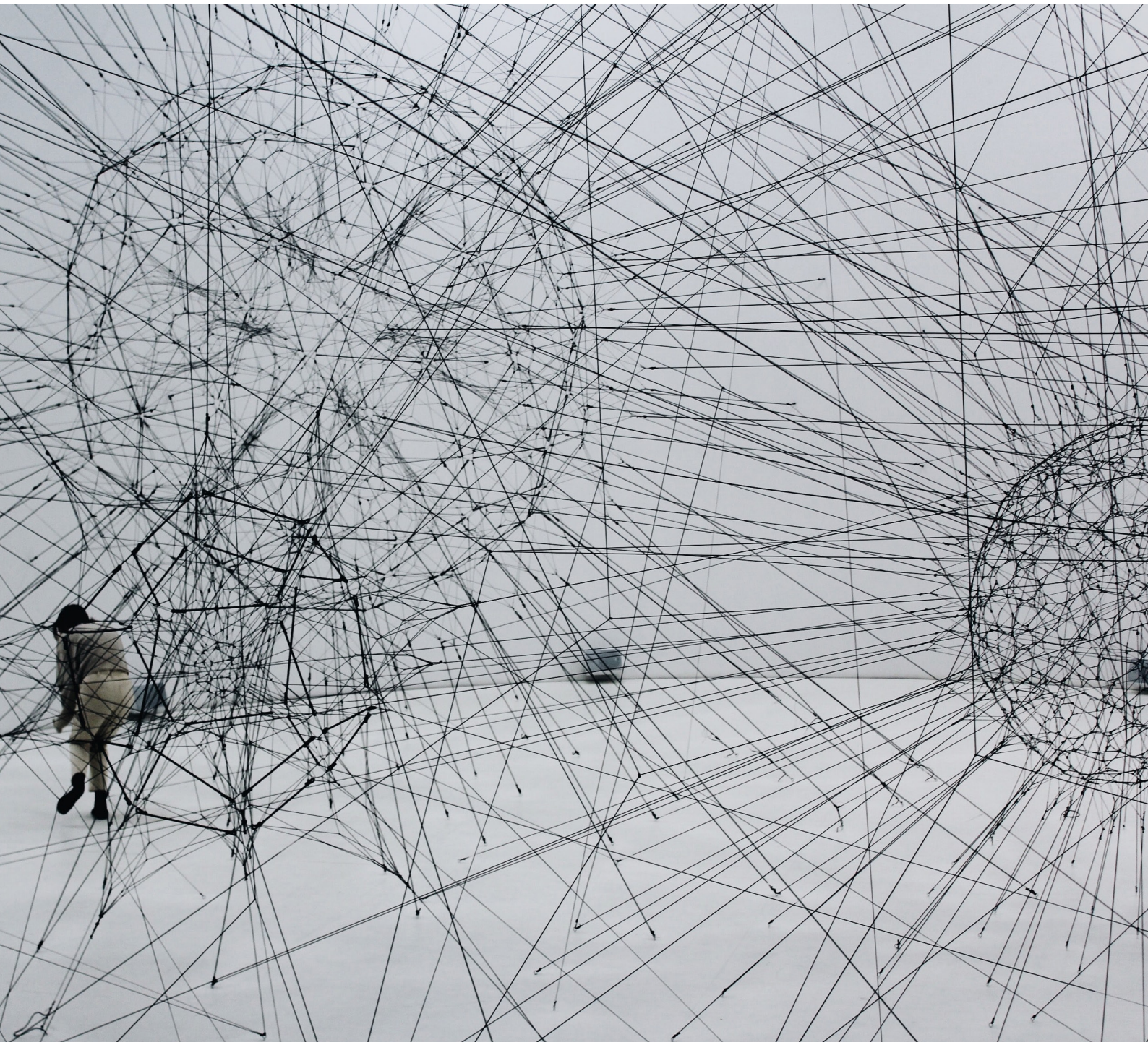
Smart Data for Multinational enterprises (MNEs) – using open source data to obtain information on multinational enterprises

FOTIS PAPAILIAS, VIRGINIA BALEA,
HIONIA VLACHOU, GEORGE KAPETANIOS

2021 edition



STATISTICAL
WORKING PAPERS

eurostat

Smart Data for Multinational enterprises (MNEs) – using open source data to obtain information on multinational enterprises

FOTIS PAPAILIAS, VIRGINIA BALEA, HIONIA VLACHOU, GEORGE KAPETANIOS

2021 edition

Manuscript completed in October 2021

Luxembourg: Publications Office of the European Union, 2021

Theme: General and regional statistics
Collection: Statistical Working Paper

PDF: ISBN 978-92-76-42218-1 ISSN 2315-0807 doi: 10.2785/415398 KS-TC-21-007-EN-N

Abstract

Traditionally, EU Member States maintain business registers for statistical purposes. The fragmented picture across EU countries on Multinational enterprises (MNE) and enterprise groups operating in the open market represents a challenge for the harmonisation between several types of statistics affected by globalisation. These include Foreign Affiliates Statistics (FATS), Foreign Direct Investment (FDI) and international trade in services and goods. Subsequently economic globalisation created new opportunities for businesses to organise their production chains, resulting in global value chains and their contributions in the production networks of resident enterprises in multiple countries. Those arrangements present challenges to macroeconomic and business statistics, including the supporting business registers ([1]). The statistical capture of aspects related to globalisation is not straightforward, as the activities are transnational or multinational statistics and are not bound anymore to the national territory of one Member State or to the aggregation of national data. Thus, the EuroGroups Register (EGR) has been set up with the aim of adjusting European business statistics to the emerging needs caused by globalisation, as well as to reduce the burden on enterprises. Trusted Smart Statistics endeavour aims to develop statistics in datafied societies, leveraging information on the Internet (Web Intelligence), using innovative data collection methods with smart devices, leveraging 'smart systems' such as smart energy, smart meters, smart transport, etc.

Within the framework of Trusted Smart Statistics and Web Intelligence, the aim of this project is to retrieve information about MNEs and enterprise groups operating in the EU and EFTA countries, including enterprise groups with headquarters located outside the European Union which are active in the common market, and to extract and deliver aggregates that might be compared with existing information.

This paper brings together the results of research undertaken within the framework of 'Smart Data for MNEs' from defining the scope and assessment of relevant data sources used to develop the necessary components of web scraped data and to process the information obtained for further analysis. Subsequently, we transform and visualise this information, exploiting relevant aspects in order to enhance and extend the availability of information on MNEs. The results of this study provide input for the developments within the Trusted Smart Statistics framework, leveraging information from the web.

The paper concludes with a series of recommendations for dealing with smart data for multinationals in official statistics, encouraging producers of statistics to focus on quality improvements.

Keywords: Smart Data; multinational enterprises; API; Web Scraping.

Authors: Fotis Papailias ([2]), Virginia Balea ([3]), Hionia Vlachou ([4]), George Kapetanios ([5]).

Acknowledgments: We would like to particularly thank Konstantinos Giannakouris, the coordinator of this study at Eurostat Unit B.1, Martin Kahlberg, Marco Stocchi, and Fernando Reis as well as Unit G.1, Agne Bikauskaite for their useful comments and feedback.

The study was carried out by GOPA, as a contractor of Eurostat for the framework contract on Methodological Support (Ref. 2018.0086).

([1]) For further information: Handbook on Accounting for Global Value Chains.
([2]) fotis.papailias@knotanalytics.com
([3]) vbalea@gmail.com
([4]) Hionia.Vlachou@gopa.lu
([5]) George.kapetanios@kcl.ac.uk

Table of contents

Index of tables

Index of figures

Abbreviations

ADIMA	Analytical Database on Individual Multinationals and Affiliates
API	application programming interfaces
EDGAR	the Electronic Data Gathering, Analysis, and Retrieval system
EFTA	European Free Trade Association
EGR	EuroGroups Register
ESS	European Statistical System
EU	European Union
FATS	Foreign Affiliate Trade Statistics
FDI	Foreign Direct Investment
GDELT	Global Database of Events, Language and Tone
GGR	Global Group Register
GLEIF	Global Legal Entity Identifier Foundation
IoT	Internet of Things
LEI	Legal Entity Identifier
MNEs	Multinational enterprises
OECD	Organisation for Economic Co-operation and Development
SEC	US Security and Exchange Commission
WIH	Web Intelligence Hub

1 Introduction

Traditionally, EU Member States maintain business registers for statistical purposes. The EuroGroups Register (EGR) is the statistical business register of multinational enterprise groups having at least one legal unit in the territory of the European Union (EU) or in the European Free Trade Association (EFTA) countries. The purpose of this network of business registers is to provide a unified and harmonised picture of the Multinational enterprise groups (MNEs) across Europe and offer official statistics suitable for further analysis on both micro and macro levels.

However, the fragmented picture across the EU countries on MNE groups operating in the open market has caused growing problems for the harmonisation between several types of statistics challenged by globalisation. Subsequently economic globalisation created new opportunities for businesses to organise their production chains, resulting in global value chains and their contributions in the production networks of resident enterprises in multiple countries. Those arrangements pose challenges to macroeconomic and business statistics, including the supporting business registers. The statistical capture of aspects related to globalisation is not straightforward, as the activities are transnational or multinational, and are not bound anymore to the national territory of one Member State or to the aggregation of national data. The EGR was set up 'with the aim of facilitating the coordination of survey frames in the European Statistical System (ESS) to produce high quality statistics on global business activities, like FATS and FDI statistics'([6]) . Moreover, it is also meant to enable European business statistics to respond to the emerging needs caused by globalisation, as well as to reduce the burden on enterprises and EU Member States ([7]). With the European business statistics regulation ([8]), the EGR has become an authoritative source within the European Statistical System (ESS).

The purpose of this project – proof-of-concept - was to retrieve for a limited number of MNEs from the Internet any available open source information about MNEs and enterprise groups operating in EU and EFTA countries, including enterprise groups with headquarters located outside the European Union who are active in the common market.

The developments in information technology during the last decades have facilitated the retrieval, organisation and analysis of vast amounts of data. Trusted Smart Statistics are based on: (1) the use of new data sources originating from digitalisation in order to complement administrative data and surveys; (2) the use of new technologies in official statistics, in principle from the web, and (3) the aim of complementing existing statistics and (4) producing new statistics.

The project described in this paper, overall, is mainly designed to investigate open sources for MNEs

([6]) https://ec.europa.eu/eurostat/web/statistical-business-registers/overview

([7]) See also under: Statistics explained - EuroGroups register –: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EuroGroups_register

([8]) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.327.01.0001.01.ENG

information, retrieve data and organise it in formatted databases, which could then be used for statistical and research analysis. This outcome matches the need for a growing demand for non-confidential information on MNEs. With this in mind, the main outcome of the project is considered as a proof of concept illustrating the need for further investigation by the Web Intelligence Hub (WIH) in the context of leveraging information of the internet.

The project aims to propose methods and program software tools (R scripts) to retrieve data in key official statistics, starting from a comprehensive review of the different data sources, the availability of certain variables and their relevance for official statistics. The project presents an output that enables specific users within Eurostat to obtain information on MNEs independently from open sources.

The project on 'Smart Data for MNEs' lines up with previous developments ([9]) within the Trusted Smart Statistics framework and the WIH, such as 'DBpedia and Eurogroups Enterprise Register and other Wikimedia projects', which aimed for the possibility of enriching the EGR and evaluate the possibility of having an open register of MNE groups publicly available sources on the largest MNE groups, like ADIMA developed by OECD or GGR which publishes webscraped data on 100 largest MNE groups.

Our study is meant to encourage the producers of statistics to actively use new open data sources (as well as new technologies), in order to obtain more detailed and frequently updated information that might be useful to produce timely data. A representative example for this need is the COVID-19 pandemic which required researchers across all fields to have immediate and easy (open) access to a large set of information.

This paper is organised as follows: Chapter 1 is an introduction and outlines the scope of the project and its relation to Trusted Smart Statistics; Chapter 2 provides an overview of the advantages of open source data, and it is followed by Chapter 4, which outlines the methods used to retrieve and process the data; Chapter 5 delivers a detailed overview of the main outcomes of the project by task (scoping, transformation, etc.); Chapter 6 concludes the paper with a scientific summary including a series of recommendations for dealing with open source data for Multinational enterprises; Finally, Chapter 7 provides some valuable insights on possible future work.

The project ran from April 2020 until February 2021, under the coordination of Eurostat Unit B.1 with the support of the external contractor GOPA (GOPA Worldwide Consultants in joint venture with GOPA Luxembourg).

([9]) See also:
https://webgate.ec.europa.eu/fpfis/wikis/display/EstatBigData/DBpedia+and+Eurogroups+Enterprise+Register

2 Smart Data for MNEs: Advantages and Disadvantages of Open Sources

Our age is characterised by a steep increase in the evolution of technology. Smart devices, and the Internet of Things (IoT), electronic networks and the ability to store mass electronic data allows the organisation and further analysis of these smart datasets. Smart statistics are seen as the future-extended role of official statistics in a world impregnated with smart technologies. Smart Data has been gaining increasing attention from researchers and statisticians over the last years. A series of papers and methodological considerations have been published on economic and social issues. Buono et al. (2017) provide a comprehensive overview of smart data types used in various applications and we refer the reader to that paper for more detailed information. In general, these applications use financial markets high-frequency data (see e.g. Degiannakis and Floros, 2015), electronic payments data (see e.g. Galbraith and Tkacz, 2007), mobile phones data (see, e.g., Smith-Clarke et al., 2014), satellite images data (see e.g. Henderson et al., 2011), scanner prices data (see e.g. Silver and Heravi, 2001), online prices data (see e.g. Cavallo, 2017) and online search data (see e.g. Choi and Varian, 2009 & 2012). Ricciato et al. (2019) outline the concept of Trusted Smart Statistics as a future evolution of official statistics.

Consequently, we can define the following: 'Trusted Smart Statistics can be seen as a service provided by smart systems, embedding auditable and transparent data life-cycles, ensuring the validity and accuracy of the outputs, respecting data subjects' privacy and protecting confidentiality' ([10]).

Our aim is to provide an MNEs data set with the following units and characteristics by obtain non-confidential data in a timely manner for:

1. Legal units: identification, demographics, control and ownership characteristics;
2. Enterprises: identification and demographic characteristics, main economic activity code (NACE([11])), number of persons employed, turnover, assets, institutional sector;
3. Enterprise groups: identification characteristics, the structure of the group, the global group head, the country of global decision centre, main economic activity code (NACE), consolidated employment and turnover as well as assets of the group.

To obtain information on units resident in EU Member States and EFTA countries, data from the respective national statistical business registers ([12]) as well as commercial source (for units outside

([10]) Eurostat, Trusted Smart Statistics in a nutshell, see under: https://ec.europa.eu/eurostat/cros/content/trusted-smart-statistics-nutshell_en and Ricciato et al (2019).
([11]) NACE Rev. 2 - Statistical classification of economic activities
([12]) See also: European business statistics methodological manual for statistical business registers — 2021 edition: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-20-006

the EU/EFTA) are used.

A Smart Data approach offers the following advantages for researchers and statisticians:

1. Since EGR data are confidential (data exchange as defined in the Regulation 2019/2152 on European business statistics Article 10 Exchange of and access to confidential data for the purpose of the European framework for statistical business registers) ([13]). Smart Data – which exists freely on the internet - could be obtained, formatted, aggregated and provided without further guarantees. The last term means that open source data are not examined, organised and/or curated by a data provider who can guarantee the authenticity or accuracy of the displayed information.
2. Information in the EGR is validated by the various national statistical institutes, as well as by Eurostat, for different roles and responsibilities. The data have been produced annually since 2008 with a time lag of fifteen months (i.e. data for reference year T is available at T+15 months). Official EGR data require time for the data to be collected/received and validated by Eurostat, whereas Smart Data could be updated on the spot (as soon as the original source has updated the information).

By using web scraped data for MNEs, collection costs may be reduced and, additionally, the frequency of information collected could be increased.

Naturally, this does not come without some constraints. Generally, all databases which collect information from various open sources on the Internet rely heavily on: (i) the provider's availability, and (ii) provider's data quality. We use the term 'Provider' to describe the organisation that publishes the database. In proprietary databases, the provider is usually assigned with the tasks of collecting, cleaning, organising, curating and publishing the database in a final format. In our context, where we focus on open-source data, the provider usually simply publishes the data (without necessarily cleaning or checking its quality) and maintains the API or another access to the database.

Provider's availability is crucial when building databases using APIs or scraping techniques as a change on the website or the API service can heavily affect the collection process. In addition, in such an approach we trust the provider that the data displayed on the web are validated, and are hence of certain quality which might or might not be correct. Nonetheless, in most cases, we do not have any information on the quality of the data and the possible validation procedures.

Taking into account all advantages and disadvantages when constructing a database based on information collected from various online sources, we believe that a 'Smart' MNEs database that could be freely provided, analysed and summarised in different ways can only add value to Eurostat's effort towards a harmonised universe of information and statistics.

([13]) REGULATION (EU) 2019/2152 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 November 2019 on European business statistics,
https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R2152&from=EN

3 Methods

This project relies heavily on data collection via the use of APIs and web scraping. The program scripts which download the datasets are written in the R and Python languages. In an effort to facilitate the data collection process, we have organised this task in separated smaller bits of code scripts. In this way, the applied researcher has the flexibility to collect the data from each source separately and/or update only one of the sources in case there are new data available.

The program scripts, which clean the data and provide a final formatted overview in the form of an 'interactive dashboard', are written in R.

The R code/program scripts developed for the data retrieval, including a user manual (with instructions on the procedures) have been provided to Eurostat. Installation of software at Eurostat premises is not a requirement, since the development of the programs is done in R, and could be run from any PC with the software installed. However, the project team paid particular attention not to build the code scripts based on too many libraries, in order to minimise any future possible maintenance costs ([14]).

The program scripts map variables and information obtained against EGR variable names (such as various identifiers, type of company, industry or sector, number of employees, status).

The user guide provided within the scope of this project should enable the user to retrieve, process and transform the data following each step of instructions accompanying the guide and obtain further insight by the examples provided.

([14]) Still, some libraries, which are mentioned in the references section, were needed to be used.

4 Main outcomes of the project

The project was organised in four subsequent steps. Each project step concluded with a report on preliminary as well as relevant pieces of software. First, there is the scoping of data sources (step 1), followed by insights into data retrieval from the different sources identified (step 2). There is a specific focus on economic/financial data. Third, variables extracted in the previous steps have been pre-processed in order to be transformed into tabular information about enterprise groups operating in the EU and EFTA countries (step 3). The last step of the project focuses on the transformation of the information in order to compute aggregate statistics related to the MNE performance at European level and per country (where available). Under step 4, we provided an operational and interactive dashboard to visualise this information, suitable to run in the most popular web browsers. Case studies incorporating information from annual enterprise reports and from specific databases (EDGAR) complemented the results of the project.

4.1. Scoping

In the context of the first step, the focus was on the scoping of various sources, which potentially offer data for European Multinational Enterprises.

To define the scope of this first step, we used the names of 199 companies and enterprises to be investigated and a list of relevant variable names used in the EGR.

A large number of sources were assessed on the basis of their advantages and drawbacks.

1. DBpedia: Includes open and freely available information on MNEs. The sources may be particularly useful to retrieve information about the legal name, number of employees and URL of businesses worldwide (multinational enterprises);
2. Global Legal Entity Identifier Foundation (GLEIF): GLEIF offers another access point to a large set of information on MNEs and their composition. GLEIF has two levels: level 1) general LEI information, and level 2) 'Who owns Whom';
3. Thomson Reuters Open PermID: The PermID APIs offers a unique entry point for MNE queries, allowing the retrieval of the Legal Entity Identifier (LEI) of the companies. It is noted that the LEI is part of the confidential data in the EGR. This information is difficult to retrieve using exclusively DBpedia and Wikidata. The Open PermID thus offers a possibility to improve completeness of the potential output of the application;
4. Open Corporates: API used to obtain information on various MNEs (but not finally adopted, given that it requires a license for more than 50 API requests);
5. Wikidata: information obtained using the MNE names as keywords;

6. Wikipedia: information scraped from Wikipedia MNE entries;
7. Annual PDF Reports: these are usually available from the MNEs. However, they are very heterogeneous and aim to inform the investor or support client relationships. Information in some cases might be included but is not always informative or presented in a straightforward way;
8. US Security and Exchange Commission (SEC) EDGAR database: This database stores files of different types and formats, periodically transmitted by US corporations. The corporations file the 'Form 10-K' every year, which offers comprehensive information on consolidated income and expenditure, balance sheets, adjusted income, operating income margin, earnings per share, effective tax rate (non-GAAP), total number of employees, subdivision per geographic area, total and tangible/intangible assets, liabilities and equity. For the project purposes, information on the Form 10-K and Form 20-F can be used, as well as one particular annex of the Form 10-k (Exhibit 21/21.1) which entails a list of subsidiary companies for a given enterprise;
9. More sources were included in the scoping and relevant weaknesses (including web scraping data retrieval limitations) were identified ([15]);
10. Some proprietary databases (Fame, Statista, Osiris) ([16]).

We have made: (i) an assessment of those data sources and their potential to be used for the purposes of the project, (ii) a description of the variables that can be further extracted and exploited, (iii) an evaluation of the data and the conditions for accessing and using the data sources in a sustainable manner.

The information obtained was analysed in a tabular form and assessed accordingly. To facilitate the comparison, we investigated the general MNEs information, information about contact details, the governance structure of the MNEs, various identifiers for MNEs, stock information, financial indicators such as financial indicators coming from Financial Statements, Income Statements, and Cash Flow statements, as well as variables aiming to capture general public interest in the MNE groups.

It is also very important to stress the fact that some of the sources contain pure textual information. Text analytics and machine learning is a rapidly growing field in the economics and business literature (see e.g. Buono et al., 2017, and the references therein for some indicative starting points in economics and finance applications). As text analytics is an extremely complex field, and consists of a separate analysis of its own, the team will comment on and investigate its usefulness, but will not perform an in-depth analysis, which is beyond the scope of the current project.

An assessment and first attempt to map the variables was conducted by analysing the list of the data sources, a detailed description of the data sources and the available data, a detailed report on the conditions and difficulties to use the data sources in a sustainable way, their potential and their limitations. The analysis showed that certain information could be obtained from multiple sources, ideally combining them (see Table 1).

After careful consideration, we have collected data from DBPedia, EDGAR (available only for a smaller number of MNEs), GLEIF, Google Trends (only used for the dashboard to illustrate public interest), Open Corporates, Wikidata, Wikipedia and Open PermID.

General information on MNEs could be obtained from multiple sources, preferably combining GLEIF, Open PermID, Wikidata and DPBedia. The number of employees could be retrieved from both DBPedia and Yahoo! Finance. Contact details could also be extracted from multiple sources. The sources which seem to have more information available are: Wikidata and DBPedia. Information regarding the Governance Structure can be obtained mainly from Wikidata; however, OpenPermID could also be useful and both sources display essentially the same information.

Wikidata also displays the majority of identifiers; therefore, it is recommended to obtain a large

([15]) In particular, we also investigated the information in OECD ADIMA, LEI Look-Up, ECTA, ESMA, financial websites (MarketScreener, Yahoo! Finance, Zacks, Bloomberg, CNBC, FT.com, Investing.com, Markets Insider, MarketWatch, Reuters, SeekingAlpha, SeekingAlpha M&A News, WSJ).
([16]) Table 1, which follows, organises all sources by availability.

number of IDs from there. Open PermID provides additional information on the stock of the MNE. Yahoo! Finance provides a large selection of company rankings and financial reporting information. Wikipedia page views and Google Trends are useful to construct an annual aggregate of web traffic/public interest.

The US Security and Exchange Commission (EDGAR) is an official source, thus the data would not need any validation. However, a drawback is that not all Europe-based MNEs have operations in the US, hence they do not need to fill in any forms. Additionally, information from EDGAR is obtained in textual format and a special package of R for web scraping, since no direct APIs could be used for data retrieval. Information from annual reports of MNEs could be obtained from the original source; however, web scraping would require a large-scale project of text analytics to obtain all relevant information.

The analysis of textual documents (e.g. PDF files of periodic financial reports) did not provide harmonised, consistent, comparable or relevant information for a number of variables for the enterprises. Annual reports were assessed as possible valuable sources of information. However, to retrieve information on the number of affiliates seems to be problematic due to the heterogeneity in the annual reporting of companies and lack of legal obligation to explicitly identify them in the financial reporting documents.

Table 1: Comparing availability of sources

Source	Available	Unclear Availability	Not Available
Annual PDF Reports	●		
DBPedia	●		
EDGAR	●		
GDELT	●		
GLEIF	●		
Google Trends	●		
LEI Look-up	●		
PermID	●		
Open Corporates	●		
OECD ADIMA	●		
Wikidata	●		
Wikipedia	●		
ECTA		●	
ESMA		●	
MarketScreener		●	
Yahoo! Finance		●	
Zacks		●	
Bloomberg			●
CNBC			●
Fame			●
FT.com			●
Google Finance			●
Investing.com			●
Markets Insider			●
MarketWatch			●

Source	Available	Unclear Availability	Not Available
Reuters			●
SeekingAlpha			●
SeekingAlpha M&A News			●
Statista			●
Osiris			●
WSJ			●

4.2. Data retrieval

With the aim of developing and documenting the data retrieval components and their specificities, scheduling periodical use/download of the data sources, and installing the software at Eurostat's premises, this second step builds on the results of identified sources while performing the first step.

There are two main ways to extract the information:

- API: R and/or Python were employed to create all program scripts for the automated data collection. The same codes also store the data in .csv or spreadsheet files (or database files);
- Scraping: investigation of the URL to be scraped, data to be extracted followed by the development of a specific R code to extract and store the data.
- The organisation of the collection process was planned in a consistent way using the list of 199 MNEs defined at the beginning of the project and under Step 1. Four main issues were identified:
- Unclear names: the supplied list of names of the MNEs did not corresponded exactly with the names found in the open data sources;
- Activity status: missing LEI would indicate a non-active status of the company. This information would need validation;
- Companies and parent companies: Parent company could not be identified for some of the provided MNEs;
- Companies not publicly listed: Some of the MNEs are not publicly listed (i.e. do not have their shares traded in a stock exchange). This means that identifying quarterly or annual reports in an automated way from standardised sources (e.g. Yahoo Finance) is not possible given the fact that the data are only widely available for publicly listed companies.
- From the initial list, 10 companies were removed due to insufficient information (e.g. LEIs, name, etc.). Since the authors did not have any access to the EGR, we assume that an internal user might consolidate the list of those companies. Conversely, it has to be taken into account that the list of companies should be revised regularly, taking as well mergers and acquisitions, separations or in-activity into account.

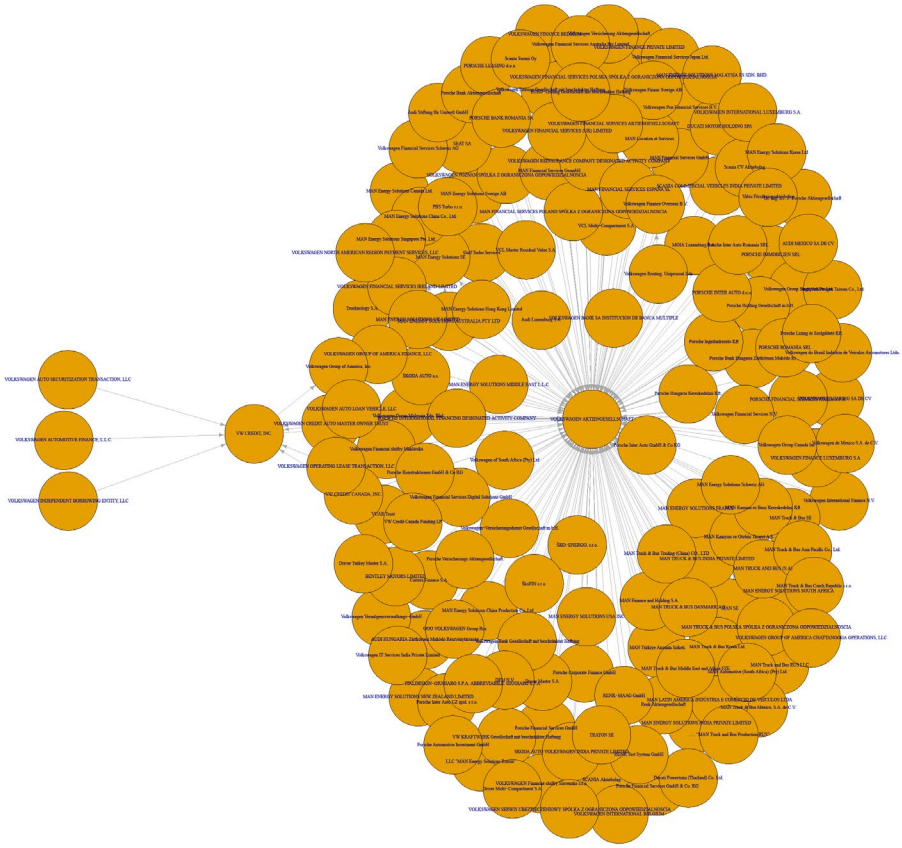
In an effort to facilitate the data retrieval process, we have organised this step in separate smaller program scripts for each step in a modular fashion. In this way, the applied researcher has the flexibility to collect the data from each source separately and/or update only one of the sources in case they think there is new data available. As a result, in this step, the researcher can download the data for each MNE (where available) in raw format and save the output in .csv files. The code retrieves data from all the open and available data sources as discussed earlier.

Essentially, all pieces of program scripts were written with the same structure (and, where appropriate, the same commands were used). Each of these program scripts reads the MNE names and/or identifier from a Masterfile (e.g. the provided MNE or a URL), visits the corresponding page

(usually via an API) and downloads the information in a raw format. This holds for almost all sources: DBPedia, Wikidata, Wikipedia, PermID and Open Corporates.

Relatively to GLEIF, we make use of the 'golden copies', which are files that contain the whole GLEIF database. The procedure is similar to the ones described above; however, we do have access to second level data (i.e. ownership information ([17])) and, thus, we can create a network of affiliated companies; see Figure 1 for an illustration using Volkswagen AG.

Figure 1: Creating a network of affiliate companies using GLEIF ([18])



Note: The figure displays a two-level analysis: (i) companies owned by the MNE (Volkswagen in this example), and (ii) companies owned by other companies which are owned by the MNE. Given the density of the information in the figure, we also provide a high-quality version to facilitate the view.

A small number of EU-based MNEs is available in the SEC EDGAR database; these are the MNEs which are also listed in the US stock exchange ([19]).

The main difficulties of creating a database from annual financial reports is the large degree of heterogeneity in the presentation of information across companies. For example, we might have

([17]) GLEIF allows users to study first level information, i.e. 'who is who', with relevant detailed information of the MNE. It also allows them to study second level information, i.e. 'who owns whom', and identify the network of parent/children companies.
([18]) Click on the picture for a more detailed view.
([19]) The associated program scripts also provide a 'demo' analysis for seven US exchange-listed MNEs. The supplied code could be applied in case more MNEs are identified in the EDGAR database. However, given the scarcity of information due to the fact that most EU-based MNEs are not listed in the US stock exchange, and consequently are not required to fill any forms, we shift our focus from EDGAR to the other sources with data availability

different ways of communicating the information across years and/or across MNEs. Writing fully automated code scripts for .pdf files and EDGAR (given the inconsistencies) can prove challenging given the heterogeneity of the information ([20]).

The R codes scripts provided highlighted the usability of data from EDGAR and annual pdf reports. The former constitute data on balance sheets/income statements presented in a relatively standardised format (however, only for some of the MNEs), whereas the latter refer to data that are very heterogeneous but require a dedicated team for processing (of course, the same data exist in various other sources).

Some of the requested information may be available (in particular, quantitative information such as financial and economic indicators, is available on the web); however, it cannot be scraped without prior agreement (e.g. Google finance does not allow web scraping). A list of data sources which display financial information in an organised manner, mainly targeting individual investors, follows. However, all these sources clearly state in their terms of service that none information should be scraped.

- Reuters Financial: Refinitiv
- Yahoo Finance: MorningStar
- WSJ/MarketWatch: Factset
- Bloomberg: Bloomberg
- Markets Insider: Factset
- Investing.com: N/A.
- MarketScreener: S&P Global Market Intelligence
- SeekingAlpha: S&P Global Market Intelligence

This issue needs to be considered in the future in order to prepare the necessary conditions for either web scraping information directly from the web or using APIs.

Perhaps a collaboration with one of these providers would allow to: (i) access all necessary data in a standardised format and build infographics, and (ii) display the data online (however, not redistribute it).

An improved way of extracting information from .pdf files could be developed in Python which allows more flexibility in manipulation of textual data. However, this is a long-scale project and it will still require a large degree of manual input regarding the parameters of each file. Therefore, room for improvement exists but the gains of this source must be carefully investigated. An alternative way would be to use the 'interactive' Shiny App 'tabulizer' which brings the .pdf on the screen and the user selects which table to extract. However, this would still require a lot of user input as the final data are not clean or in any way consistent. Finally, another software which can parse information from .pdf files to text is the ABBYY FineReader which, as the above Shiny app, lets the user select a particular table and transforms it into an Excel table. Again, this requires a team to work on this full time as it involves a large degree of labour work.

4.3. Data Pre-processing

The previous Step 2 provides the tools to retrieve the raw and unformatted data. Consequently, Step 3 is concerned with cleaning the data and organising the information to obtain a more structured format. As mentioned above, at a minimum, the resulting records will contain general information (LEI, name of company, URL of the official web portal of the company, headquarters information) and performance information where available (number of employees turnover, assets, liabilities and

([20]) As in the case of EDGAR database, where the associated scripts provide a demonstration on seven MNEs, we have also created program scripts which demonstrate how information from an annual financial report can be obtained.

investments, number of subsidiaries and a list of the countries under which they operate).

In this part of the coding automation, the researcher is required to provide some information manually. In particular, the following procedure is adopted: (i) screen the obtained fields, (ii) remove non-relevant fields, (iii) transform relevant fields to corresponding variables, (iv) provide details for these variables. This procedure is repeated for all sources (apart from EDGAR). Once the fields have been mapped to variables, the remaining analysis is automated. For example, we see in Figure 2 that the field 'asset.value' might contain import information on the value of assets and is, therefore, assigned to a variable called 'Assets1' and 'Assets2'.

Figure 2: DBPedia data example

	Old Value	New Value [If -BLANK- then not used]	New Value
87	http://dbpedia.org/property/areaServed.value	AreaServed3	AreaServe
35	http://dbpedia.org/ontology/assets.value	Assets1	Assets1
638	http://dbpedia.org/property/assets.value	Assets2	Assets2
36	http://dbpedia.org/ontology/assets.datatype	AssetsCurrency	AssetsCurr
589	http://dbpedia.org/ontology/assetUnderManagement.value	AssetsUnderManagement	AssetsUnd
590	http://dbpedia.org/ontology/assetUnderManagement.datatype	AssetsUnderManagementCurrency	AssetsUnd
55	http://dbpedia.org/property/assetsYear.value	AssetsYear	AssetsYear
83	http://dbpedia.org/ontology/locationCity.value	City1	City1
697	http://dbpedia.org/property/hqLocationCity.value	City2	City2
106	http://dbpedia.org/ontology/locationCountry.value	Country1	Country1

As explained under Step 1, the researcher needs to carefully check each source's fields and identify the most relevant ones to be reported. These are the input files for Step 4. It is also important to highlight the difficulties identified with EDGAR and annual reports. The associated output files for this step include separately retrieved dates, fields and variable list files as well as files with the data organised by source.

The EDGAR public database allows free access to a company's financial information. Two forms contain relevant annual information about MNEs. The forms are 10-K and 20-F. The form 10-K has to be submitted by domestic public companies. Domestic public companies must fill out the annual reports, as well as current reports when certain events occur that require prompt disclosure. The annual reports include financial statements for the relevant period.

The form 20-F is linked to an annual report for foreign private companies. A foreign private issuer must file their annual report on this Form within the four months after the end of the fiscal year covered by the report. The company has to provide selected historical financial data regarding the company in the same currency as the financial statements.

The form contains a description of the nature of the company's operations and its principal activities, stating the main categories of products sold and/or services performed. Consolidated financial statements include: balance sheet, statement of comprehensive income, statement showing changes in equity, cash flows statement and related notes.

In EDGAR, we found information about 40 groups out of 199 of our initial company list. For 29 groups, there is information available according to 20-F form, while the remaining 11 groups filled in 10-K form. Not the same documents are available for all 40 groups but for most of them we could find consolidated statements of income, consolidated balance sheets, and consolidated statements of cash flows, consolidated statements of shareholders' equity and notes to financial statements.

Notes to financial statements are quite heterogeneous. For an important number of groups, there is information on operating segments. Fiscal year is similar to calendar year for most of the groups, while for some it is different (ends March 31st). To facilitate the comparison, we have created specific 'EDGAR' variables so that the user can simply filter out all EDGAR-related information.

The annual reports available on the groups' websites differ from one group to another in terms of the format of the reports (tables, text presentation, graphs/charts). However, for the majority of the MNE, the annual reports contain information on consolidated income statement, consolidated balance sheet and consolidated cash flow. These are presented in a table format. Concerning the notes to

the consolidated financial statements, for some of the groups the information is presented in tables while for others it is available as text or charts/info graphics.

Below are the variables and the available data sources, as well the priority to be given in using them (more details in Annex I).

Table 2: Overview of information available on MNE groups

Source	Share of MNE/Groups in the list for which information is available	Structured information	Level of information
Annual reports	• Almost 100 %		Group data
EDGAR	• About 20%	•	Group data
GLEIF	• About 90%	•	Legal unit and Group identification (name, address and identifiers)
PermID	• About 90%	•	Legal unit and Group identification (name, address and identifiers)
Wikipedia	• About 80%	•	Group (identification)
DBPedia	• About 60%	•	Legal unitand Group identification (name, address)
Open Corporates	• About 25%	•	Legal unit and Group identification (name, address)
Wikidata	• About 88%	•	Group identification (name, address)

The table below presents the variables identifying the legal units of the MNE. As shown in the table, most of the data related to the LEIs are available from GLEIF data source.

Table 3: Overview of information available for the Legal Unit (LEU)

Description of the variables	Sources
Frame reference year	EGDAR, Annual Reports, for the LEU of the group head
LEI number of the legal unit	GLEIF; PermID, Wikidata for the LEU of the group head
Primary national ID of the legal unit	GLEIF; EDGAR for the LEU of the group head
Name of the legal unit	All (DBpedia, EGDAR, Annual Reports, GLEIF, PermID, Wikidata, Wikipedia). GLEIF, PermID contains info about affiliates. The other sources mainly for the LEU of the group head
Address details of the legal unit (number, street)	GLEIF; EGDAR; Annual Reports for the LEU of the group head
City of the address	GLEIF; DBpedia, Annual Reports and EDGAR for the LEU of the group head
Postal code of the address	GLEIF
Country code of the legal unit	GLEIF; can be derived in case of DBpedia, Annual Reports and EDGAR for the LEU of the group head
Telephone number of the legal unit	No source
Email address of the legal unit	No source
Website of the legal unit	DBpedia, EGDAR, Annual Reports, PermID, Wikidata, Wikipedia for the LEU of the group head
Legal form of the legal unit	No source

Description of the variables	Sources
Type of the legal unit (branch or legal unit)	No source
Activity status of the legal unit	GLEIF and PermID
SPE code of the legal unit (SPE or not)	No source
Date of incorporation of the legal unit	EGDAR for the LEU of the group head (partially), Annual Reports
Date of liquidation of the legal unit	No source
4-digit NACE Rev 2 main activity code of the legal unit	Can be derived from EGDAR for the LEU of the group head
Number of persons employed by the legal unit	DBpedia, EGDAR, Annual Reports, Wikidata, Wikipedia for the LEU of the group head
State name of the legal unit	No source

The table below presents the variables identifying the global enterprise group of the MNE.

Table 4: Overview of information available for the Enterprise group

Description of the variables	Annual reports	EDGAR	GLEIF	PermID	Wikipedia	DBPedia	Open Corporates	Wikidata
Frame reference year	●	●						●
Official name of the global enterprise group	●	●	●	●	●	●	●	●
NACE Rev 2 main activity code of the global enterprise group in EGR at 2-digit level	●	●						
Number of persons employed by the global enterprise group	●	●			●	●		●
Number of persons employed in activities outside the EU and EFTA countries	partially	partially						
Net turnover of the global enterprise group in millions	●	●						
Currency of the net turnover of the global enterprise group	●	●						
Total assets of the global enterprise group in millions	●	●			●	●		●
Currency of the asset of the global enterprise group	●	●			●	●		●
Website of the global enterprise group	●	●			●	●		●

Apart from the identification variables, the data sources analysed provide additional information regarding the employment and the economic performances (income, expenditures, etc.) of the MNE.

Table 5: Overview of information available for the constituent Enterprise

Description of the variables	Sources
Frame reference year	No source identified for any of these variables
Country code of the enterprise	
The start date from which the enterprise exists	
The end date when the enterprise ceased	
Name of the enterprise	
Activity status code of the enterprise	
Institutional sector code of the enterprise	
NACE code of the enterprise	
Number of persons employed by the enterprise	
Net turnover of the enterprise	
Currency code of the net turnover value	
LEID number of the legal unit which is the reporting unit of the enterprise for statistics	

Based on the definition of the enterprise, which is the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources, there were no free data identified to be used for constituent enterprises of the MNE groups.

4.4. Data transformation and visualisation

Step 4 is concerned with the organisation and production of the final database, which combines the information for all indicated MNEs across sources and corresponding variables. In particular, Step 4 stacks the data output from Step 3 in a common database and attempts to homogenise it.

The scope focuses on the transformation of the data in a standardised format to be effectively used in the production of aggregate statistics or visualisation of information. To that end, we loaded the final output in Google Sheets and designed a simple dashboard which presents a summary of information combining multiple sources.

4.4.1. DATA CLEANING

In the previous step, the most important fields were transformed to corresponding variables. Step 4 consists of a set of codes that take all individual output from multiple data sources and check its consistency. In cases where information is not available, this is kept as non –available (NA), instead of omitting the entry.

Given the very large amount of data (and also the number of different sources), at first, it might prove difficult to browse through the dataset. However, opening the database using a spreadsheets software and applying default filters allows for easy browsing.

In accordance with Step 4, we have created an online dashboard on Google Sheets; examples have already been presented in Figures 1 and 2 earlier.

4.4.2. DATA TRANSFORMATION

The next step in our procedure is to transform remaining data with multiple entries (for example, scraping the data from Wikipedia might give four values for MNEs predecessors which are all stored as a unique entry) into single entries, i.e. splitting and transforming the data to a standardised format. Once this process is completed, we end up with two output files: (i) one with the actual data, and (ii) one with the actual data and some band categorisation.

These additional bands facilitate the navigation through the dataset and characterise the data according to their availability. We have also added a specific 'EGR-Type' variable to filter out the data which are not potentially relevant to the EGR. Finally, we illustrate one of many ways to combine information from various sources and display it together in the form of a dashboard.

Figure 3: Example of final database in Excel

	A	B	C	D	E	F	H	I	J	K	L
1	Src	RtrvDate	MNE Name	MNE_Band	Var_Ctg	Var_Type	Var	Var_Band	Value	VarDetailType	VarDetailValue
176	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	General Relations	Character	Affiliation	Var_Band4(101-150)			
177	Wikipedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Operations	Character	AreaServed1	Var_Band3(51-10(	Worldwide		
178	Wikipedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Operations	Character	AreaServed2	Var_Band1(0-25)			
179	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Operations	Character	AreaServed3	Var_Band3(51-10(	Worldwide		
180	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Operations	Character	AreaServed3	Var_Band3(51-100)			
181	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Various	Character	ArtworkInCollecti	Var_Band1(0-25)			
182	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Numeric	Assets1	Var_Band3(51-100)			
183	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Numeric	Assets2	Var_Band1(0-25)	billion		
184	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Character	AssetsCurrency	Var_Band3(51-100)			
185	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Numeric	AssetsUnderMana	Var_Band1(0-25)			
186	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Numeric	AssetsUnderMana	Var_Band1(0-25)			
187	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Character	AssetsUnderMana	Var_Band1(0-25)			
188	DBPedia	08.10.2020	Ab Volvo	MNE_Band4(201-	Financials	Date	AssetsYear	Var_Band1(0-25)			
189	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	General Relations	Character	AssociatedWith	Var_Band3(51-100)			
190	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Various	Character	Award	Var_Band2(26-50)			
191	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Identifiers	Numeric	BaFinInstitute	Var_Band1(0-25)			
192	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Identifiers	Numeric	BiblNatFranc	Var_Band2(26-50)	12016707f		
193	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Identifiers	Numeric	BIC	Var_Band3(51-100)			
194	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Social Networking	Character	Blog	Var_Band1(0-25)			
195	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Corporate Govern	Character	BoardMembers	Var_Band1(0-25)			
196	Wikidata	02.10.2020	Ab Volvo	MNE_Band4(201-	Operations	Character	Brand	Var_Band1(0-25)			

For example, we see in Figure 3 that the user has specified two fields, 'MNE' and 'Var_Ctg', which filter the data for the preferred MNE (Ab Volvo in this example) and the general category of financial variables. This allows the user to browse across specific variables and sources.

4.4.3. DATA VISUALISATION

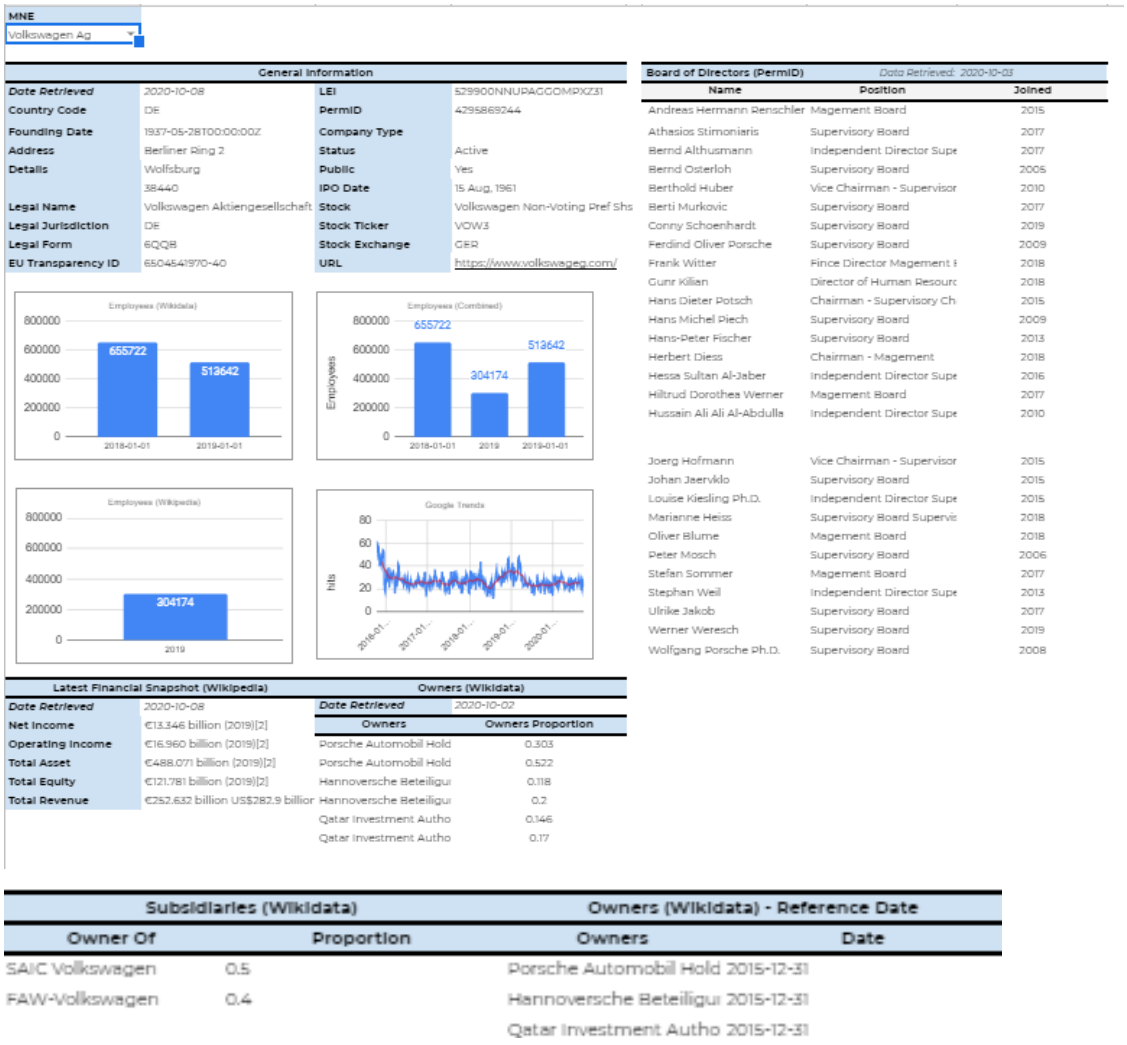
We have provided an organised dataset which could be used in different ways to analyse or visualise the data. Based on our work with the dataset, we have identified that, perhaps, some sources are better at providing particular information. For example, PermID provides a detailed list of board members and information on the MNE, Wikipedia provides a snapshot of latest financial information data, while Wikidata might have some entries regarding the stakeholders of the company, etc.

After careful consideration, we loaded the dataset in a Google Sheets file and prepared a dashboard which combines information from different sources and summarises it together. It is important to note that the dashboard does not display all the information, but rather illustrates a snapshot with the most important, or most easily found, information across multiple sources and gives a general overview of the MNE. Obviously, it is subject to data availability from the original source.

In Figure 4, the dashboard provides a summary for the Volkswagen Ag company combining information from multiple sources. In the top left panel, we have some general information for the company sources from GLEIF and PermID. In the top right panel, we have a detailed list of board members sourced from PermID. Then, we have some figures which display the source of information for the number of employees: (i) Wikidata, (ii) Wikipedia, and (iii) their combination. We also provide

a snapshot of the Google Trends (red line indicates the rolling average) to illustrate public's interest in this MNE for the past years. Finally, the bottom panels display financial information from Wikipedia, as well as stakeholders and subsidiaries information from Wikidata.

Figure 4: Example of Volkswagen AG data visualisation using the dashboard



The interactive dashboard which tries to collect the information across multiple (available) sources and provide a summary to the user.

Figure 5: Interactive Dashboard using two MNEs as examples

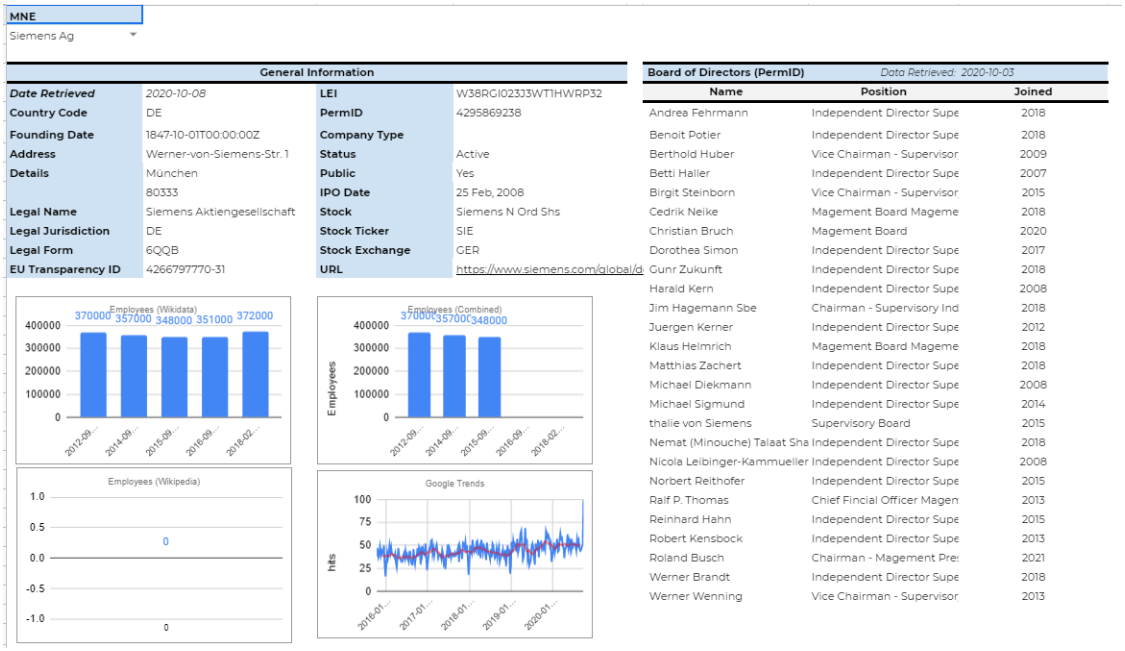
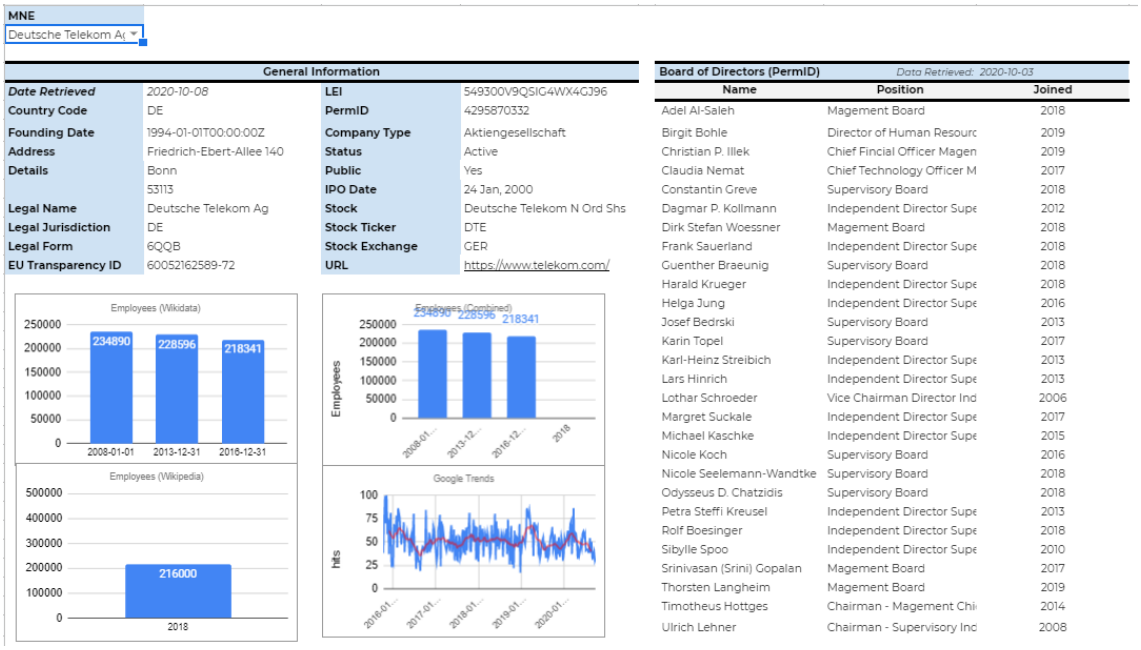


Figure 6: Interactive Dashboard using two MNEs as examples



It is important to highlight that checking the quality and validity of this information is not part of this project. Instead, our aim is to obtain information from multiple sources across the internet and organise it into a standardised database.

4.4.4. DATA AVAILABILITY AND STATISTICS

Having organised all the data into a database, we provide some availability statistics and output.

Table 6: Output by source

Output 1: By Source	
Source	Observations
DBPedia	2 889
EDGAR	1 173
GLEIF	3 310
OpenCorp	4 980
PermID	14 926
Wikidata	11 585
Wikipedia	2 318

Table 6 presents an overview of the total observations across all variables, types, categories and MNEs per source. These figures correspond to the universe of variables (not only the EGR-related). Observations refer to both quantitative and qualitative data. As we see, PermID and Wikidata seem to have the largest availability. It is important not to misinterpret the above number. PermID has a very large collection of values available simply because it provides information on the names of board members, characteristics of the MNE, etc.

Similarly, in Table 7, we have the Top10 MNES ([21]) with most available information across sources. These variables are based on the resulting open source database.

Table 7: Information by MNE

Variable	Observations
Position	4 002
StartDate	4 002
Title	4 002
Subsidiary	1 241
officer.name	1 182
officer.position	1 182
Location1	912
OwnerOf	870
officer.start_date	864

([21]) Defined variables from open data sources.

Industry1	833

Data presented above, as expected, are all corporate finance or generic information (e.g. position of a board member, start date and title of each board member, etc.). Given the very large amount of data (and also the number of different sources), it has proven difficult to browse through the dataset. However, as also mentioned earlier, applying default filters in a spreadsheet software or browsing some information using our dashboard can help the researcher.

To further shed light on data availability statistics, we tried to provide some summary tables for the EGR-related variables and show data by source. We have also generated a long table with the same information by MNE; however, it has been omitted for presentation purposes.

Table 8: Availability of variables by source

Variable	Wikipedia	Wikidata	GLEIF	PermID	DBPedia	Open Corp	Total
Assets1+Assets2+AssetsUnderMgt	0	1	0	0	82	0	**83**
City1+City2	0	0	183		57	84	**324**
Country1+Country2	0	184	183	185	37	42	**631**
Employees1+Employees2+Employees3	134	199	0	0	119	0	**452**
HQ1:HQ5	0	0	870	0	0	0	**870**
Inactive	0	0	0	0	0	46	**46**
IncorpDate	0	0	0	0	0	29	**29**
LegalForm	0	0	183	0	0	0	**183**
LegalName	0	0	183	0	0	0	**183**
LEI	0	78	183	182	0	0	**443**
Name	0	63	0	0	123	50	**236**
Postal	0	0	182	0	0	28	**210**
Status	0	0	183	185	0	46	**414**
TotalAssets	92	145	0	0	1	0	**238**
URL1+URL2	145	212	0	184	142	0	**683**
Sum	**371**	**882**	**2150**	**736**	**561**	**325**	**5025**

Table 9 provides availability statistics for a selection of variables (or sum of variables) which are useful for a more detailed comparison of existing datasets. ‘Total’ refers to the total number of data available across sources and MNEs for the specified EGR variables. Information on the variables summarised in Table 8 can be found in the annex.

5 Conclusions and recommendations

In this project, we investigated the possibilities towards an open source database to potentially enrich the information contained in the EuroGroup's Register as well as to enable an internal quality assessment of this information by internal users in Eurostat in order to ensure confidentiality aspects. It is important to highlight that this project may serve as a proof of concept, though the final output should not be considered as a complete database.

We have reviewed a large number of sources and retrieved data from those that offer free APIs and allow web scraping. A prototype version of this database is illustrated via an interactive dashboard. There is heterogeneity of the information from different sources, even for basic information (e.g. addresses).

The data retrieved from several free data sources might be used, to some extent, as a source to update the EuroGroup Register. Most of the free available data sources refer to MNEs as consolidated figures. There is no information available at constituent enterprise level. Information at this level does not seem to be available on the Internet. This is mainly due to the fact that most sources aim to serve investors or researchers due to accounting practice who focus on the parent organisation and not the corresponding subsidies.

It is also important to highlight that any web scraping activity/code script or APIs will be accurate as long as the original data source keeps the format of the webpage/data environment or API process unchanged. If this changes, then the supplied scripts will require maintenance, which is out of the scope of this work. Furthermore, our efforts are exclusively based on open source data freely available over the web. This data is not organised, curated, checked, confirmed or maintained by a third-party who could guarantee its authenticity or accuracy.

Regarding the information at legal unit level, the main data sources are GLEIF, PermID and OpenCorp (although their free API has a limitation of 40–50 requests). However, the free data sources cover required EGR variables only partially.

Concerning the relevant to the economic variables (turnover, assets) and employment, the most extensive data sources are the annual reports of the MNE that are available on the webpages of the MNE or on annualreport.com website. Almost all MNE present their annual reports and main yearly results. Overall, limited quantitative information is available free of charge. Financial information (retrieved e.g. from Yahoo Finance) is limited without any special arrangements with the providers.

It is important to note that, after reviewing the source, we are not in a position to assess the actual quality, simply because we do not have access to a benchmark. However, we can generally say that sources such as Wikipedia, Wikidata, EDGAR, which provide reference dates and some sort of time series for specific variables, are more useful in the analysis of quantitative data. For qualitative data, the best structured information comes from GLEIF and OpenPermID.

6 Future work

This project has been used as a proof-of-concept to investigate whether information freely available on the web could be retrieved and put together in a meaningful and standardised way. However, since the information in the EGR is confidential, we were not able to assess the quality of the retrieved information. As a next step, it would be important to understand how far the information retrieved fits with the information already available at Eurostat in order to improve the scripts and programmes. Additionally, further investigation might focus on a specific list of MNEs (e.g. selection and ranking by turnover) and on whether their annual financial reports contain supplementary information.

The authors emphasised that part of the requested information may be available (in particular, concerning quantitative information, i.e. financial and economic indicators); however, it cannot be scraped without prior agreement (e.g. with Google finance, scrapping is not allowed). A way forward would be to investigate whether an agreement with a third-party source that already has the necessary information (e.g. Yahoo Finance) is a viable option to enrich the dataset on MNEs. This needs to be considered in the future in order to prepare the necessary conditions either to scrape information directly from the web or to use APIs. The additional information would enable further quality assessment of the information available in the EGR.

In line with the project, the authors recommend to concentrate future research also on the Business Registers Interconnection System (BRIS) infrastructure facilitating access to information on EU companies for the public and capturing aspects of cross-border mergers and foreign branches.

While the current work focuses on the collection of basic information regarding the structure of the enterprise groups, future research could complement the database by monitoring MNEs' activity and help increase emphasis on financial market variables (such as the stock price or corporate bond yields in the secondary market), as well as on real-time event monitoring. This could provide Early Warning signals for MNEs, which attract investors' attention, or attention from the media, and could potentially indicate future mergers, acquisitions, changes in board of directors and corporate governance, signs of profitability and leadership in the market, etc.

7 References

Breton, R., Swiel, N., O'Neil, R. (2015). 'Using Web Scraped Data to Construct Consumer Price Indices', New Techniques and Technologies for Statistics, Eurostat Conference, 9–13 March 2015.

Buono, D., Mazzi, G.L., Kapetanios, G., Marcellino, M., Papailias, F. (2017), *Big data types for macroeconomic nowcasting*, EURONA, pp. 1–4.

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

Choi, H. and H. Varian (2009), 'Predicting initial claims for unemployment benefits', Google Working Paper.

Choi, H. and H. Varian (2012), 'Predicting the Present with Google Trends', *Economic Record*, 88(1), pp. 2–9.

Degiannakis S., Floros C., 2015. 'Introduction to High Frequency Financial Modelling', *Modelling and Forecasting High Frequency Financial Data*, Palgrave Macmillan, London.

European Commission, Eurostat, 'European business statistics methodological manual for statistical business registers — 2021 edition'. Available link here: URL

European Commission, Eurostat, 'NACE Rev.2- Statistical classification of economic activities in the European Community'. Available link here: URL

European Commission, Eurostat, 'Trusted Smart Statistics in a nutshell'. Available link here: URL.

European Commission, Eurostat, 'Statistics explained –EuroGroups register. Available link here: URL

Galbraith, J.W., Tkacz, G. (2007), 'Analyzing Economic Effects of Extreme Events using Debit and Payments System Data', *CIRANO Scientific Series*, Working Paper 2011s–70.

Henderson, J.V., Storeygard, A., Weil, D.N. (2011). 'A Bright Idea for Measuring Economic Growth', *American Economic Review: Papers & Proceedings*, 101(3), pp. 194–199.

Silver, M., Heravi, S. (2001), 'Scanner Data and the Measurement of Inflation', *The Economic*

Journal, 111, F383–F404

Smith-Clarke, C., Mashhadi, A., Capra, L. (2014), 'Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks', CHI2014 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 511–520.

OECD (2019), Analytical Database on Individual Multinationals and Affiliates (ADIMA), Available link here: URL

Regulation (EC) No 177/2008 of the European Parliament and of the Council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing Council Regulation (EEC) No 2186/93. Available link here: URL

Regulation (EU) 2019/2152 of the European Parliament and of the Council of 27 November 2019 on European business statistics, repealing 10 legal acts in the field of business statistics (Text with EEA relevance). Available link here: URL

REGULATION (EU) 2019/2152 of the European Parliament and of the Council of 27 November 2019 on European business statistics. Available link here: URL

Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F. and Skaliotis, M. (2019), 'Trusted smart statistics: Motivations and principles'. *Statistical Journal of the IAOS* 35 (2019) 589–603589DOI 10.3233/SJI-190584IOS Press. Available link here: URL.

R Packages used in the program scripts: igraph, gtrendsR, httr, pdftools, purrr, rjson, rlist, rvest, selectr, tidyverse, WikidataQueryServiceR, XML, xml2.

Python packages used in the program scripts: bs4 (BeautifulSoup), numpy, os, pandas, selenium (webdriver), selenium.common.exceptions (TimeoutException), selenium.webdriver.common.by (By), selenium.webdriver.support (expected_conditions), selenium.webdriver.support.ui (WebDriverWait), time.

United Nations Trade Statistics (2018), *Handbook on Accounting for Global Value Chains*. Available link here: URL

8 Annex

Table 9: Information on variables

Category	Variable	Type (*)	EGR code	Description
Financials	Assets1	N	GEG_T_ASSET	Assets Value (not necessarily Total)
	Assets2	N	GEG_T_ASSET	
	AssetsUnderManagement	N	GEG_T_ASSET	Assets (not necessarily Total) under Management
General Information	City1	C	LEU_CITY_NAME	MNEs registered city
	City2	C	LEU_CITY_NAME	
	Country1	C	LEU_COUNTRY_CODE	Country of the MNE
	Country2	C	LEU_COUNTRY_CODE	
	Employees1	N	LEU_PERS_EMPL	Number of employees occupied by the MNE
	Employees2	N		
	Employees3	N		
	HQ1	C	LEU_ADDRESS	Headquarters address details
	HQ2	C		
	HQ3	C		
	HQ4	C	LEU_COUNTRY_CODE	
	HQ5	C	LEU_POSTAL_CODE	
	Inactive	B	LEU_STA_CODE	True or False
	IncorpDate	D	LEU_DATE_INC	Date of incorporation of the MNE
	LegalForm	N	LEU_LFORM	Legal form of the MNE

Category	Variable	Type (*)	EGR code	Description
	LegalName	C	LEU_NAME	Legal Name of the MNE
Identifiers	LEI	N	LEU_LEI	LEI number of the MNE
General Information	Name	C	LEU_NAME	Legal Name of the MNE
	Postal	C	LEU_POSTAL_CODE	Headquarters address details
	Status	C	LEU_STA_CODE	Activity status of the MNE

(*) N=Numeric, C=Character, B=Binary and D=Date

Table 10: Information on availability of EGR LEU variables by sources (*)

EGR field - Legal Unit (LEU)	Description	Data source	Availability
LEU_FRAME_RYEAR	Frame reference year	OpenCorp	Generally NO
LEU_LEI	LEI number of the legal unit	OpenCorp	YES
LEU_NAT_ID	Primary national ID of the legal unit	OpenCorp	Generally NO
LEU_NAME	Name of the legal unit	OpenCorp	Generally YES
LEU_ADDRESS	Address details of the legal unit (number, street)	OpenCorp	Generally YES
LEU_CITY_NAME	City of the address	OpenCorp	Generally YES
LEU_POSTAL_CODE	Postal code of the address	OpenCorp	Generally YES
LEU_COUNTRY_CODE	Country code of the legal unit	OpenCorp	Generally YES
LEU_TEL_NUMBER	Telephone number of the legal unit	OpenCorp	NO
LEU_EMAIL	Email address of the legal unit	OpenCorp	NO
LEU_WEB	Website of the legal unit	OpenCorp	NO
LEU_LFORM	Legal form of the legal unit	OpenCorp	NO
LEU_TYPE	Type of legal unit (branch or legal unit)	OpenCorp	NO
LEU_STA_CODE	Activity status of the legal unit	OpenCorp	NO
LEU_SPE_CODE	SPE code of the legal unit (SPE or not)	OpenCorp	NO
LEU_DATE_INC	Date of incorporation of the legal unit	OpenCorp	Generally YES
LEU_DATE_LIQ	Date of liquidation of the legal unit	OpenCorp	Generally YES
LEU_NACE_CODE	4-digit NACE Rev 2 main activity code of the legal unit	OpenCorp	NO
LEU_PERS_EMPL	Number of persons employed by the legal unit	OpenCorp	NO
LEU_STATE_NAME	State name of the legal unit	OpenCorp	NO

(*) The table presents only OpenCorp because for all the other sources there is no information available for any of the variables.

Table 11: Information on availability of EGR GEG variables by sources

EGR field - GEG	Description	Data source	Availability
GEG_FRAME_RYEAR	Frame reference year	DBPedia	NO
		OpenCorp	NO
		Wikidata	Partially
		Wikipedia	NO
		EDGAR	YES
		Annual reports	YES
GEG_NAME	Official name of the global enterprise group	DBPedia	YES
		OpenCorp	YES
		Wikidata	YES
		Wikipedia	YES
		EDGAR	YES
		Annual reports	YES
GEG_NACE_CODE_DIV	NACE Rev 2 main activity code of the global enterprise group in EGR at 2-digit level	DBPedia	NO, for some of the groups there is information about operations/segments
		OpenCorp	NO
		Wikidata	NO, for some of the groups there is information about operations/segments /products
		Wikipedia	NO, for some of the groups there is information about operations/segments /products
		EDGAR	YES, SIC 4 digit code that can be converted into NACE 2 digit
		Annual reports	It might be retrieved/derived from the group presentation (business at a glance)
GEG_PERS_EMPL	Number of persons employed in the global enterprise group	DBPedia	Generally YES
		OpenCorp	NO
		Wikidata	Generally YES
		Wikipedia	Generally YES
		EDGAR	YES

EGR field - GEG	Description	Data source	Availability
		Annual reports	YES
GEG_PERS_EMPL_ACT_OUT_EU	Number of persons employed in activities outside EU and EFTA countries	DBPedia	Generally NO
		OpenCorp	NO
		Wikidata	NO
		Wikipedia	NO
		EDGAR	For most of the groups
		Annual reports	For some of the groups
GEG_TURNOV	Net turnover of the global enterprise group in millions	DBPedia	Partially YES (Revenues, sales)
		OpenCorp	NO
		Wikidata	Partially (Revenues/Sales)
		Wikipedia	Proxy indicators (total revenues, total sales)
		EDGAR	YES
		Annual reports	YES, either as net turnover or as net sales
GEG_TURNOV_CUR_CODE	Currency of the net turnover of the global enterprise group	DBPedia	YES
		OpenCorp	NO
		Wikidata	NO
		Wikipedia	YES
		EDGAR	YES
		Annual reports	YES
GEG_T_ASSET	Total assets of the global enterprise group in millions	DBPedia	Generally YES
		OpenCorp	NO
		Wikidata	Partially
		Wikipedia	YES
		EDGAR	YES
		Annual reports	YES
GEG_T_ASSET_CUR_CODE	Currency of the asset of the global enterprise group	DBPedia	YES
		OpenCorp	NO
		Wikidata	NO
		Wikipedia	YES
		EDGAR	YES
		Annual reports	YES
GEG_WEB	Website of the global enterprise group	DBPedia	YES
		OpenCorp	YES
		Wikidata	NO
		Wikipedia	YES
		EDGAR	NO
		Annual reports	YES

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Smart Data for Multinational enterprises (MNEs) – using open source data to obtain information on multinational enterprises

KS-TC-21-007-EN-N

Trusted Smart Statistics endeavour aims to develop statistics in datafied societies, leveraging information from the web (Web Intelligence), using innovative data collection methods and "smart systems". Within the framework of Trusted Smart Statistics and Web Intelligence, this paper presents the results of a proof-of-concept of retrieving information about MNEs and enterprise groups from open source data from the web. It brings together the results of research undertaken within the framework of "Smart Data for MNEs" including scoping and assessment of relevant data sources for developing the necessary components of web scraped data, processing the information for further analysis, transforming and visualising this information, exploiting relevant aspects in order to enhance and extend the availability of information on MNEs. The results of this study provide input for the developments within the Trusted Smart Statistics framework, leveraging information from the web.

For more information
https://ec.europa.eu/eurostat/

Publications Office of the European Union

ISBN: 978-92-76-42218-1