to change the group identifier, while in the other 2.5% of cases, a new group identifier was required.

The transformation step includes a quality evaluation of the data source for each variable considered.

In the loading step, all transformed, updated and ranked information is uploaded into the information database and made directly available in the business intelligence tool or directly integrated into the EGR.

5. Quality evaluation of the data sources

In the quality evaluation of the loading step, each value of each transformed variable was compared with the information in the EGR to measure the quality level. Following this, a ranking order of each variable source loading was established.

To set a quality indicator for our quantitative variables, we first determined the relative difference between the public data source and the EGR values. We then defined our final quality indicator as: the number of observations where the relative difference between the public variable and the EGR variable was in the range of $[-0.5, 0.5]$, divided by the total number of observations, i.e. cases with available data from the public source. The final information database was built based on an integration process that contained the information at group level from both the public sources and the official data from the EGR.

In what follows, the quality evaluation for each variable will be analysed and evaluated.

5.1. Country code of the group's global decision centre

The information on the country code of the Global Decision centre (GDC) was present in all the public sources used; Table 1 summarises the results for the main sources considered. In GLEIF, this information was available for 93% of the groups analysed, while for Wikidata and Wikipedia the rate was 53% and 88%, respectively. In 91% of the units, the country code of the GDC from GLEIF was the same as the EGR 2020 country code, while the rate for Wikidata and Wikipedia was 88% and 83%, respectively. Due to its relatively higher accuracy, GLEIF was thus assigned with the highest priority (quality) among the three sources examined overall.

The last row in Table 1 describes the integrated result obtained when using the information on the GDC country code from all the three sources. The integrated data delivers a slight increase of the coverage of the GDC country code to 94%.

In the integration process, we used for each variable the highest quality data first and then the lower ones, in sequence, whenever the MNE group's data were not present in the highest priority source. If the last available public source had no information, the data value was null.

5.2. Variables for number of persons employed, turnover and total assets

In the EGR, the three variables employment, turnover and assets are not always available. For each of these variables and sources, we applied the following methodology to calculate the quality indicator.

We first defined the Relative Difference of the Variable (*RDVar*) between the Public Variable value (*PVar*) and the EGR Variable value (*EgrVar*):

$$RDVar = (PVar - EgrVar)/EgrVar \quad (1)$$

Table 1
Sources considered for the country code of the Global Decision centre and integrated sources results

Public sources	Variable coverage	Same country code	Different country code	Priority level
GLEIF	93%	91%	9%	1
Wikidata	53%	88%	12%	2
Wikipedia	88%	83%	17%	3
Integrated sources	94%	91%	9%	–

Therefore, the final quality indicator was defined as the number of cases with a relative difference in the range of $[-0.5, 0.5]$, that is cases where the value provided by the public sources does not deviate by more than 50% from the EGR value, divided by the total number of Variable Coverage occurrences (*#VarC*):

$$\textit{Quality Indicator} = \textit{\#RDVar}[-0.5, 0.5]/(\textit{\#VarC}) \quad (2)$$

All occurrences of the public source with relative differences in the interval $[-0.5, 0.5]$ contribute to the numerator of the quality indicator: i.e. the values that are greater than half of the value of the corresponding EGR variable or smaller than 1.5 times the EGR variable.

The percentage of person employed that is not null in the 2020 EGR set is very high, equal to 98%. From the public sources Wikipedia, Wikidata and DBpedia, employment information is also present and easily usable, as shown in Table 2. In this table, 'variable coverage' is the ratio between the number of matched MNE groups with available information on the number of persons employed present in both the public sources and the 2020 EGR frame.

In Table 2, the percentages of overlapping MNE groups and variables when we integrated the sources are noticeably higher than the highest value of each single public source. This reflects that the public sources do not cover the same MNE groups and therefore the integrated data can optimise the coverage.

On the contrary, the quality indicator for all integrated sources is 43%, which is lower than the value of Wikipedia on its own (47%). This means that the coverage of the variable increased at the expense of the quality, with values outside the acceptable quality range. This may reflect the fact that the higher coverage of DBpedia with respect to Wikidata is offset by lower quality.

Figure 1 shows the number of persons employed according to the EGR versus the integrated public sources' values for each MNE group. The two straight lines are the range boundaries used for the employment quality indicator.

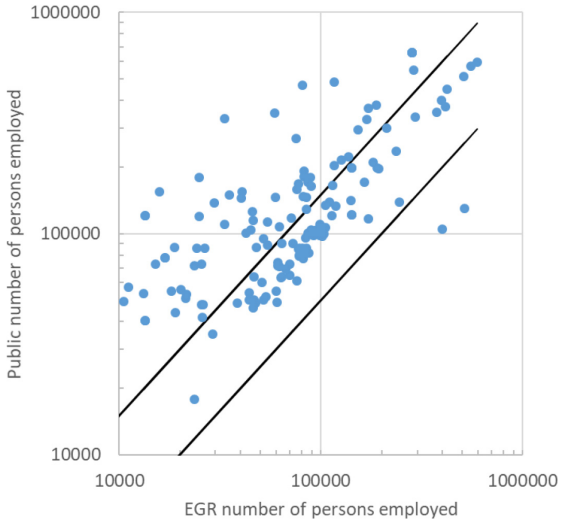

Fig. 1. Log-log scatter plot of number of persons employed, EGR versus public source.

From the scatter plot, it is evident that the scraped values tend to overestimate employment, compared with EGR values. It is not easy to interpret this upward shift. One likely reason could be that EGR employment is accurate for the MNE groups inside the EU (because data are provided by NSIs), but not for those outside the EU (where only commercial sources are used and coverage is partial). To prove this hypothesis, the full set of MNE groups should be split between those fully operating in the EU market and the others, to assess a possible reduction of the shift in the former group. This could be a future quality analysis on MNE groups.

Concerning the variable turnover,[2] which must be the consolidated value excluding intra-group sales, the completeness rate in the EGR is only 16% in total. It was possible to retrieve information only from DBpe-

[2]It comprises the total invoices of the MNE group during the reference period, and this corresponds to market sales of goods or services supplied to third parties. Turnover also includes all other charges (transport, packaging, etc.) passed on to the customer, even if these charges are listed separately in the invoice. Turnover excludes VAT and other similar deductibles as well as all duties and taxes on the goods or services invoiced by the unit.

Table 2
Number of persons employed in public sources and integrated results

Public sources	Retrieved groups	Variable coverage	Quality indicator	Priority level
Wikipedia	71%	72%	47%	1
Wikidata	54%	39%	43%	2
DBpedia	62%	64%	42%	3
Integrated sources	83%	86%	43%	–

Table 3
Sources considered for the turnover and integrated sources results

Public sources	Retrieved groups	Variable coverage	Quality indicator	Priority level
DBpedia	46%	16%	79%	1
Wikipedia	73%	18%	67%	2
Integrated sources	78%	20%	69%	–

dia and Wikipedia; Table 3 shows the percentage of information retrieved.

According to the same quality indicator, defined as for employment, DBpedia was emerging as the first priority source and Wikipedia as the second despite Wikipedia having a higher coverage than DBpedia. When we integrated the sources, it was possible to obtain information on the turnover for 78% of the MNE groups, out of which 20% covers the information about turnover values in the EGR. The quality indicator for the integrated turnover was 69%. Like the number of persons employed, this value was lower than the maximum quality value obtained from DBpedia on its own, reflecting that the source with the highest coverage on turnover had a lower quality.

The analysis of turnover shows that for 65% of the MNE groups for which EGR data are not available, it is possible to obtain this information from the public sources.

Concerning the total assets variables, which comprises the total economic resources controlled by MNE groups, the overall conclusion is like that for turnover. In the EGR, data on total assets are rarely available. For the considered sample of MNE groups, the EGR total assets variable is available only for 10% of cases, whereas from the public sources it was possible to retrieve information only from Wikipedia and DBpedia.

Table 4 shows the percentage of MNE groups retrieved from each public source. The variable overlap is very small due to missing values and DBpedia is defined as the first source.

The integrated source row in Table 4 shows that the percentage of retrieved MNE groups is higher than the value of each single public source. The variable overlap remains 13% and the quality indicator increases slightly. Naturally, these values are influenced by the small number of observations and therefore are subject to high variability even for small variations. The analysis should be repeated with a wider set of MNE groups.

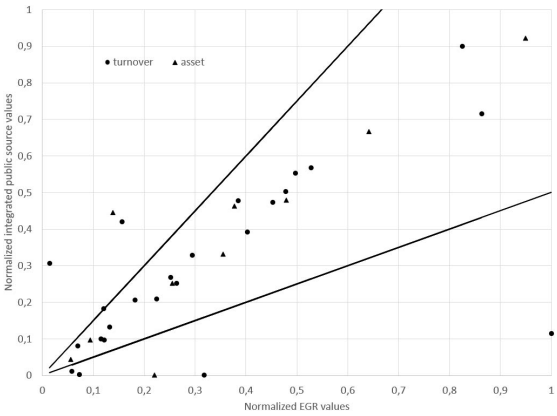

Fig. 2. Normalised scatter plot of turnover and asset EGR versus public source.

The analysis on the asset variable shows that, for 53% of the MNE groups, it is possible to obtain information from public sources when the corresponding values in the EGR are not available.

Figure 2 shows the scatter plot of the EGR versus the integrated public source values for turnover and total assets values. Each dot represents a group with its normalised values and the two straight lines identify the range boundaries of the quality indicator.

In the scatter plot, the overall limited presence of usable information is evident, even if it is possible to recognise an acceptable fitting of the integrated public sources with the EGR for both variables, which shows that the quality indicator of the integrated sources is high (Tables 3 and 4). Unlike the persons employed variable, in these cases there are no systematic data shifts but, instead, dots outside the quality range lines are randomly distributed. The results show that turnover and total assets retrieved from public available sources,

Table 4
Sources considered for the asset and integrated sources results

Public sources	Retrieved groups	Variable overlap	Quality indicator	Priority level
DBpedia	41%	6%	80%	1
Wikipedia	49%	13%	75%	2
Integrated sources	60%	13%	82%	–

when available, are of acceptable quality and could be used to replace the missing values in the EGR.

To overcome the issue regarding the assessment of the public data sources due to the unavailability of EGR data, like for the cases of the economic variables (turnover and total assets), we have explored the official source, EDGAR [18]. The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database system is an online system operated by the U.S. Securities and Exchange Commission (SEC). It is currently the primary repository for electronic filings of financial documents submitted by publicly traded companies in the US stock exchanges and other entities required to file reports with the SEC.

In the cases for which EGR information is incomplete, EDGAR can serve as a benchmark to evaluate the economic variables collected from public sources.

The benefits from using EDGAR rely mostly on the official status of the source, which is collected by the SEC. However, EDGAR focuses mostly on US companies while the EGR has a focus on EU-EFTA MNE groups.

There are several cases in which the two systems overlap. Most of these cases are very large MNE groups that are listed in several stock exchanges around the globe. In numbers, from the cases in which we were able to match the information in EDGAR and the EGR, we found that more than 80% of the cases are MNE groups based in the US, while only around 5% were based in an EU-EFTA country. This matching is done using the Levenshtein string distance algorithm; a total of 1 274 multinational enterprise groups can be matched between EDGAR and the EGR. Out of them, 886 are perfect matches (i.e. 100% string matching), while the others are partial matches that require additional input. Using the information from EDGAR in these 886 cases, a total of 751 values for turnover and 842 values for total assets can be collected to complement the data in the EGR.

5.3. Number of legal units in the group

The last point analysed is the perimeter of a group, i.e. the list of legal units managed directly or indirectly by the global group head, as well as the hierarchical structure. We have treated this variable separately from others because the conclusion is radically different.

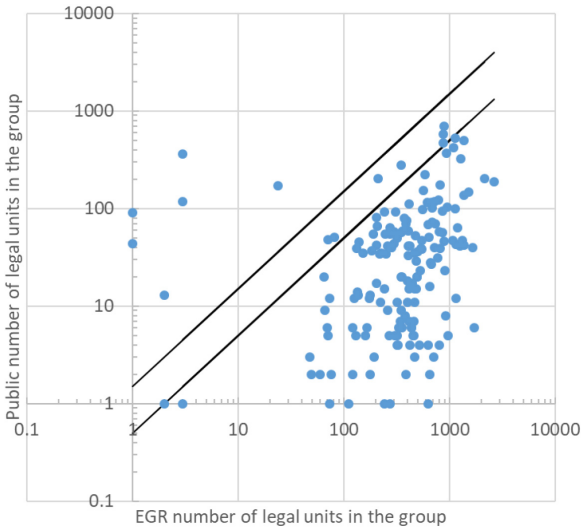

Fig. 3. Log-log scatter plot of number of legal units in the group, EGR versus public source.

Table 5 shows the information retrieved from each public source on the number of legal units. Information about the structure of the group can be found only in GLEIF and Wikidata. In GLEIF, we found very good coverage as there is information on the legal units owned by the MNE group in 93% of cases. In Wikidata, it was possible to find information in 80% of cases, which is also an acceptable coverage. On the contrary, the values of the quality indicator for the number of legal units are very scarce for each source. However, the GLEIF source is better than Wikidata and is used as the first source in the integration process.

The integrated source row shows the possibility of obtaining information on the number of legal units for 93% of cases. The quality indicator of integrated sources is 7%, which is the lowest value obtained from the analysis. This small value indicates a bad matching with the public sources, probably due to the complexity of the information.

Figure 3 shows the scatter plot on the number of legal units in the selected MNE groups of the EGR versus the integrated sources. It illustrates that the public sources tend to underestimate the number of legal units com-

Table 5
Sources considered for the number of legal units in the group and integrated sources results

Public sources	Retrieved groups	Quality indicator	Priority level
GLEIF	93%	7%	1
Wikidata	80%	3%	2
Integrated sources	93%	7%	–

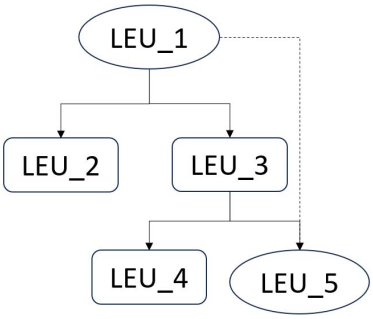

Fig. 4. Group structure example.

pared with the EGR. The main reason could be related to the more accurate official sources used by the EGR.

In any case, the comparison of the structure is a complex task. A missing and/or additional relationship from the integrated public sources with respect to the EGR might make the task difficult.

In this part, the analysis on the legal units was extended to MNE group structures to try to understand the differences on the number of legal units, but a more detailed analysis is beyond the scope of this paper. At this stage, we would like to provide an interpretation of the mismatch by using the information found on GLEIF.

Firstly, GLEIF can be used as a source for an additional legal entity identifier (LEI) beyond already existing records in the EGR. This information may be useful to NSIs, to identify the legal units involved in their relationships. This is the main goal of GLEIF: to provide a unique identifier to any legal entity. The use of LEIs in all registers will improve the treatment of data in the EGR; it may happen that some relationships known by NSIs cannot be included in the EGR due to the absence of any matching with NSI databases.[3] In this way, some information is lost; that could be solved with LEI.

Secondly, the comparison between the EGR and GLEIF raises the issue of coverage.

Figure 4 shows a typical mismatch, where the coverage of the EGR is higher than the coverage of the integrated sources. The GDC information is confirmed by the information available in the EGR, but the perimeter is hugely different between the two sources. This can imply that the coverage of GLEIF is extremely limited, compared with the EGR.

The ellipses represent legal units known by both sources; squared boxes are those known only by the EGR. The dashed lines are known only by GLEIF, while the solid lines are known only by the EGR.

To give more details about Fig. 4, we briefly compare GLEIF data and EGR data in terms of unit coverage. The GLEIF database contains 2.5 million units[4] (compared with 1.6 million in the EGR[5] [19]) of which 1.5 million have a usable identifier. We focused on EU/EFTA and UK units (1.4 million) for which the comparison between a unit in the GLEIF database and one in the EGR is more reliable.[6] 75% of these units are known from the EGR, but, according to the EGR, only 11% are part of a group operating in the EU. GLEIF provides information about relationships in the second part of its database ('who owns whom'); we deduce that GLEIF contains around 43 000 groups, 28 000 of which operate in the EU/EFTA. Using this information, we find around 300 000 units, of which 160 000 located in the EU or EFTA, involved in a group operating in the EU according to GLEIF. A more detailed analysis must be carried out to determine control relationships known to GLEIF and unknown to the EGR, in particular for the non-EU part.

6. The approach to harness Wikipedia as a data source on MNE groups for official statistics

In the following section we present our approach to harnessing Wikipedia as a data source on MNE groups

[3]EuroGroups register identification service - Statistics Explained (europa.eu) – https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EuroGroups_register_identification_service.

[4]Reference snapshot: July 2023.

[5]The number of 1.6 million units includes legal units in foreign-controlled groups (schematically, a unit for which only the country of the controller is known).

[6]Matching based on country and name only is not considered in this paragraph in order to obtain more accurate results. Some identifiers provided by GLEIF could not be used either because they do not meet the national rules for identifiers or because they are identifiers not available in the EGR; further analysis is needed to use these identifiers.

for official statistics. We discuss the methodology used to identify and match Wikipedia articles with the MNE groups in our sample, and how to collect and process the content from Wikipedia to prepare the final dataset for statistical purposes. We also discuss methodological and technical challenges, and insights gained from using Wikipedia as a data source on MNE groups.

6.1. The Web Intelligence Hub platform

To achieve its goals, the WIH has developed a platform where statisticians from statistical authorities of the ESS can easily run web crawlers to collect data from the web. Those crawlers can be configured through a graphical user interface (GUI), where the user can provide the URL (website) to be scraped or queried, together with other parameters and launch the collection of content of that website. The user can configure, among other parameters, whether the website being crawled is dynamic (in which case Selenium is used) or not, the interval between data acquisitions (e.g. in case of retries due to errors), text extractor parameters like regular expressions (regex), and how deep in terms of subpages the crawler will crawl to search for data. The GUI of the WIH platform can be seen in Figs 5 and 6.

6.2. Methodology of content retrieval from Wikipedia

Wikipedia offers various APIs to query the content of its articles and can return the content in various formats, with JSON being the recommended one. For example, the query to the English Wikipedia API[7] would return the content of the English Wikipedia article with title 'Siemens' (https://en.wikipedia.org/wiki/Siemens) in JSON format. This returns the raw content of the Wikipedia article as written in Wikitext, Wikipedia's own markup language to edit and format pages. An example of the raw content of a Wikipedia article, in Wikitext format is presented in Fig. 7.

We used the WIH platform to do the querying and to acquire the content, which we then saved to an OpenSearch [20] index. An example of the raw content from Wikipedia for an MNE in an OpenSearch index is seen in Fig. 8.

[7] https://en.wikipedia.org/w/api.php?action=query&prop=revisions&rvprop=content&format=json&titles=Siemens&rvslots=main.

Table 6
Wikipedia categories, used to identify articles on MNEs

Wikipedia categories
Conglomerate_companies_by_countr
Companies_listed_on_the_New_York_Stock_Exchang
Multinational_companies_by_country
Corporations
Multinational_companie
Companies_by_country
Banks_under_direct_supervision_of_the_European_Central_Ban
Companies_in_the_Euro_Stoxx_50
Companies_by_stock_exchang
Financial_services_companies_by_year_of_establishmen
Government-owned_energy_companie
1991_establishments_by_country
Manufacturing_companies_by_country
Engineering_companies_of_Germany

6.3. Methodology of matching Wikipedia articles with names of MNE groups

One of the challenges we faced early in the project was the identification of the Wikipedia articles that correspond to the MNE groups in our sample, from now on *EGR MNE*.

According to Wikipedia itself, 'as of 12 February 2024, there are 6 782 641 articles in the English Wikipedia'.[8] This is a very large number of articles to be matched to a potentially large number of MNE groups. To facilitate this matching, we used Wikipedia's categorisation to identify articles about MNE groups. We used in total 14 such categories, which can be seen in Table 6.

We then queried the names of all Wikipedia articles under these categories and their subcategories, using Wikipedia's API service. This returned a pool of around 300 000 articles related to MNE groups or companies from the English Wikipedia.

We then found the closest match of the Wikipedia articles and the respective names of the MNE groups in our sample, the EGR MNE groups. Our goal was to identify the Wikipedia article that referred to the *parent company* or the *group head*, i.e. the enterprise that owns and manages all subsidiary companies that comprise the corporate group.

We followed a heuristic approach to match the Wikipedia articles with the EGR names in our sample. We used a semi-automatic identification using R's stringdist [21] library, followed by a manual validation of the final articles for our MNE groups. We tried various distance metrics to calculate the distance be-

[8] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

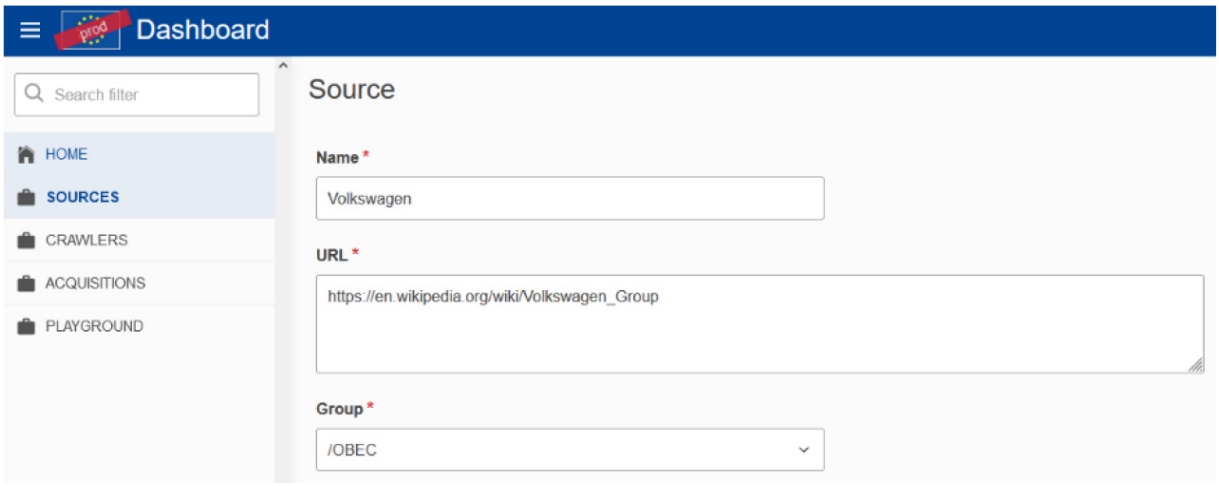

Fig. 5. The WIH Platform, creation of a source.

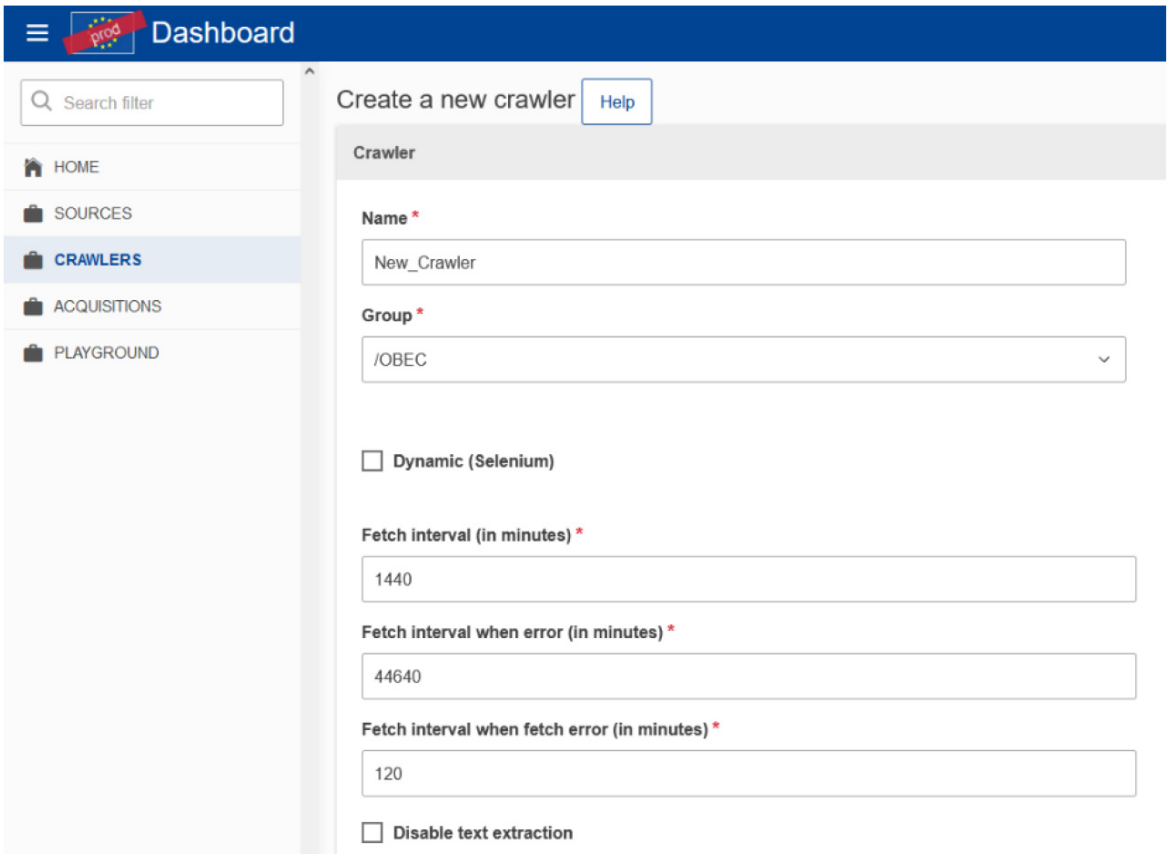

Fig. 6. The WIH Platform, creation of a new crawler.

tween the two names (Wikipedia name vs EGR name), namely Jaro-Winkler, Optimal String Alignment, Levenshtein, and q-gram with 2 and 3 grams. The method that worked best for our data was the Jaro-Winkler method, with a prefix value of $p = 0.2$. With this method we identified and matched accurately around 50% of the Wikipedia articles. To increase the final matching accuracy, we further checked and validated

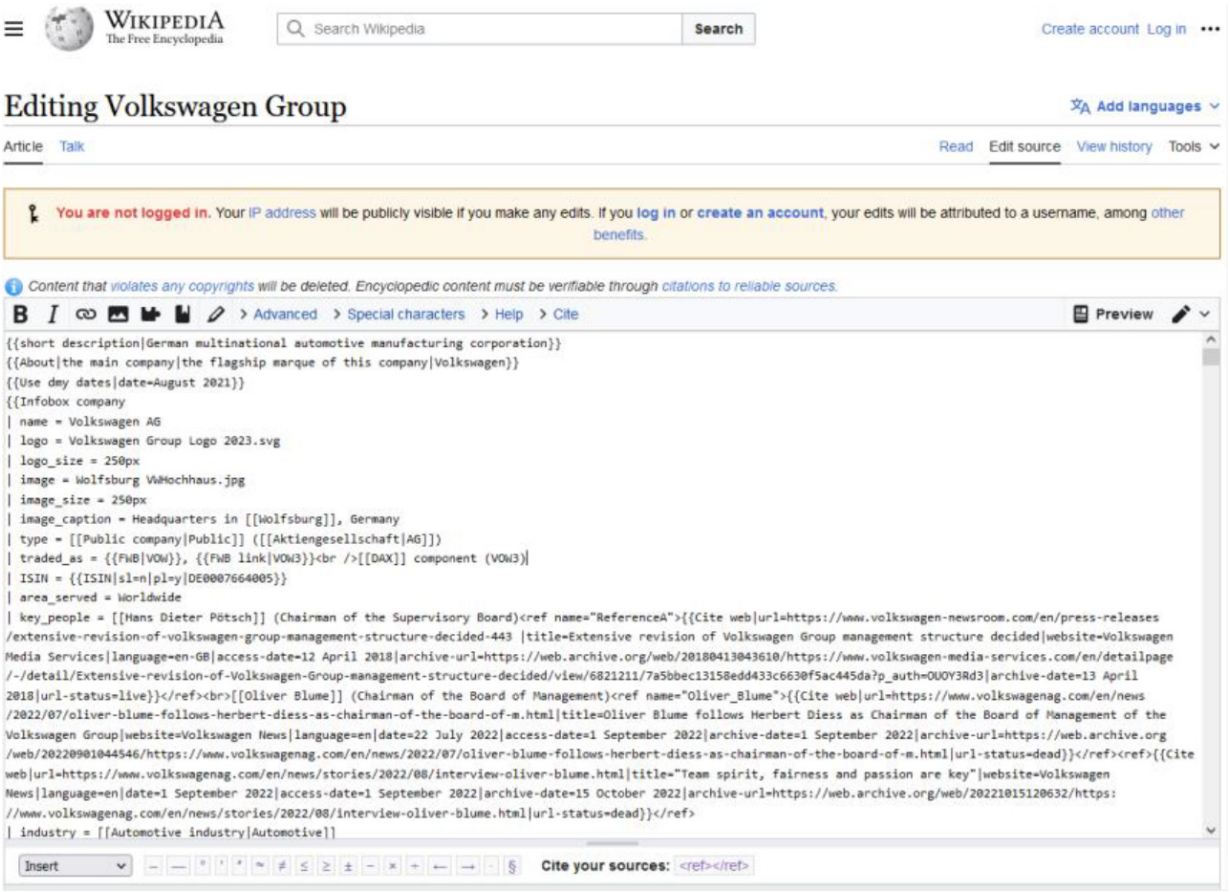

Fig. 7. A Wikipedia article in Wikitext markup language (edit mode).


```
"_index" : "",
"_type" : "_doc",
"_id" : "Fuxhe40Bndz2i3Z-EONJ",
"_score" : 1.0,
"_source" : {
  "id" : "a548aa706ac9e5b37fc2fb267e45b5171f813ebd9df0c80fa20abd449a7dafc3",
  "url" : "https://en.wikipedia.org/wiki/Volkswagen_Group?action=raw",
  "fetched_date" : "Thu, 09 Nov 2023 14:20:05 GMT",
  "data" : {
    "name" : "Volkswagen AG",
    "ISIN" : "{{ISIN|sl=n|pl=y|DE0007664005}}",
    "area_served" : "Worldwide",
    "industry" : "[[Automotive industry|Automotive]]",
    "products" : "[[Automobiles]], [[commercial vehicle]]s, [[internal combustion engines]], [[motorcycles]], [[turbomachinery]]",
    "services" : "[[Banking]], [[financing]], [[fleet management]], [[insurance]], [[leasing]]{{Sfn|Volkswagen AG|2012|p=110}}",
    "revenue" : """{{increase}} [[Euro|€]]250.200 billion (2021)<ref name="AR21a" />""",
    "operating_income" : """{{increase}} €20.126 billion (2021)<ref name="AR21a" />""",
    "net_income" : """{{increase}} €15.428 billion (2021)<ref name="AR21a" />""",
    "assets" : """{{increase}} €528.609 billion (2021)<ref name="AR21a" />""",
    "equity" : """{{increase}} €144.449 billion (2021)<ref name="AR21a" />""",
    "owners" : """* [[Porsche SE]]: 31.4% of equity, 53.3% of votes (as of 31 December 2021)<ref name="AR21a" />,
    "num_employees" : """{{increase}} 667,647 employees (average during 2021)<ref name="AR21a"/>""",
    "founder" : "[[German Labour Front]]",
    "location_city" : "[[Wolfsburg]], [[Lower Saxony]]",
    "foundation" : "[[Berlin]], Germany<br>({{Start date and age|1937|05|28|df=yes}})",
    "location_country" : "Germany",
    "locations" : "[[list of Volkswagen Group factories|100 production facilities across 27 countries]]",
    "homepage" : "{{URL|volkswagen-group.com}}"
```


Fig. 8. Sample of a raw content of a Wikipedia article as stored in an OpenSearch index.

Table 7
Data extraction and decomposition

Data extraction and decomposition		
Assets	Assets_currency	Assets_year
15 428 million	€	2021

manually the matched articles. For those MNE groups for which the match was inaccurate we identified the right article manually by searching Wikipedia.

This validated dataset of Wikipedia articles now serves as a benchmark for improving our string-matching algorithm in the future. All the above analysis was performed using R statistical software [22].

6.4. Content extraction

Wikipedia has a specific template for presenting information boxes in articles of companies. In this template, called 'infobox company', editors add information (variables) related to companies in structured fields. Using this information, Wikipedia then provides an infobox in the article.

We identified that all variables of our interest are included in this template, namely revenue, net income and website, among others. The infobox company template constitutes a quasi-structured environment for the data of our interest, where parameters (or variables) of that template are set. For example, the parameters 'revenue' and 'revenue_year' of the infobox company template refer to the financial concepts of revenue and its respective reference year, while 'num_employees' refers to the number of employees. For a complete list of variables and their definition on Wikipedia's infobox company template, see: https://en.wikipedia.org/wiki/Template:Infobox_company/doc.

Having queried the complete content of a Wikipedia article as described previously, we then extract the part of the Wikipedia article that contains only the infobox company template. We further filter and select the content of the variables in scope. Until this point, we have not modified the raw content of Wikipedia. From this point onwards though, we break down the information available in the infobox company template to relevant data and metadata of our interest. For this step, we used the wikitextparser [23] library of Python [24] and we further developed our own parsing routines in Python to extract the relevant pieces of information.

For example, from the string: 'assets = {{increase}} €15 428 billion (2021)<ref name = 'AR21a' />' we extract the following variables (Table 7).

Similar processing is performed for all the variables of interest.

After this step, we proceed to the final formatting or standardisation of the data. In this step, we format the values of certain variables to comply with the Eurostat standards and code lists. Here we format currencies according to the Eurostat code list 'currency', (e.g. '€' is formatted to code 'EUR') and we format the values of the economic variables to full numbers (e.g. 15 428 million is formatted to 15 428 000 000).

6.5. Final statistical data asset

To further comply with ESS standards, we export the final data in an SDMX-CSV format. This format respects the SDMX[9] recommendations and best practices on the structure of the dataset, the naming conventions of variables and the use of codes from standard code lists used by Eurostat and the ESS.

The overall implemented pipeline to extract and process the content from Wikipedia can be seen in Fig. 9.

6.6. Lessons learnt – possible improvements

Our research revealed that around 80% of the MNE groups in our sample had a corresponding article on English Wikipedia. The main economic variables are largely available in the structured infobox company template. Inconsistencies in how Wikipedia editors enter data in the infobox company template posed challenges in the parsing of the content into appropriate values, thus sometimes undervaluing the quality of the data.

Very often, more than one Wikipedia article exists for the same MNE group. For example, a search of the term *'Volkswagen'* on the English Wikipedia returns at least two articles referring to the company Volkswagen or to the group.[10] A similar search of the term 'Volkswagen Wikipedia' in Google, returns the former URL article as the first result. However, in the above example, our target article would be the latter one, as it refers to the group head or parent company of Volkswagen.

Often the EGR name was quite different from the respective name of the MNE group on Wikipedia. This discrepancy can be due to one of the two names being outdated, or an abbreviated version of the full name being available, e.g. 'Volkswagen' versus 'VW', or a

[9] https://sdmx.org/.

[10] https://en.wikipedia.org/wiki/Volkswagen and https://en.wikipedia.org/wiki/Volkswagen_Group.

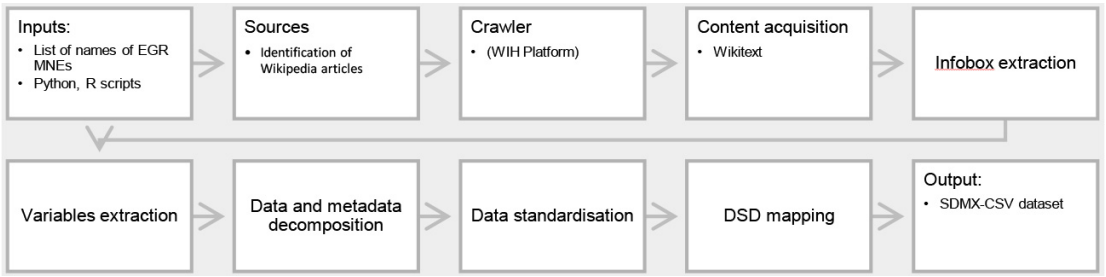

Fig. 9. Pipeline to extract and process content from Wikipedia.

difference between the legal name of an EGR MNE group and its branding name.

Lastly, Wikipedia pages exist in various languages. For some MNEs we identified Wikipedia articles in a non-English language. However, since our focus was on the English Wikipedia only, it meant that we had to leave out any Wikipedia article not appearing in English Wikipedia. Furthermore, the presence of MNE group articles in non-English languages would call for translation or variable mapping.

Our future work will focus on refining the data extraction algorithm and exploring the potential of non-English Wikipedia content. We also envisage to develop quality gates (checks) during the content extraction to ensure a high quality of the data. Another development may be in the direction of monitoring if Wikipedia articles are edited or updated to collect new data.

7. Conclusion

Based on the analysis of the results, this study concludes that public sources can be used as an additional source of information for the support of users in improving the quality of data of the MNE groups. Moreover, public sources can be considered when complementing missing EGR information on MNE groups, but their contribution needs to be precisely qualified.

Regarding most of the attributes of an MNE group, the gain seems to be positive for the country of the Global Decision centre, turnover and total assets. The conclusion on employment is more moderate and further analysis is needed to understand the employment gap, which could be attributed to the lack of coverage of information from outside EU/EFTA in the EGR. If confirmed, the employment data from the public sources could well complement any missing data in the EGR for the countries outside the EU.

7.1. Further work

The study and the quality analysis carried out was based on a limited number of MNE groups only. Following the positive assessment of the data acquired from the proof-of-concept study, Eurostat started a web data collection for the whole population of the biggest MNE groups (more than 1600 groups) hosted under its Web Intelligence Hub. This was to extend the coverage and verify the possibility to implement the results for the purpose of the production of EGR data. This work is ongoing and the first dataset from Wikipedia for the entire population of the biggest MNE groups is expected during 2024. At the same time, Eurostat is also working on using the data from EDGAR and GLEIF, connecting through an API.

In parallel, Eurostat is working on a proof of concept for a data visualisation and comparison module of the EGR, where the data of the public data sources will finally be uploaded. They will then be compared automatically with the EGR data, and the results will be made available in a specific dashboard. There, the EGR producers or profilers will be able to visualise and compare the data from public data sources with the EGR data and decide to use the data that they consider will improve the completeness and accuracy of the EGR MNE group data.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily represent the official position of the European Commission.

References

[1] Gereffi G. The organization of buyer-driven global commodity chains: How US retailers shape overseas production networks. In: Global Value Chains and Development. Cambridge University Press. 1994; 95-122. doi: 10.1017/9781108559423.003.

[2] Gereffi G. Global value chains in a post-Washington Consensus world. Review of International Political Economy. 2014; 21(1): 9-37. doi: 10.1080/09692290.2012.756414.

[3] Gereffi G, Fernández-Stark K. Global Value Chain Analysis: A Primer. 2011; 3-34. Available from https://www.globalvaluechains.org/wp-content/uploads/Primer_1stEd_2011.pdf.

[4] Menghinello S, Faramondi A, Laureti T. The future role of official statistics in the business data arena. Statistical journal of the IAOS. 2020; 36: 519-533. Available from https://api.semanticscholar.org/CorpusID:214116981.

[5] Bikauskaite A, Götzfried A, Völfinger Z. The EuroGroups Register. Statistika: Statistics and Economy Journal. 2019; 99(1): 69-76.

[6] Eurostat. European Business Profiling, Recommendations Manual. 2020 edition'; 2020. Available from https://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-GQ-20-002. doi: 10.2785/110879.

[7] Bianchi G, Laureti Palma A, Quaresma S. Prepare your Data Warehouse for a Big Future, by including Big Data. In: European Conference on Quality in Official Statistics 2018. Kraków, Poland; 2018.

[8] Ricciato F, Wirthmann A Giannakouris K, Reis F, Skaliotis M. Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS. 2019; 35(4): 589-603. doi: 10.3233/SJI-190584.

[9] Degiannakis S, Floros C. Introduction to High Frequency Financial Modelling. In: Modelling and Forecasting High Frequency Financial Data. London: Palgrave Macmillan UK. 2015; 1-23. doi: 10.1057/9781137396495_1.

[10] Galbraith JW, Tkacz G. Analyzing Economic Effects of Extreme Events Using Debit and Payments System Data. CIRANO – Scientific Publications; 2011; 2011s-70. doi: 10.2139/ssrn.1963812.

[11] Smith-Clarke C, Mashhadi A, Capra L. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. Conference on Human Factors in Computing Systems – Proceedings. 2014 April; 511-520. doi: 10.1145/2556288.2557358.

[12] Henderson V, Storeygard A, Weil DN. A Bright Idea for Measuring Economic Growth. American Economic Review. 2011 May; 101(3): 194-199. doi: 10.1257/aer.101.3.194.

[13] Silver M, Heravi S. Scanner Data and the Measurement of Inflation. The Economic Journal. 2001 December; 111(472): 383-404. doi: 10.1111/1468-0297.00636.

[14] Cavallo A. Are Online and Offline Prices Similar? Evidence from Large Multi-channel Retailers. American Economic Review. 2017 January; 107(1): 283-303. doi: 10.1257/aer.20160542.

[15] Ascheri A, Nagy AMK, Marconi G, Mészáros M, Paulino R, Reis F, et al. Competition in Urban Hiring Markets: Evidence from Online Job Advertisements: 2021 Edition. Statistical working papers/Eurostat. Publications Office of the European Union; 2021. doi: 10.2785/667004.

[16] Descy P, Kvetan V, Wirthmann A, Reis F. Towards a shared infrastructure for online job advertisement data. Statistical Journal of the IAOS. 2019; 35(4): 669-675. doi: 10.3233/SJI-190547.

[17] Eurostat. Smart Data for Multinational enterprises (MNEs) – using open-source data to obtain information on Multinational enterprises. Eurostat; 2021. Available from https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-tc-21-007. doi: 10.2785/415398.

[18] EDGAR the Electronic Data Gathering, Analysis, and Retrieval system. Available from https://www.sec.gov/edgar.shtml.

[19] Eurostat. Structure of multinational enterprise groups in the EU; 2024. Available from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Structure of multinational enterprise groups in the EU.

[20] OpenSearch, distributed search and analytics service. Available from https://opensearchorg/.

[21] van der Loo MPJ. The stringdist Package for Approximate String Matching. The R Journal. 2014; 6(1): 111-122. doi: 10.32614/RJ-2014-. 011.

[22] Core Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria. Available from https://www.R-project.org.

[23] Wikitextparser. A Wikitext Parsing Library for MediaWiki. Available from https://pypi.org/project/wikitextparser/.

[24] Python Software Foundation. Python Language Reference, version 311. 2023. Available from https://www.python.org/.