Group of Experts on Business Registers

Online meeting, 26 – 29th September 2022

Alexandre Depire, Antonio Laureti Palma
Eurostat

Session 2: The use of administrative data, big data, and other data sources

Web Intelligence on multinational enterprise groups data

Abstract

Economic globalisation creates new opportunities for businesses to organise their production chains using multinational enterprise (MNE) groups. In this context, following the highly volatile evolution of the dynamics of such global organizations is extremely challenging for the statistical business registers that need to support the production of macroeconomic and business statistics with the highest possible quality.

The EGR is the European statistical register on MNE groups with limited resources available, and growing requests for better quality, higher completeness, and improved timeliness, as well as requests for new statistics, the well-established process of EGR has to be adapted to continue delivering on new user requirements in innovative ways.

In search for additional data sources that could be used to increase the coverage of EGR, especially for the extra-EU populations, new technologies and tools, like web intelligence and web scrapping are being explored. The information available in the World Wide Web can be analysed to verify its potential use for EGR.

To explore these capabilities and answer to the assumption that these technologies and the available public data can be used for EGR, a pilot study has been set up by Eurostat. The study delivered a set of data web scraped from public sources for a limited set of groups during the end of 2020.

This paper presents a methodology to assess the quality of the information retrieved with the pilot study based on an ex-post comparison with the information contained in the EGR. The results provide some preliminary indications about the possibility to use web scraped information to complement EGR data.

Keywords: Multinational enterprise groups, Statistical business registers, web intelligence

1. Introduction

Economic globalisation creates new opportunities for businesses to organise their production chains using multinational enterprise groups. In the last twenty years, the globalization has grown greatly as a consequence of political decisions and technological evolutions, creating challenges to official business statistics to record organisations and activities beyond their national borders.

The production of high-quality business statistics on economic globalisation depends to a large extent on the quality of the information available in the statistical business registers for the Multinational Enterprise (MNE) groups. Delineating a MNE group correctly is a challenge for all NSIs, as they tend only to capture the domestic part of it and its cross-border transactions, but not the complete structure beyond the national boundaries.

The EGR is the European statistical registers on MNE groups created by the European Statistical System and managed by Eurostat. It receives input data from the National Statistical Institutes (NSIs) of the European Union Member States and EFTA countries and a commercial data provider, consolidates them and makes it available for statistical purposes [1].

The EGR needs to continuously adapt to new user requests for improved timeliness, higher completeness, and better quality. With limited resources different options are being explored, and a strategy set by the European Statistical System is being followed.

Among the different solutions, and in the context of trying to find innovative alternatives, one is to improve quality and completeness of EGR using public available data from the World Wide Web and implementing new techniques to gather and process them. One of them is Web Intelligence which is a relatively new area of scientific research that makes use of big data tools for extracting and exploring information from the World Wide Web [2].

This paper focuses on the results of a pilot study run by Eurostat to investigate the possibility to use public data source through Web Intelligence to build up the variables of interest for EGR. This approach defines a quality ranking for each variable considered from each of the identified public source, based on an ex-post comparison with the information contained in the EGR. In this way, the variables from each source are ranked based on their quality and could be used to build a new input source on MNE groups for EGR.

2. Web intelligence process on public sources

The pilot study 'Smart Data for MNEs' [3] analysed the challenges related to Web Intelligence on a selected number of MNE groups mainly operating in EU and EFTA countries and including some whose headquarters is based outside the European Union.

The study was developed in two phases: in the first phase (discovering phase), public and open sources for MNE groups were investigated and in the second one (implementation phase), an information database (DB) was built using public data obtained from the web.

This paper focuses on the quality evaluation of the results. For each MNE group, each transformed variable was compared with the information in EGR to evaluate the quality level and to decide the ranking order of the source for the DB.

The overall study focused on a sample of 200 MNE groups, extracted from EGR, for which public information was extracted from the web at the end of 2020. Only 6% of the sample was not available from the web.

3. Discovering phase

The first phase identified the public sources that could be used for MNE groups, looking at information on the control structure of the groups, their global group heads, the country of global decision centres, the main activity codes, the consolidated persons employed, turnover and assets.

The study assessed a large number of sources according to their pros and cons. For the identification process, GLEIF[1], Wikipedia, Wikidata and DBpedia resulted the most relevant ones. GLEIF was used

[1] Established by the Financial Stability Board in June 2014, the Global Legal Entity Identifier Foundation (GLEIF) is tasked to support the implementation and use of the Legal Entity Identifier (LEI). The foundation is backed and overseen by the

both for information on the LEI register and for the 'who owns whom' information. Wikipedia and Wikidata, projects owned by the Wikimedia foundation, a non-profit company created to fund several wiki projects, were also used as sources. Wikipedia is a multilingual free online encyclopaedia written and maintained by a community of volunteers through open collaboration and it is considered as a relatively stable and updated 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 also regarding the key functions in the MNE groups, headquarters geographical coordinates (which can be particularly useful for geo maps), and unique identifiers (including the LEI identifiers in some cases). Both Wikipedia and Wikidata allow several ways for extracting information, in particular they offer an API for free extraction, under the terms of the Creative Commons Attribution Share-Alike license.

Finally, DBpedia is a project started by the Free University of Berlin and Leipzig University in collaboration with OpenLink Software and 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 query. The source is particularly useful for retrieving information about the legal name, number of persons employed and URL of businesses.

4. Implementation phase

The information Database (DB) was created on the basis of EGR key variables at group level: ultimate controlling institutional unit (UCI) and its country code (CC), number of persons employed, turnover, assets and number of associated legal units[2].

The DB implementation process was articulated for each key variable and was based on the three standard steps: extraction of each piece of information from the web sources, transformation of the information to reconcile it with the EGR variables, loading of the reconciled information into the DB.

In the extraction step, the provider's availability is 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 a group is present in the public source, there is no guarantee that all information will be stably present.

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 record linkage of the extracted information with the proper EGR group identifier. In 5% of the cases, a change in group name or country code was identified: in 2.5% of them, this happened without group identifier changes, while in the other 2.5% of cases, a new group identifier was required.

The loading step needed first a quality evaluation of the data source for each variable considered and then the proper loading process into the DB based on the identified level of quality of each source associated with the variable.

Regulatory Oversight Committee, representing public authorities from around the globe that have come together to jointly drive forward transparency within the global financial markets.

[2] The ultimate controlling institutional unit is the institutional unit at the top of a chain of control that is not controlled by another institutional unit. The number of persons employed is the total number of persons who work in the MNE group, including wage earners and self-employed persons (directors, administrators or any other people involved in the group board). The turnover comprises the totals invoiced of the MNE group, and this corresponds to market sales of goods or services supplied to third parties. The assets comprise total economic resources controlled by the MNE group entity as a result of past events. The number of legal units is the number of legal units that are part of a MNE group.

5. Quality evaluation of the data sources

In the quality evaluation, for each MNE group, each transformed variable was compared with the information in EGR to evaluate the quality level and to decide the ranking order of the source for the DB loading. In what follows, the specific result for each variable will be described.

5.1 Ultimate controlling institutional unit country code

The information on the country code of the ultimate controlling institutional unit was present in all the public sources used: table 1 summarizes the results for the main sources considered. In the GLEIF source (respectively in Wikidata, in Wikipedia), it was possible to find the UCI information in 93% (w.r.t. 53%, 88%) of the groups analysed, in which 91% (w.r.t. 88%, 83%) of the cases, these UCIs had the same country code as in the EGR 2020. Thus, the GLEIF has the highest priority.

Table 1: sources considered for the ultimate controlling institutional unit country code and integrated sources results.

Public sources	variable coverage	same CC	different CC	priority level
GLEIF	93%	91%	9%	1
Wikidata	53%	88%	12%	2
Wikipedia	88%	83%	17%	3
Integrated sources	*94%*	*91%*	*9%*	-

The last row in Table 1 describes the integrated result that can be obtained using the information on the UCI country code from all the three sources for the creation of the data in the information DB, applying the priority level of each source. The integration process allows a slight increase of the UCI country code variable coverage to 94%.

In the integration process, each variable was populated using the highest quality data first and then the lower one in sequence: a trigger made the lower data available whenever the MNE group's data was not present in the above quality level source. If the last available public source had no information, the data value was null.

5.2 Number of persons employed, Turnover, Asset variables

In the EGR, not all variables (employment, turnover, assets) have an actual value. For each variable and source, 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):

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

Therefore, the variable quality indicator is the cumulative frequency of the number of public occurrences having relative difference values in the range *-1/2* to *1/2* divided by the total number of Variable Coverage occurrences (#VarC):

$$Variable\ Quality\ Indicator = \#RDVar\_Val\ [-1/2,\ 1/2]\ /\ (\ \#VarC\ ) \qquad (2)$$

All occurrences of the public source with relative differences in the interval *[-1/2, 1/2]* contribute to the numerator of the variable 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 employment that is not null in the 2020 EGR frame 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. Variable coverage is the ratio between the number of matched

MNEs groups with information on persons employed present in the public sources and in the 2020 EGR frame.

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%*	-

In table 2, the percentage of overlapping MNE groups and variables of the integrated sources are noticeably higher than the highest value of each single public source reflecting that the public sources do not cover the same MNE groups and therefore the integration process can optimize the coverage. On the contrary, the quality indicator for all integrated sources equal to 43%, is lower than the value of Wikipedia (47%), because 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 the lower quality.

Figure 1: Log-log scatter plot of number of persons employed, EGR versus public source

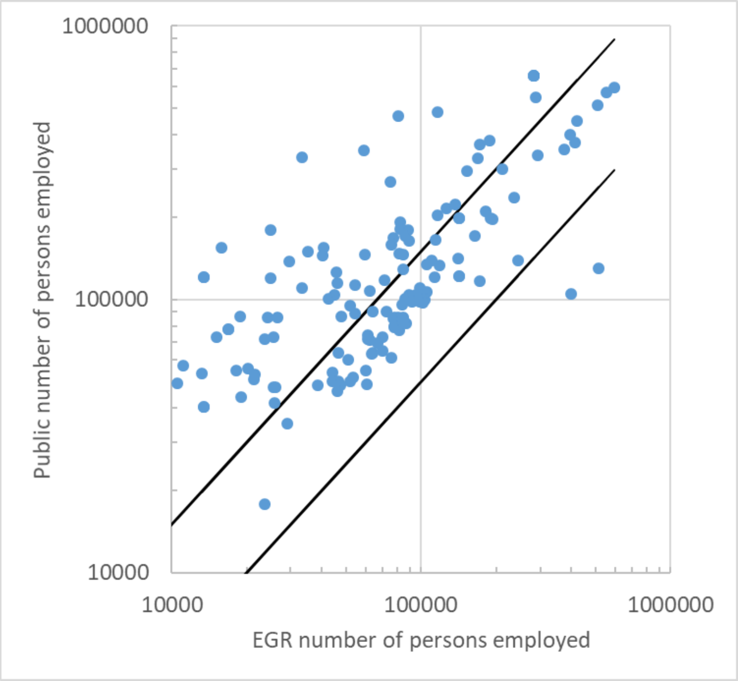


Figure 1 shows the number of persons employed according to EGR versus the integrated public sources' values for each MNE group. The two straight lines are the range boundaries used for the employment quality indicator.

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

Then, concerning the turnover[3] which must be the consolidated value, excluding the intra-group sales, it has values in EGR only in 16% of considered cases. It was possible to retrieve information only from DBpedia and Wikipedia, table 3 shows the percentage of information retrieved.

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%	-

According to the variable quality indicator, defined as for employment, DBpedia was emerging as the first priority source and Wikipedia as the second source, because even if Wikipedia had more coverage the quality of DBpedia was better. The integrated source row in the table shows that it is possible to obtain information on Turnover for 78% of the MNE groups and that for 20% of them it cover the information about turnover values in EGR. The quality indicator for the integrated turnover was 69%. Similar to the number of persons employed analysis, this value was lower than the maximum quality value available, reflecting that the most available information source had a lower quality indicator.

From the analysis of turnover results that in 65% of the MNE groups it is possible to obtain information from public sources even when the corresponding values in EGR are null.

Concerning the Assets which comprises total economic resource controlled by MNE groups, the overall conclusion is similar to the turnover's one variable. In the EGR the values related to Asset are very rarely present. For the considered sample the EGR asset variable is present only in 10% of cases and 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.

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%*	-

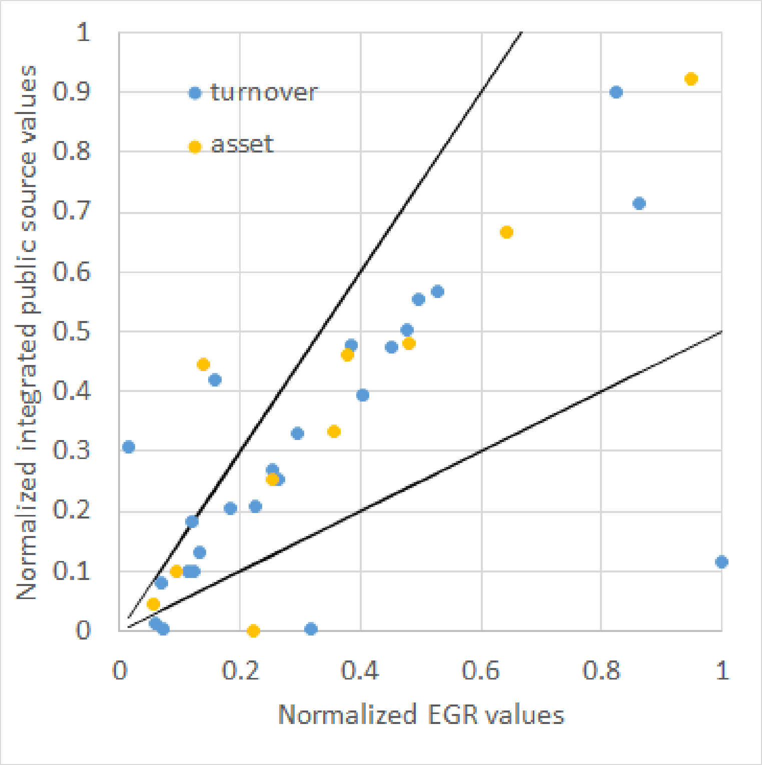
The integrated source row in table 4 shows that the percentage of retrieved MNE groups is higher than the value of each single public source. This is due to the wide availability of public information usable. The variable overlap remains 13% and the quality indicator increases slightly. Naturally, these 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.

From the analysis on asset results that in 53% of the MNE groups it is possible to obtain information from public sources even when the corresponding values in EGR are null.

[3] 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 deductible as well as all duties and taxes on the goods or services invoiced by the unit.

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

Figure 2: Normalized scatter plot of turnover and asset EGR versus public source



In the scatter plot, the overall limited presence of usable information is evident, even if it is possible to recognize an acceptable fitting of the integrated public sources with EGR for both variables, which reflects that the quality indicator of the integrated sources is high (table 3 and 4). Differently from the persons employed variable, in these cases there are not systematic data shifts but instead, dots outside the quality range lines are randomly distributed. The results show that turnover and asset retrieved from public available sources, when available is of acceptable quality and could be used to fulfil the missing value 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 controlled directly or indirectly by the global group head, as well as the hierarchical structure. We have treated this variable separately from other because the conclusion is radically different.

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 unit owned by the MNE group list 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 number of legal unit quality indicator values 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.

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%*	-

The integrated source row shows the possibility of obtaining information on number of legal units for 93% of cases. The quality indicator of integrated sources results of 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: Log-log scatter plot of number of legal units in the group, EGR versus public source

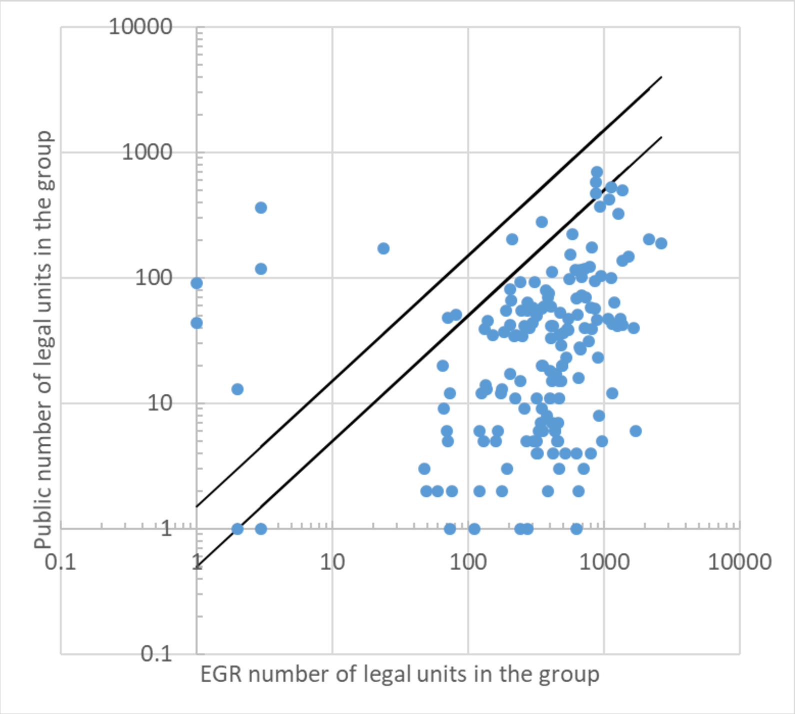
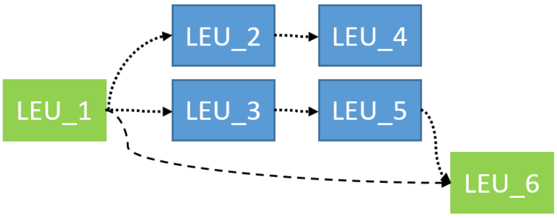


Figure 3 shows the scatter plot on Number of legal units in the selected MNE groups of EGR versus the integrated sources, it illustrates that the coverage of the EGR is better than the coverage of public sources. The main reason could be related to the more accurate official sources used by EGR.
Anyway, the comparison of the structure is a very complex task. A missing or additional relationship from the integrated public sources with respect to EGR might make the task difficult and the consequence of each modification may be significant.

In this part, the analysis on the legal units was extended to the MNE groups' structures to try to understand the differences on the number of legal units, but a more detailed analysis is outside the scope of this paper. At this stage, we would like to provide a possible interpretation of the mismatch found through the GLEIF source.
Figure 4 shows a typical mismatched found, where the coverage of EGR is higher than the coverage of the integrated sources. The UCI information is confirmed by the Information DB but the perimeter

is very different between the two sources, the coverage of the GLEIF is extremely limited compared to the EGR.

Figure 4: group structure examples:



Green boxes represent legal units known by both sources, blue ones only known by EGR. Dashed line only known only by GLEIF and dotted lines only known by EGR.

To give more details about the Fig 4, we quickly compared the GLEIF data and EGR data in terms of unit coverage. The GLEIF database contains 2.1 million units[4] (respectively 1.6 million in EGR[5] [4]) of which 1.3 million have a usable identifier. We focused on European and UK units (1.1 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 EGR but only 11% are part of a MNE group operating in EU according to EGR. We used the relationships provided by GLEIF in its second part of its database ('Who owns whom'), we find only 34000 units (out of around 700 000) involved in a group operating in EU according to GLEIF, it represents an improvement of 0.2% to the EGR database. A more detailed analysis must be carried out in order to exploit control relationships known to GLEIF and unknown to the EGR, in particular for the non-EU part.

6. Conclusion

Based on the analysis of the results, this study concludes that public sources can be taken into account when complementing EGR missed information on MNE groups, but however their contribution needs to be precisely qualified.

With regard to most of the attributes of a MNE group, the gain seems to be positive for the country of the ultimate controlling unit (UCI), turnover and 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 Europe in EGR. If confirmed, the employment data from the public sources could well complement the null data of EGR for the countries outside the EU.

Finally, information on legal units requires a fine-tuned approach tailored to the aims sought, which could be to complete the existing structure of the MNE groups with more relationships and with the UCI information and improve the general quality of the EGR.

Public sources appear to be interesting to fill in any missing values and to assess and consolidated information of the MNE groups in the EGR. They can also be used as triggers for early detecting changes in the EGR and for actions needed for the maintenance and/or for profiling activities dedicated to the construction of the enterprise statistical unit.

[4] Reference snapshot: 01/04/2022

[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 on the basis of 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 exploited 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 understand these identifiers.

The pilot study and the quality analysis carried out in this paper are based on a limited number of MNE groups and the next step is to extend the work on a more substantial number of MNE groups to get robust results.

The intention is to implement the use of public sources to complement EGR missed information on MNE groups in order to further extend its coverage in particular for the countries outside the European Union.

7. Bibliography

[1] Eurogroups register - Statistical business registers - Eurostat (europa.eu)

[2] G. Bianchi, A. Laureti Palma, S. Quaresma, "Prepare your data warehouse for a Big Future, by including Big Data", European Conference on Quality in Official Statistics 2018, Kraków, Poland. 26-29 June 2018

[3] Smart Data for Multinational enterprises (MNEs) – using open source data to obtain information on Multinational enterprises — 2021 edition - Products Statistical working papers - Eurostat (europa.eu)

[4] Structure of multinational enterprise groups in the EU - Statistics Explained (europa.eu)